MEASI
Institute *of* Management

BUSINESS ANALYTICS USING PYTHON COURSE MATERIAL

SUBJECT CODE: 534E2C

VISION & MISSION STATEMENTS OF
THE INSTITUTE VISION

- To emerge as the most preferred Business School with Global recognition by producing most competent ethical managers, entrepreneurs and researchers through quality education.

MISSION

- **Knowledge through quality teaching learning process**; To enable the students to meet the challenges of the fast challenging global business environment through quality teaching learning process.
- **Managerial Competencies with Industry institute interface**; To impart conceptual and practical skills for meeting managerial competencies required in competitive environment with the help of effective industry institute interface.
- **Continuous Improvement with the state of art infrastructure facilities**; To aid the students in achieving their full potential by enhancing their learning experience with the state of art infrastructure and facilities.
- **Values and Ethics;** To inculcate value based education through professional ethics, human values and societal responsibilities.

PROGRAMME EDUCATIONAL OBJECTIVES (PEOs)

PEO 1 - Placement; To equip the students with requisite knowledge skills and right attitude necessary to get placed as efficient managers in corporate

companies.

PEO 2 - Entrepreneur; To create effective entrepreneurs by enhancing their critical thinking, problem solving and decision-making skill.

PEO 3 - Research and Development; To make sustained efforts for holistic development of the students by encouraging them towards research and development.

PEO4 - Contribution to Society; To produce proficient professionals with strong integrity to contribute to society.

PROGRAM OUTCOME

PO1 - Problem Solving Skill; Apply knowledge of management theories and practices to solve business problems.

PO2 - Decision Making Skill; Foster analytical and critical thinking abilities for data- based decision making.

PO3 - Ethical Value; Ability to develop value based leadership ability.

PO4 - Communication Skill; Ability to understand, analyze and communicate global, economic, legal and ethical aspects of business.

PO5 - Individual and Leadership Skill; Ability to lead themselves and others in the achievement of organizational goals, contributing effectively to a team environment.

PO6 - Employability Skill; Foster and enhance employability skills through subject knowledge.

PO7 - Entrepreneurial Skill; Equipped with skills and competencies to become an entrepreneur.

PO8 - Contribution to community; Succeed in career endeavors and contribute significantly to the community.

PROGRAM SPECIFIC OBJECTIVES

PSO 1: Finance: The students should demonstrate proficiency in analyzing financial statements, evaluating investment opportunities and making financial decision to maximize shareholders' value.

PSO 2: Marketing: Students should be able to create a comprehensive marketing plan that integrates effective communication strategies, leading to customer success and the accomplishment of marketing objectives.

PSO 3: Logistics: Students will acquire knowledge of inventory management for domestic and global supply chains, thereby developing problem-solving skills in logistics to optimize supply chain efficiency.

PSO 4: Business Analytics: The students should able to analyze data, communicate insights, take data-driven decisions and solve business problems effectively.

SYLLABUS

	Subject Name		L	T	P	O			Marks

Subject Code		Category					Credits	Inst. Hours	CIA	External	Total
3	**Business Analytics Using Python**	Elective	3	-	-	-	3		25	75	100

Course Objectives	
C1	Business data analysis techniques and their theoretical foundations
C2	Visualizations using tableau
C3	To understand business models
C4	Analyse various models
C5	Applications of Marketing Analytics

UNIT	Details	No. of Hours	Course Objectives
I	**Introduction** Introduction to Business Analytics - Evolution of Business Data and Analytics timeline - Types of Analytics - Marketing Analytics Applications - Summarizing & Reporting Marketing Data using Excel	9	C1
II	**Visualizing Business Data using Tableau -** Visualizations Using Python & R - Understanding the Metrics across domains - Developing Metrics - Flowchart for Metric Creation	9	C2
III	**Business Models & Strategies** Business Models - Marketing Engineering – Segmentation Analytics – Clustering Algorithms - Positioning Analysis - Data Mining applications	9	C3
IV	**Marketing Mix Analytics**: New Product development decisions - Pricing the Product - Forecasting the Sales – Allocating the Retail	9	C4

	space & Sales Resource – Consumer Attribution Modelling Methods		
V	**Marketing Mix Analytics Applications** Customer Churn Modelling – Purchase Behaviour Prediction Models- social media Listening and Sentimental Analysis – Market Basket Analysis – RFM Analysis – Recommender Systems development	9	C5
	Total	**45**	

Course Outcomes		
Course Outcomes	On completion of this course, students will;	**Program Outcomes**
CO1	Understand and explain key principles, concepts and terms associated with marketing analytics including the Marketing Metrics, web analytics, big data analytics, social media analytics and analytics trends	PO1, PO6
CO2	Construct a metric identifying the areas to be measured for the individual or corporate and how it makes sense to the business managers.	PO1, PO2, PO5
CO3	Analyse marketing situations using appropriate instruments to formulate marketing strategies and plans, and to evaluate their impact	PO4, PO6
CO4	Analyse marketing situations using appropriate instruments to formulate marketing strategies and plans, and to evaluate their impact	PO4, PO5, PO6
CO5	Apply the marketing Instruments and quantitative methods providing students with an image of the complexity and pitfalls of typical marketing situations and problems	PO2, PO6

Reading List	
1.	https://bedford-computing.co.uk/learning/wp-content/uploads/2015/10/Python-for-Data-Analysis.pdf
2.	https://cfm.ehu.es/ricardo/docs/python/Learning_Python.pdf
3.	Van Rossum G, others (2016). Python Programming Language. URL http://www.python.org/.
4.	Jesus Rogel-Salazar,Data Science and Analytics with Python, 2017

References Books		
1.	"R for Marketing Research and Analytics", Chris Chapman,Springe Publications, 1st Edition, 2015.	
2.	"Business Analytics", Dinesh Kumar U Wiley India, 1st Edition, 2017.	
3.	"Marketing Metrics: The Definitive Guide to Measuring Marketing Performance", Paul W Farris,Pearson Education, 2nd Edition, 2010.	
4.	"Business Analytics- Texts and Cases", Tanushri Banerjee & Arindham BanerjeeSage Publications, 1st Edition, 2019.	
5.	"Marketing Analytics – Data Driven Techniques with Microsoft Excel", Wayne L Winston,Wiley Publications, 1st Edition, 2015..	
Methods of Evaluation		
Internal Evaluation	Continuous Internal Assessment Test	25 Marks
	Assignments	
	Seminars	
	Attendance and Class Participation	
External Evaluation	End Semester Examination	75 Marks
	Total	100 Marks
Methods of Assessment		
Recall (K1)	Simple definitions, MCQ, Recall steps, Concept definitions	
Understand/ Comprehend (K2)	MCQ, True/False, Short essays, Concept explanations, Short summary or overview	
Application (K3)	Suggest idea/concept with examples, Suggest formulae, Solve problems, Observe, Explain	
Analyze (K4)	Problem-solving questions, Finish a procedure in many steps, Differentiate between various ideas, Map knowledge	
Evaluate (K5)	Longer essay/ Evaluation essay, Critique or justify with pros and cons	
Create (K6)	Check knowledge in specific or offbeat situations, Discussion, Debating or Presentations	

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8
CO 1		S				M		
CO 2	S	S			M			
CO 3				M		S		
CO 4				S	M	M		
CO 5		S				S		

S-Strong M-Medium L-Low

UNIT 1

INTRODUCTION TO BUSINESS ANALYTICS

Business analytics is one of the most growing fields in the modern era. Due to the deadly combination of statistics and computer science the scope of business analytics has been growing wider and wider. This evolution of business analytics has resulted in various kinds of career opportunities. That's why it is very important to understand the meaning and the importance of business analytics.

In Business Analytics, we first have to understand the term 'analytics.' Now, analytics generally refers to the science of manipulating data by applying different models and statistical formulae on it to find insights. These insights are the key factors that help us solve various problems. These problems may be of many types, and when we work with data to find insights and solve business-related problems, we are actually doing Business Analytics. The tools used for analytics may range from spreadsheets to predictive analytics for complex business problems. The process includes using these tools to draw out patterns and identify relationships. Next, new questions are asked and the iterative process starts again and continues until the business goal is achieved.

One of the biggest essentials of business analytics is categorized as descriptive analytics, which analyzes historical data to determine particular reactions from a bunch of people. Business analytics refers to a subset of several methodologies, such as data mining, statistical analysis, and predictive analytics, to analyze and transform data into useful information. Business analytics is also used to identify and anticipate trends and outcomes. With the help of these results, it becomes easier to make data-driven business decisions.

The use of business analytics is very popular in some industries such as healthcare, hospitality, and any other business that has to track or closely monitor its customers. Many high-end business analytics software solutions and platforms have been developed to ingest and process large data sets.

Business Analytics Examples

Some of the examples of Business Analytics are:

- A simple example of Business Analytics would be working with data to find out what would be the optimal price point for a product that a company is about to launch. While doing this research, there are a lot of factors that it would have to take into consideration before arriving at a solution.
- Another example would be applying Business Analytics techniques to identify and figure out how many and which customers are likely to cancel the subscription
- One of the highly appreciated examples of Business Analytics is working with available data to figure out and assess how and why the tastes and preferences change of customers who visit a particular restaurant regularly.

Components of Business Analytics

Modern world business strategies are centered around data. Business Analytics, Machine Learning, Data Science, etc. are used to arrive at solutions for complex and specific business problems. Even though all of these have various components, the core components still remain similar. Following are the core components of Business Analytics:

- **Data Storage**– The data is stored by the computers in a way that it can be further used in the future. The processing of this data using storage devices

is known as data storage. Object storage, Block Storage, etc. are some of the storage products and services.

- **Data Visualization**– It is the process of graphically representing the information or insights drawn through the analysis of data. Data visualization makes the communication of outputs to the management easier in simple terms.
- **Insights**– Insights are the outputs and inferences drawn from the analysis of data by implementing business analytics techniques and tools.
- **Data Security**– One of the most important components of Business Analytics is Data Security. It involves monitoring and identifying malicious activities in the security networks. Real-time data and predictive modeling techniques are used to identify vulnerabilities in the system

Types of Business Analytics

There are various types of Business Analytics that are performed on a daily basis across many companies.

Descriptive Analytics

Whenever we are trying to answer questions such as “what were the sales figures last year” or :what has occurred before”, we are basically doing descriptive analysis. In descriptive analysis, we describe or summarize the past data and transform it into easily comprehensible forms, such as charts or graphs.

An example would be finding out the percentage of leads that we couldn’t convert and the potential amount of business that we lost due to this.

Predictive Analytics

Predictive analytics is exactly what it sounds like. It is that side of business analytics where predictions about a future event are made. An example of predictive analytics is calculating the expected sales figures for the upcoming

fiscal year. Predictive analytics is majorly used to set up expectations and follow proper processes and measures to meet those expectations.

Prescriptive Analytics

In the case of prescriptive analytics, we make use of simulation, data modeling, and optimization of algorithms to find answers to questions such as “what needs to be done”. This is used to provide solutions and identify the potential results of those solutions. This field of business analytics has recently surfaced and is on heavy rise since it gives multiple solutions, with their possible effectiveness, to the problems faced by businesses. Let’s say Plan A fails or there aren’t enough resources to execute it, then there is still Plan B, Plan C, etc., in hand.

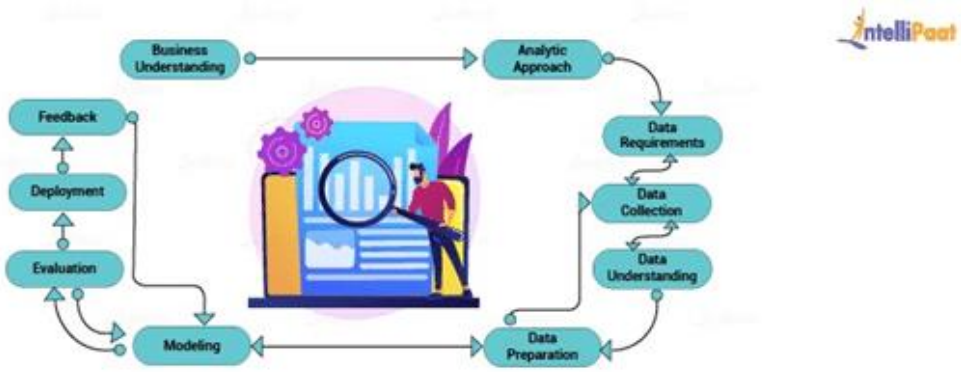
Key Steps in Business Analytics Process

Just like any other thing in business, there is a process involved in business analytics as well. Business analytics needs to be systematic, organized, and include step-by-step actions to have the most optimized result at the end with the least amount of discrepancies.

Now, let us dive into the steps involved in business analytics:

- **Business Problem Framing:** In this step, we basically find out what business problem we are trying to solve, e.g., when we are looking to find out why the supply chain isn’t as effective as it should be or why we are losing sales. This discussion generally happens with stakeholders when they realize inefficiency in any part of the business.
- **Analytics Problem Framing:** Once we have the problem statement, what we need to think of next is how analytics can be done for that business analytics problem. Here, we look for metrics and specific points that we need to analyze.

- **Data:** The moment we identify the problem in terms of what needs to be analyzed, the next thing that we need is data, which needs to be analyzed. In this step, not only do we obtain data from various data sources but we also clean the data; if the raw data is corrupted or has false values, we remove those problems and convert the data into usable form.
- **Methodology selection and model building:** Once the data gets ready, the tricky part begins. At this stage, we need to determine what methods have to be used and what metrics are the crucial ones. If required, the team has to build custom models to find out the specific methods that are suited to respective operations. Many times, the kind of data we possess also dictates the methodology that can be used to do business analytics. Most organizations make multiple models and compare them based on the decided-upon crucial metrics.
- **Deployment:** Post the selection of the model and the statistical ways of analyzing data for the solution, the next thing we need to do is to test the solution in a real-time scenario. For that, we deploy the models on the data and look for different kinds of insights. Based on the metrics and data highlights, we need to decide the optimum strategy to solve our problem and implement a solution effectively. Even in this phase of business analytics, we will compare the expected output with the real-time output. Later, based on this, we will decide if there is a need to reiterate and modify the solution or if we can go on with the implementation of the same.



Applications of Business Analytics

Business analytics is a very useful process that is used in different sectors. Whether it be the IT sector, the healthcare domain, or any other type of business, business analytics can help improve them immensely. Hence, there are a vast number of applications for business analytics. Some of the notable examples of business analytics are:

- Optimization of supply chains
- Forecasting revenue
- Pinpointing reasons for employee attrition
- Fraud detection
- Recommendation systems
- Finding out the number of cabs required in a region
- Price point comparison

Business Analytics vs. Data Analytics

Business Analytics means performing data analysis to draw business insights and offer solutions to complex business problems. It specifically involves dealing with business insights, unlike Data Analytics.

Data Analytics refers to the analysis of already existing data to draw conclusions about the information contained in the data. It is a broader concept and involves business analytics too.

EVOLUTION OF BUSINESS DATA AND ANALYTICS TIMELINE

Evolution of Business Analytics

Today, business analytics has become a buzzword for companies around the globe. Every business, irrespective of its size, is on a lookout for different ways to make sense of the vast amount of raw data available. This is because business analytics has been transforming the way companies function for over a decade now. From targeting the right customers and increasing sales to helping HR personnel select the right candidates and reducing overhead costs; there is hardly any sector where data analytics has failed to tap in.

By now, we are sure that you might have a brief idea of why business analytics plays such a vital role. Now, let us take a step back and analyse how did this all start? Why has this field gained popularity just recently? Has the way businesses use data evolved over the years? What did business analytics look like 20 years ago?

Now, lets get started with the history of business analytics with the help of a timeline, starting from the 1800s to 2020.

How Business Analytics Has Evolved Over the Years

BA in the 1800sThe need to stay ahead

The first use of data to stay ahead of his competitors dates back to 1865. During this time, Mr Richard Miller Devens described in his book how Sir Henry Furnese, a banker, was always one step ahead by actively gathering information and acting on it before any of his competitors. This makes it clear that professionals such as Sir Furnese relied more on data and empirical evidence, rather than gut instinct.

BA in the late 1800s The Advent of Scientific Management

During this time, Frederick Taylor introduced the first-ever system of business analytics in the United States of America, and he called it scientific management. The purpose of this system was to analyze the production techniques and labourers body movements to identify greater efficiencies.

BA in the early 1900sThe Transformation of the Manufacturing Industry

Frederick Taylors scientific management system inspired Henry Ford, who hired Taylor as his consultant. Ford was willing to measure the time each component of his Ford Model T took to complete on his assembly line. This analysis transformed his work and the manufacturing industry across the globe.

BA in the 1950sThe first hard drive disk by IBM

Computers werent accessible in the early 1900s but had a massive demand during World War II. As they were still rudimentary, punch cards or tapes were used to store information. However, in 1956, the tech giant, IBM invented the first hard disk drive. This allowed users to save a vast amount of data with better flexibility.

BA in the late 1900sThe Emergence of Business Intelligence

Owing to the lower prices for storage space and better databases, the next generation of business intelligence solutions was all set to step in. By now, there was a considerable amount of data available but not a centralized place to store it. To address this problem, Ralph Kimball and Bill Inmon proposed similar strategies to build data warehouses (DW).

BA in the new MilleniumAvailability of different analytical solutions

By this time, medium and large-sized businesses had already realized the value of business intelligence solutions. Companies such as IBM, Microsoft, SAP, and

Oracle were at the forefront of offering such solutions to change the way businesses function.

BA in 2005Accessibility of Data for the Common People

Considering the extensive usage of data, companies started directing their efforts on improving the speed at which the information was available. New business analytics tools were introduced to ensure technical as well as non-technical people were able to mine the data and gain insights.

Around this time, the increasing interconnectivity of the business world led to the need for real-time information. This was when Google Analytics was introduced. Google wanted to provide a free and accessible way for users to analyze their website data.

BA from 2005 to 2020The Bread and Butter for Companies globally

With the internet available to almost everyone and the increasing data, companies needed better solutions to store and analyze all the information. Building computers with more storage capacity and better speed wasnt possible for many, so companies resorted to using several machines at the same time. This was the beginning of cloud computing.

Since the last decade, big data, cloud computing, and business analytics have become integral for almost all companies. The new advancements have made these technologies even better. Now, data analytics and science are known to be the future. From advertising and marketing to recruiting and planning operational activities, these terms are tossed around in every field.

Summary

From the advent of business analytics in the early 1800s to it being the part-and-parcel for every business in 2020, we have covered almost every element to help you understand how it has evolved over all these years. If you are willing to make a career in a field which is expected to rule the corporate world, it is time to sign up for our PGDM in Research and Business Analytics provided in association with IBM.

1. Evolution of Business Analytics

In recent times, business analytics has evolved into a much more advanced set of tools and techniques assisted by automation and big data. Initially, business analytics was limited to a few corporate applications used by only the major MNCs. The first adoption of computing for business was noticed in the use of report-building, presentations and data entry using applications such as Microsoft Excel. Later on, more advanced applications that involved multi-dimensional data processing and data analytics were seen using add-ons such as PowerPivot in Excel. This was still a long way from BI or business intelligence, which the highly accredited firm, Gartner.

Once the concept of BI was introduced into the market, corporations around the globe wanted their hands on it. BI used various digital instruments, technologies, and metrics to evaluate business performance and helped companies get valuable insights. This facilitated more informed decision-making and ensured that all the important organizations during that time adopted software such as Tableau, SAS, or Microsoft Power BI to support their business activities and operations. With BI, concepts such as 'web questions,' collaboration, data security, and sourcing data from databases and distributed file systems came into mainstream use as well. By now, business analytics was not just used by large MNCs and conglomerates but also medium-level and much smaller enterprises as well. This introduced the world to an era where analytics helped create research

models, design models, and simulators that further helped companies use data to forecast and predict future outcomes more accurately than ever.

2. Business Analytics Historical Facts

Analytics and visualizations have been used throughout history without the support of computers and software. This was done by manually plotting graphs using statistical methods and manually recording data. This was quite different from the business analytics that we recognize and know about. The more modern version of business analytics was only used much later in the 20th century to identify trends during the Second World War. This process of identifying trends helped code-breakers use data from encrypted messages such as destination (recipients of the messages), origin, and the time and date of these messages to find out what information these contained. This is a more modern use of analytics to predict information. However, we have also seen business analytics being used a bit earlier in history. Sir Henry Furnese, a well-documented banker, had been extensively using data during the 1860s to stay ahead of his competition.

Here are some more historical facts about business analytics:

- During the early 1900s, Henry Ford, inspired by Frederick Taylor's scientific management system, hired him in order to measure the performance of the assembly line of his famous Ford Model T. This led to a series of events that had transformed the manufacturing industry and production lines across the world. This also helped Henry Ford make his assembly line as efficient as possible.
- In 1956, IBM introduced the first hard disk drive that allowed users to store data that can be used for business or corporate purposes.
- During the 1970s, Bill Inmon started discussing the concept of a data warehouse to solve the problem of storing vast amounts of data for business intelligence.
- During the 1980s, the first business data warehouse was developed by IBM researchers Barry Devlin and Paul Murphy.
- In this period between the 1990s and early 2000s, various solutions and software were introduced, such as business intelligence tools by companies like SAP, Microsoft, SAS and IBM alongside relational databases.
- After the early 2000s, common people started using data more proactively for personal purposes. This also led to more corporate use of data through employees extensively using organisational data. More tools were also introduced during this time with which individuals can use business intelligence tools without extensive training. Eventually, Google Analytics was introduced that allowed website owners to analyse statistics about their website, such as trends in website visits.
- After 2010, business intelligence and analytics truly took off, being adopted worldwide by companies and businesses around the world. This also

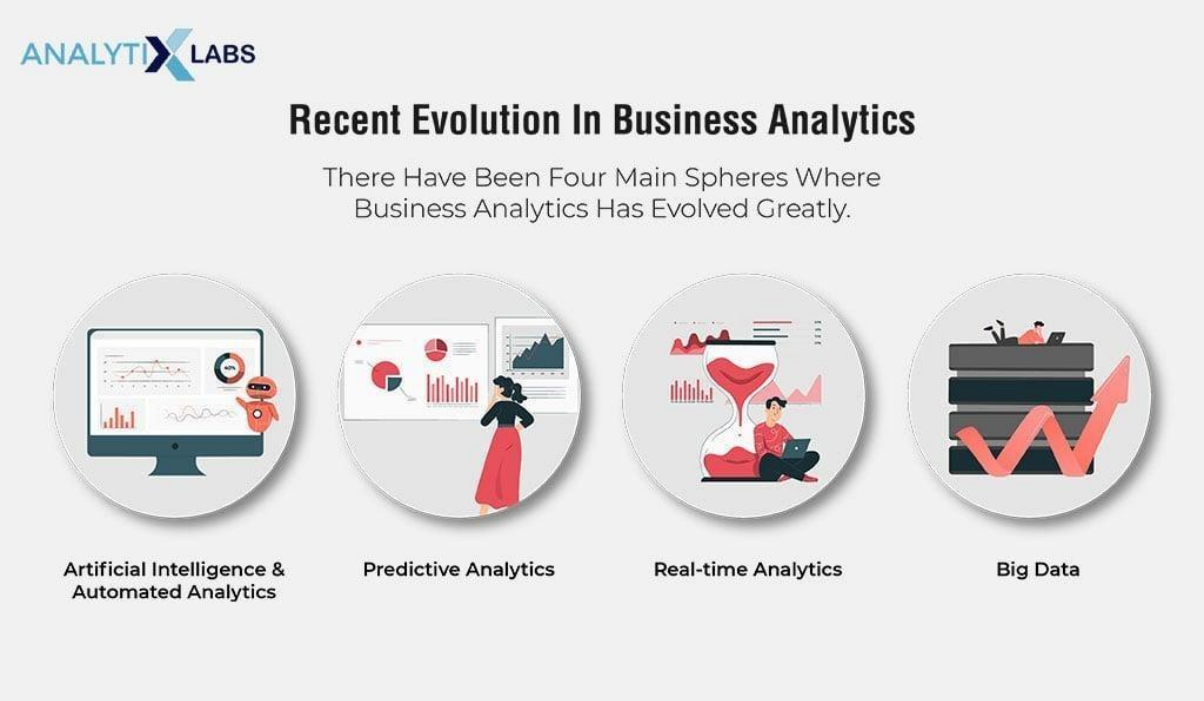
pushed us to an era of cloud computing and extensive use of Artificial Intelligence or automation.

3. Recent Evolution in Business Analytics

The recent evolution that business analytics has experienced can be fundamentally traced back to the introduction of automation in analytics and the concept of big data. The advent of big data meant that analytics along with various data sources should become more scalable and more powerful. This helped in introducing more advanced tools and systems that are compatible with large volumes of data. The emergence of cloud technologies also meant that data did not need to be on-site. There was also a huge demand for automating analytical tools by this time due to the massive amount of data that needed to be worked upon. All of this motivated companies to upgrade their existing software into more capable applications that can process massive datasets rapidly and from multiple sources such as from the cloud and distributed file systems rather than just the traditional RDBMS. Business analysts were also now armed with predictive and forecasting abilities that were now more accurate than ever with the help of modern business analytics. This is where businesses truly understood the importance of data analytics in business. All this technology had already existed, but the industry's growing requirement encouraged businesses of all sizes to start incorporating data analytics into daily operations.

There have been four main spheres where business analytics has evolved greatly, these are:

- Artificial Intelligence and Automated Analytics
- Predictive Analytics
- Real-time Analytics
- Big Data



4. Why Business Analytics Is Essential To Build A Competitive Business?

Here are some reasons why Business Intelligence and analytics are essential for a successful business:

- Business analytics is very helpful in reducing risk in business operations.
- It increases revenue and helps companies churn out more profit.
- Analytics allow companies to make better and more informed business decisions that are data-driven.
- It helps increase the operational efficiency of projects or processes.
- It also helps in effectively using resources such as human assets.
- It helps in reducing wastage and in cutting operational costs.
- Business analytics directly allows companies to stay ahead of their competition and outperform them.
- Analytics optimises processes and helps in business process automation.
- It also allows companies to replicate successful results and understand how the success was achieved.

- It helps in forecasting and prediction of outcomes.
- Business analysts help monitor performance and evaluate operations through identifiable metrics or Key Performance Indicators(KPI).
- It helps in the identification of anomalies and factors that affect market and customer behaviour.
- Real-time business analytics helps in taking rapid data-centric decisions.
- It helps in improving the sustainability of businesses and projects in the long run.

5. Big Data – Overview

Data is the most important aspect of business analytics. The Meaning of 'big data is, fundamentally, data that is too complex, and more importantly, too massive to be processed by traditional methods or software. Modern business analytics is highly focused on mining, analyzing, and extracting insights from large datasets. This is why big data is a huge deal in this day and age. Companies collect massive chunks of data on a daily basis as they keep operating, and this data consists of multiple complex pieces of information associated with customers, products, performance, finance, etc., that must be maintained properly and then effectively used for future operations. For instance, businesses can use past data collected over many years to predict customer behavior, such as buying patterns. Or, banks can use customers' credit history when deciding upon eligibility. Histories of other customers can also be studied to predict the likelihood of certain events, such as buying substitute products.

Big data is a treasured part of business analytics, and analysts have started getting comfortable working with big data in organizations of all sizes. Big data has also directly encouraged the use of Machine Learning (ML) when working with massive

datasets. ML allows data to be extracted without errors, structured and utilized much more effectively.

6. Advantages of Big Data For Business Analytics

Here are some advantages of using Big Data for Business Analytics:

- **More Informed Decision Making:**

Big data allows companies to work with a huge amount of user or customer data that allows them to make better decisions. A lot of past or historical data also helps companies predict more accurately.

- **Faster Decision Making:**

Using machine learning, big data can predict and identify patterns rapidly. This speeds up decision-making and allows businesses to use a lot of data very fast. The compiled data can also be processed alongside fresh incoming data, which has helped companies make informed decisions in real-time.

- **Cost Reduction:**

Generating and collecting your own data is the most cost-effective method, as sourcing data from third parties is much more expensive. Also, saving past data enables you to generate only the relevant data and not waste time on generating the same leads or information again. Also, big data, in general, helps businesses cut costs. Thus, the importance of data analytics in business is truly seen when companies increase their revenue with the help of big data.

- **Providing Better Service to Customers:**

With the help of big data, companies can understand customers' buying behavior and can then effectively provide the after-sales services or complementary services that they might require. With the help of an enormous amount of data such as posts, comments, and messages on social media, companies can even

figure out customer sentiments and how satisfied these customers are with the company's products or services.

- **Finding New Opportunities:**

Businesses can discover new opportunities with such huge volumes of data. With the help of this data, businesses can come up with major solutions to problems or come up with fresh concepts and product ideas. They might also be able to figure out how to improve sales or reach out to more consumers finally.

- **Having a Competitive Edge over Competition:**

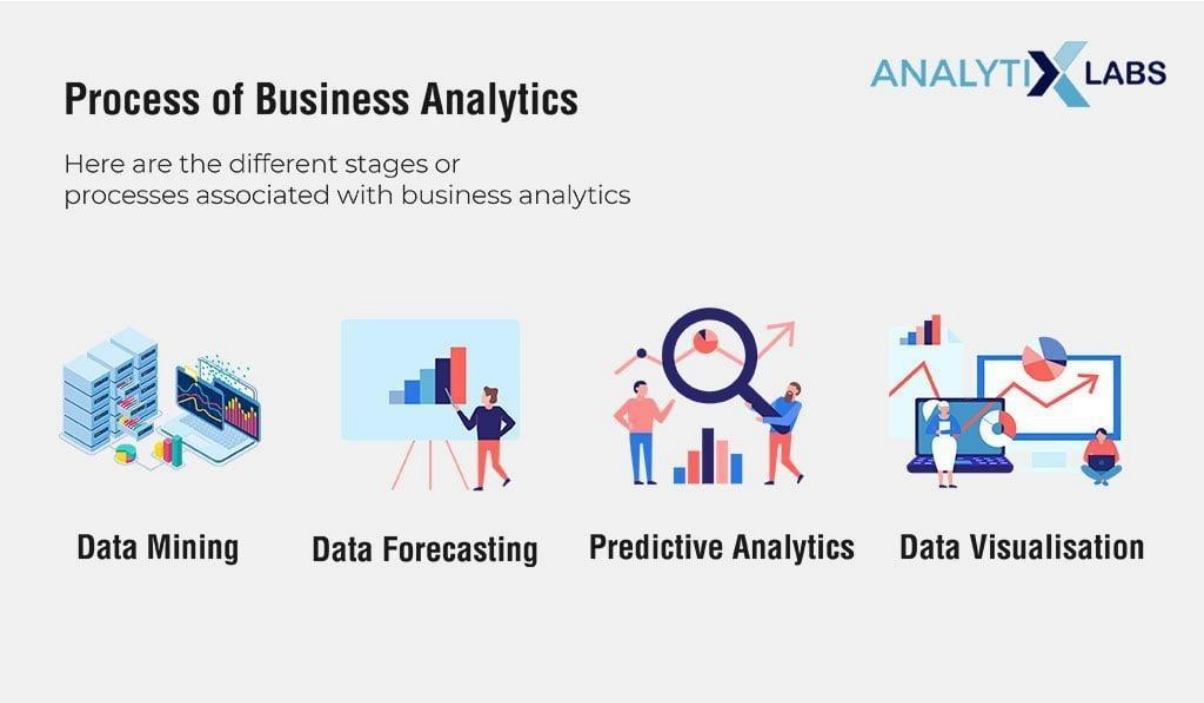
Businesses that use big data are able to project more superior business models and make better predictions that allow them to have the edge over their competitors who are using more traditional methods. Forecasting is more accurate with big data, and companies can directly increase their performance in the market with the help of these advanced insights.

7. Process of Business Analytics

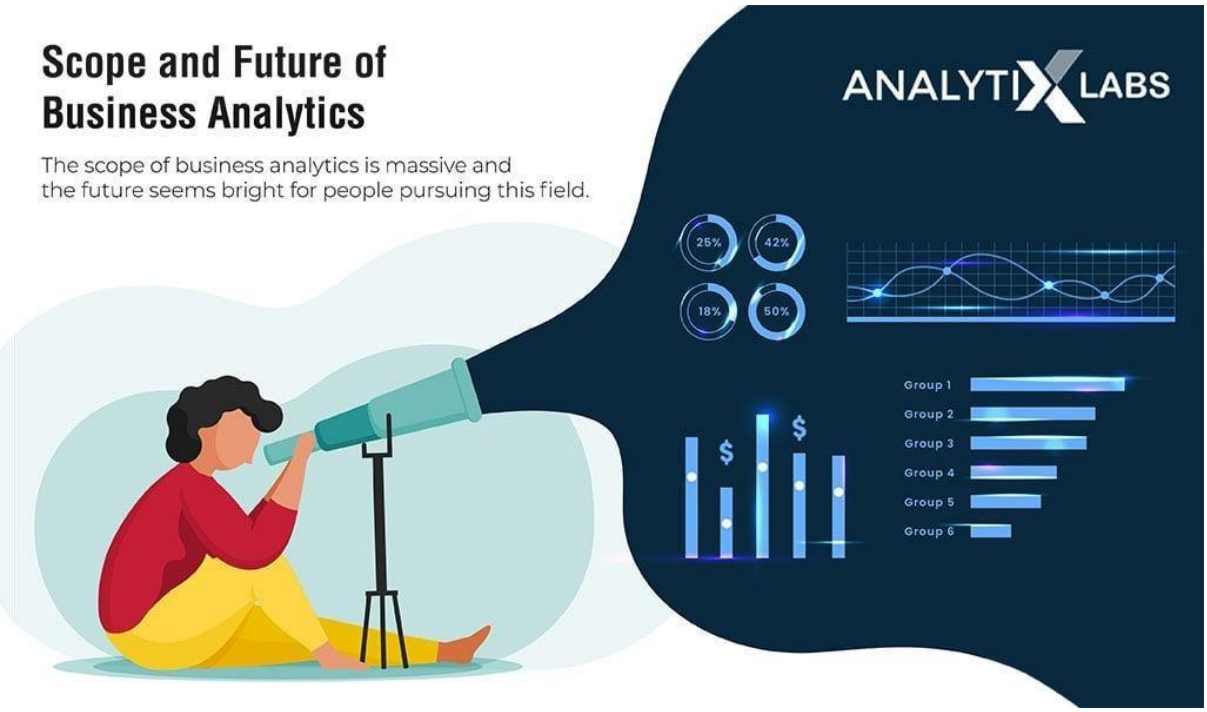
Here are the different stages or processes associated with business analytics:

- **Data Mining:** Mining data is one of the most important things in business analytics, meaning that without data mining, businesses would not be able to extract cognizable information from structured and especially unstructured data. Data mining is performed with the help of statistical tools or machine learning in order to identify past trends and hidden patterns in data by analysing them. Once the data is processed or mined, it is transferred into data warehouses where it can be shared with multi-dimensional databases. Business analysts are then tasked with examining these patterns and presenting them in charts, plots or graphs.

- **Data Forecasting:** Forecasting is a very important process that involves deep analytics of past or historical data in order to estimate the performance of products, the condition of the market or customer behaviour. This can also help forecast events that require an allocation of budget. Machine learning is crucial for accurate forecasting. Data forecasting is highly used for forecasting market conditions, financial statuses, asset requirements and other future requirements. For more details, please read ***Business Forecasting: Meaning, Methods & More***
- **Predictive Analytics:** This is similar to data forecasting, however, this is even more advanced as it uses Artificial Intelligence to evaluate risk and potential abnormalities. This allows businesses to make better decisions. Predictive analytics facilitates informed operational decisions that take aberrations, anomalies or potential events into account. Predictive analytics can also help cut ahead of the competition with the power of data. Predictive analytics help marketers and advertisers recommend products and services to customers more effectively as well.
- **Data Visualisation:** This is one of the most important processes that represent insights in cognizable formats such as diagrams, tables, graphs and charts. This helps identify patterns in data manually with ease and makes reporting to non-technical staff and stakeholders easier. Visualisation is also important for projecting the important information in data using the best possible (compact but informative) method from massive datasets.



8. Scope And Future of Business Analytics



The scope of business analytics is massive, and the future seems bright for people pursuing this field. More than 2.5 quintillion bytes of data are generated every day, thus ensuring that businesses cannot function without business analytics anymore. The world produces more data now in two days than we did in total

from the beginning of civilization till 2003. This simply reinforces the fact that if we wish to effectively use this data to enhance the performance of our website, business, product, service, we must rely on business analytics.

53% of companies use business and data analytics in their daily operations, while many smaller enterprises have started adopting analytics ever since the Covid-19 pandemic started. Business analytics is also one of the most desirable jobs now, with companies always in dire need of skilled analysts. The growing dependency on data and the reliance on business analytics will only increase with more advancements in technology and a more fast-paced world where the competition is armed equally well with business intelligence and data.

TYPES OF DATA

WHAT IS DATA ANALYTICS IN BUSINESS?

Data analytics is the practice of examining data to answer questions, identify trends, and extract insights. When data analytics is used in business, it's often called business analytics.

You can use tools, frameworks, and software to analyze data, such as Microsoft Excel and Power BI, Google Charts, Data Wrapper, Infogram, Tableau, and Zoho Analytics. These can help you examine data from different angles and create visualizations that illuminate the story you're trying to tell.

Algorithms and machine learning also fall into the data analytics field and can be used to gather, sort, and analyze data at a higher volume and faster pace than humans can. Writing algorithms is a more advanced data analytics skill, but you don't need deep knowledge of coding and statistical modeling to experience the benefits of data-driven decision-making.

WHO NEEDS DATA ANALYTICS?

Any business professional who makes decisions needs foundational data analytics knowledge. Access to data is more common than ever. If you formulate strategies and make decisions without considering the data you have access to, you could miss major opportunities or red flags that it communicates.

Professionals who can benefit from data analytics skills include:

- Marketers, who utilize customer data, industry trends, and performance data from past campaigns to plan marketing strategies
- Product managers, who analyze market, industry, and user data to improve their companies' products
- Finance professionals, who use historical performance data and industry trends to forecast their companies' financial trajectories
- Human resources and diversity, equity, and inclusion professionals, who gain insights into employees' opinions, motivations, and behaviors and pair it with industry trend data to make meaningful changes within their organizations

4 KEY TYPES OF DATA ANALYTICS

1. Descriptive Analytics

Descriptive analytics is the simplest type of analytics and the foundation the other types are built on. It allows you to pull trends from raw data and succinctly describe what happened or is currently happening.

Descriptive analytics answers the question, "What happened?"

For example, imagine you're analyzing your company's data and find there's a seasonal surge in sales for one of your products: a video game console. Here,

descriptive analytics can tell you, “This video game console experiences an increase in sales in October, November, and early December each year.”

Data visualization is a natural fit for communicating descriptive analysis because charts, graphs, and maps can show trends in data—as well as dips and spikes—in a clear, easily understandable way.

2. Diagnostic Analytics

Diagnostic analytics addresses the next logical question, “Why did this happen?” Taking the analysis a step further, this type includes comparing coexisting trends or movement, uncovering correlations between variables, and determining causal relationships where possible.

Continuing the aforementioned example, you may dig into video game console users’ demographic data and find that they’re between the ages of eight and 18. The customers, however, tend to be between the ages of 35 and 55. Analysis of customer survey data reveals that one primary motivator for customers to purchase the video game console is to gift it to their children. The spike in sales in the fall and early winter months may be due to the holidays that include gift-giving.

Diagnostic analytics is useful for getting at the root of an organizational issue.

3. Predictive Analytics

Predictive analytics is used to make predictions about future trends or events and answers the question, “What might happen in the future?”

By analyzing historical data in tandem with industry trends, you can make informed predictions about what the future could hold for your company.

For instance, knowing that video game console sales have spiked in October, November, and early December every year for the past decade provides you with ample data to predict that the same trend will occur next year. Backed by upward

trends in the video game industry as a whole, this is a reasonable prediction to make.

Making predictions for the future can help your organization formulate strategies based on likely scenarios.

4. Prescriptive Analytics

Finally, prescriptive analytics answers the question, "What should we do next?"

Prescriptive analytics takes into account all possible factors in a scenario and suggests actionable takeaways. This type of analytics can be especially useful when making data-driven decisions.

Rounding out the video game example: What should your team decide to do given the predicted trend in seasonality due to winter gift-giving? Perhaps you decide to run an A/B test with two ads: one that caters to product end-users (children) and one targeted to customers (their parents). The data from that test can inform how to capitalize on the seasonal spike and its supposed cause even further. Or, maybe you decide to increase marketing efforts in September with holiday-themed messaging to try to extend the spike into another month.

While manual prescriptive analysis is doable and accessible, machine-learning algorithms are often employed to help parse through large volumes of data to recommend the optimal next step. Algorithms use "if" and "else" statements, which work as rules for parsing data. If a specific combination of requirements is met, an algorithm recommends a specific course of action. While there's far more to machine-learning algorithms than just those statements, they—along with mathematical equations—serve as a core component in algorithm training.

USING DATA TO DRIVE DECISION-MAKING

The four types of data analysis should be used in tandem to create a full picture of the story data tells and make informed decisions. To understand your

company's current situation, use descriptive analytics. To figure out how your company got there, leverage diagnostic analytics. Predictive analytics is useful for determining the trajectory of a situation—will current trends continue? Finally, prescriptive analytics can help you consider all aspects of current and future scenarios and plan actionable strategies.

Depending on the problem you're trying to solve and your goals, you may opt to use two or three of these analytics types—or use them all in sequential order to gain the deepest understanding of the story data tells.

Strengthening your analytics skills can empower you to take advantage of insights your data offers and advance your organization and career.

MARKETING ANALYTICS APPLICATION

Marketing analytics is the process of collecting, analyzing, and interpreting data related to marketing efforts and activities. The primary goal of marketing analytics is to gain valuable insights and understanding into the performance of marketing strategies and campaigns. By examining data, businesses can measure the effectiveness of their marketing initiatives, identify patterns and trends, and make data-driven decisions to optimize their marketing efforts.

Marketing analytics can provide valuable information on various metrics, such as customer behavior, conversion rates, return on investment (ROI), customer acquisition costs, and overall campaign performance. This data-driven approach enables businesses to allocate resources more efficiently, identify areas for improvement, and ultimately enhance their marketing strategies to reach and engage their target audience effectively.

Benefits of marketing analytics

In today's marketing environment, the significance of precise data cannot be overstated. Consumers have grown increasingly discerning when it comes to

selecting the branded content they interact with, while filtering out irrelevant media.

To captivate the attention of the ideal customer, brands must heavily depend on accurate data to craft personalized advertisements tailored to individual interests, rather than relying on generalized demographic assumptions. By doing so, marketing teams can deliver the most relevant ads precisely when and where they are needed, effectively guiding consumers through the sales journey. Here are some of the benefits of using marketing analytics.

Understanding consumer behavior

Marketing analytics allows you to track consumer behavior and see what they are doing, so you can plan for the future. Every customer goes through a unique journey before making a purchase decision. By leveraging marketing analytics, businesses can gain valuable insights into how customers interact with their brand across various touchpoints, such as social media and search. These insights allow marketers to provide personalized information and tailored experiences at each stage of the customer journey, ultimately driving growth and fostering customer loyalty.

Making informed decisions

You can't make informed decisions if you don't have the information to back them up. And when it comes to marketing, there are so many variables involved that it's impossible to know what works unless you have something to compare it to. Marketing Analytics gives you access to information that helps you make better decisions about your marketing plan. Without it, you'd be making choices based on gut instinct or personal preference alone—and those aren't always great bases for decision-making!

Integrating marketing data for comprehensive analysis

Marketing teams often run campaigns across different platforms and teams. Gathering data from various channels can be challenging. With the use of marketing analytics, you can simplify processes by centralizing all marketing data for comprehensive analysis. This integration of marketing data empowers businesses to make data-driven decisions and optimize their marketing strategies effectively.

Enhancing brand awareness

In today's competitive market, brand awareness plays a crucial role in attracting customers. Marketing analytics enables businesses to analyze key metrics across social channels and compare their brand awareness with competitors. By examining search volumes, web traffic, and social media sentiment, marketers can develop strategies to improve brand awareness and gain a competitive edge. With marketing analytics, businesses can track their brand's performance and make data-driven decisions to enhance brand visibility and customer perception.

Staying ahead of competition

In today's competitive landscape, businesses must leverage every advantage to stay ahead. Marketing analytics provides businesses with a competitive edge by enabling data-driven decision-making, optimizing campaigns, and enhancing customer experiences. By harnessing the power of marketing analytics, businesses can gain valuable insights into market trends, customer behavior, and competitor strategies. This knowledge empowers businesses to make informed decisions, refine their marketing strategies, and outperform their competitors.

Application of marketing analytics with examples

Marketing analytics is a powerful tool that can help you make good decisions, but it's not magic. Here are some of the best examples of marketing analytics applications –

Use of marketing analytics to improve its website

Amazon has long been a leader in marketing analytics, using its data-driven approach to continue innovating its website and product offerings. For example, they have used data from their customers to help them improve their search functionality, including personalizing results based on each user's browsing history. This personalized experience helps visitors find what they're looking for faster and more efficiently, which keeps them returning to Amazon.

Using marketing analytics to make content recommendations

Netflix is another company that extensively uses marketing analytics to create content recommendations more likely to appeal to each viewer's tastes. They collect data about what people watch (and don't watch) every month and use this information to inform their recommendations algorithmically.

Customer insights

Walmart has an enormous amount of data about its customers' shopping habits. It knows where they live when they shop there, how much money they spend at the store each year, and much more! This information allows Walmart to target specific customers with coupons or other offers that appeal directly to them.

Marketing analytics skills

Here are some of the key marketing analytics skills you must know

Data analysis

One of the critical marketing analytics skills is data analysis. This involves using statistical methods to analyze large data sets and draw conclusions from them. A good marketer must identify patterns in their data and determine how to use them to improve their marketing strategy.

Analytics tools

Next, analytics managers must be onboarded and comfortable with various automation tools and analytics platforms, because of the vital role these tools play in reducing the time from consumer engagement to consumer insight.

Spreadsheet skills

Businesses use spreadsheets for many things, including managing budgets and sales figures. Marketers need good spreadsheet skills to effectively manage all these different aspects of their role within an organization's hierarchy.

Data visualization

Data visualization can take many forms, including charts, graphs, maps, etc. It's important to choose a format that's appropriate for your audience. If you are working with people who are unfamiliar with the data or industry-specific terminology, you may want to stick to a simple horizontal bar graph.

Marketing analytics tools

Marketing analytics tools are software programs that help marketers and business owners gather and analyze their marketing campaign data. They measure how well a campaign performs, so you can see what kind of return you are getting on your investment. Here are some of the top marketing analytics tools you can use.

Google Analytics

Google Analytics is one of the most widely used and powerful marketing analytics tools. It provides information on how many people are visiting your website, what they are doing there, and how you can improve your site for better conversions.

Semrush

Semrush functions as a marketing analytics solution designed to grant insights into your competitors' actions. By utilizing SEMrush, you can gain a competitive edge in search outcomes and easily oversee rivals' social media profiles and brand activities.

MixPanel

MixPanel is a marketing analytics tool that focuses on user engagement. It allows you to see who uses your product or service and helps you understand how they interact with it so you can optimize their experience.

Heap Analytics

Heap Analytics is a tool that allows you to see how your users interact with your website, and it also helps you understand why they are doing what they are doing. With this information, you can make data-driven decisions about your marketing strategy based on user behavior rather than assumptions or guesses.

Cyfe

Cyfe is a dashboard tool similar to Heap Analytics in many ways, but it also has some unique features that make it worth checking out. It allows you to create custom dashboards and visualize data in new ways—and it integrates with other marketing analytics tools like Google Analytics to compare and contrast data sets across multiple platforms.

Read more: Why choose MBA in Marketing?

How does marketing analytics help businesses?

The biggest advantage of marketing analytics that helps businesses is that it allows you to take a more scientific approach to the market. This means that you can measure the effectiveness of your campaigns and then use that data to make smarter decisions about how you spend your money.

With the advantages of marketing analytics, businesses can find out which of their ads work best and which ones don't. This allows them to optimize their advertising budget by only spending on the ads that are getting the best results.

Marketing Analytics also helps with customer acquisition. By analyzing trends in customer behavior and preferences, businesses can tailor their marketing efforts toward those customers who are most likely to buy from them. This makes it easier for companies to attract new customers without spending as much on advertising or other promotional materials.

SUMMARIZING & REPORTING MARKETING DATA USING EXCEL

Why Marketing Analytics Using Excel

The reason for preference of Excel for marketing analytics industry wide is threefold:

Easy to Use: Excel does not involve stringent coding techniques. It has commands that are simple and user friendly which guide you through the process. It is the first step for someone looking to enter the field of analytics.

Range of Application: Excel is one tool capable of storing, sorting, arranging, computing, analyzing and then presenting data in the most lucid manner. This is the reason why it has crossed industries and departments with its utility.

Entry Point to Complex Data Analytics: Learning excel is fundamental for any professional or academic career in data analytics. Its visual, its transparent and it keeps things simple and hence its wide application.

Marketers handle tons of data whether it be related to sales, product pricing, customer feedbacks and other customer insights. All firms are engaging in data driven marketing in order to optimize marketing efforts to the empowered consumer. Analytics is the key to convert this data into information.

Marketing analytics using Excel can be easily done with the varied functionalities of excel in analytics domain. Let's touch upon those domains and try to get a fair understanding of how to do marketing analytics using excel.

Excel to Summarize Marketing Data

Think of a scenario where you are the Area Sales Manager at a footwear firm. You are to analyze the following:

- Investigate the sales volume and % Sales for every month and for different products.
- Analyze the impact of weekends and festive seasons on your sales volume.
- Study the impact of a marketing promotion on sales.

Do these scenarios seem relevant to you? They should because these are the kinds of analysis most of us would be doing once, we get to our jobs.

The data that is made available is in the form of rows and columns of random numbers and text. It is important to slice and dice this data to gain insights.

Pivot Tables help with summarizing large detailed data set into compressed insights. Not only that, with the help of slicers it is easy to navigate through data by quick filtering. Pivot tables also help in data visualization which makes it easy to study as well as present data.

Some other important functions that help in data summarization are as follows:

- Counting and Summing Functions: COUNTIF, COUNTIFS, SUMIF, SUMIFS, AVERAGEIF, AVERAGEIFS
- Statistical Functions: AVERAGE, STDEV, RANK, PERCENTILE, PERCENTRANK
- Array Functions: TRANSPOSE, FREQUENCY

Pricing

Understanding product pricing and how that particular price impacts customer is one of the most important and frequent issues for a marketing manager. In order to do so, marketers should have a fair understanding of the consumer demand curve.

Linear and Power Demand Curves

These curves are used to study the consumers purchase trend with respect to prices keeping the elasticity of the product in mind. Pricing analytics is one important domain of marketing analytics using excel.

Linear Demand curves represent a straight-line relationship between price and demand.

Power Demand curve forms an arc between the price and demand.

Excel Solver to Optimize Price

Excel has some add-in functions that work as readily available tools for complex analysis. One of them is the solver add-in. Often in marketing, we tend to have projects that require us to **maximize profits or minimize one or the other thing**.

In case the price elasticity for a particular product is not known, then a demand curve for a product can be obtained by identifying the lowest and the highest price obtainable for the product. The lowest, highest and a mid-way price can be used to estimate product demand.

Price Bundling

Until now we have discussed linear pricing models. The customer pays equal amount whenever he makes a purchase of the product. But what happens when the products are bundled together and the **price of the bundle is not equal to the summation of prices of individual products**.

Why do companies go for product bundling?

For the simple reason which is to **motivate the customer to purchase several products and services from the same company**. In bundled price, companies try to sell products for a lower price than what would be charged if products would be bought individually.

How to determine that price which is attractive for the customers and is profitable to the company?

The pricing of bundles is based on the assumption of consumer surplus. **Consumer surplus** is basically the value customer attaches to a product minus the actual cost of the product. The consumer is most likely to choose a product combination that has the highest consumer surplus.

Forecasting

Often marketers need to establish relationship between different variables. Whether it be the effect of pricing on sales, the impact of advertisements on product demand or how is allocation of a particular shelf space related to sale of that product.

Linear and Multiple regression models are used to help forecast the impacts of such variables. Regression models can be run with the Data Analysis add-in in excel. This is one important applications of marketing analytics using excel.

Analysis of What Customer Wants

Marketing analysts more often than not have to figure out **what aspect of a particular product are more appealing to customer** and which ones are least.

Conjoint Analysis

Conjoint analysis helps in **determining which product attributes drives sales for a particular product**. What are the factors that influence sales of apparels in a showroom?

- Is it the price of the clothes?
- Is it the look and feel?
- Is it the promotional discounts?
- Is it the personal assistance provided by sales team in the show room?

Conjoint analysis is done by listing the various product options available and then weighing them based on listed attributes. The consumer is asked for their preference of particular product attributes and rank them in order.

After the data is gathered, regression analysis can be used to highlight the importance of product attributes. This is another great representation of marketing analytics using excel.

Estimating Customer Value

If a company keeps on spending more to acquire customers than the value customers generate, then such companies will struggle to remain in business. Here it becomes important to analyze **customer lifetime value** and use these calculations to drive profit for the company.

Monte Carlo Simulation

Monte Carlo simulation is used when the **profitability is to be calculated for outcomes in a process that is unpredictable due to intervention of random variables**.

Think of yourself as the marketing manager or a product manager of a firm. You are looking to introduce a new product in the market and are looking to estimate the ideal profits from the product. There are various random variables involved namely the market scenario that will determine the price of the product and its sales and unit cost. In such scenarios, Monte Carlo simulation leads the way.

Optimizing Retention and Acquisition Spending

Firms generally have two ways in which they can increase their revenues. They can either go for **retaining their existing customers** which comes at a cost or work on **acquiring new customers** which again has an associated cost.

It is the balance of the two that firms generally want to achieve.

Excel solver can be effectively used to solve such problems and reach at a conclusive solution.

Retailing

The billing desk at a retail shop is an epicenter of truckloads of data that can be used to gain useful customer insights. This data if used properly can drive growth for a particular business. This is yet another application of marketing analytics using excel.

Marketing research studies have found that **customers usually buy pairs or sets of some products together**. This insight can be used to place these two products close to each other to increase total store sales.

Market Basket Analysis using Excel

Market basket refers to the **list of products a customer purchases on a visit to a retail store**. There are useful insights that can be drawn from a customer's basket.

For e.g.; Most customer, when they are buying a torch, will also buy batteries because the torch batteries may die anytime and for such situations the customer wants to keep themselves stocked.

Lift is the most widely used tool for market basket analysis. Lift for two products purchased together is calculated by the formula:

= (Actual no. of times the combination occurs)/ (Predicted no. of times combination occurs if items were purchased independently)

Let's say for instance lift is to be calculated for a combination of bread and eggs brought together.

Let's say the actual no. of transactions where eggs and bread were purchased together are 200. There is total 1000 transactions from the store where the no. of

times egg was purchased is 300 and the fraction of times bread was purchased is 600.

Fraction of times eggs were purchased= 300/1000= 0.3

Fraction of times bread was purchased= 600/1000= 0.6

Lift for a combination of bread and eggs= 200/ (1000*0.3*0.6) = 200/180= 1.11

Lift much greater than 1 indicate a tendency of customers for items to buy these items simultaneously.

Multiple two-way and three-way combinations can be analyzed for a store with the help of functionalities of excel.

The lift concept can be used to optimize store layouts as it can suggest various product combinations that have a higher tendency to be bought together in order to increase store sales.

UNIT 2

Visualizations Using Python & R

The human brain processes visual data better than any other kind of data, which is good because about 90% of the information our brains process is visual. Visual processing and responses both occur more quickly compared to other stimuli. Ever wonder why you can pick out detail in an image with ease while staring at spreadsheets makes your head hurt? The brain processes data in visuals or images faster than data in text or rows of numbers.

You're probably tired of hearing that information is proliferating at a rate that humans can barely comprehend, let alone keep up with. The good news is, you don't have to! Machine learning and advanced analytics are helping humans make sense of large amounts of structured and unstructured data by leaning into our

natural ability to make a better sense of visuals than the raw data we want to understand. This where the power of visualizations is apparent.

Both Python and R are advanced coding languages that can produce beautiful images that allow humans to understand vast datasets with ease. In this article, we'll look at the ways both languages do it and give you some code you can use to create visuals of your own!

What are data visualizations?

Simply put, data visualizations allow humans to explore data in many different ways and see patterns and insights that would not be possible when looking at the raw form. Humans crave narrative and visualizations allow us to pull a story out of our stores of data.

The phrase "A picture is worth a thousand words" is expressly true when turning huge piles of data into images a viewer can actually understand and derive meaning from. Children's storybooks contain lots of images, but very few words. As kids, we don't know many words, but the visuals allow us to easily understand the story.

In our modern digital world, we have huge amounts of data all around us. Data scientists and ML engineers get most of the data they deal with data in a structured or unstructured data format, however, it's difficult for humans to understand and analyze this. Data visualizations (or graphical representations of data) are vital for understanding the data. They help users explore data through visual elements like charts, graphs, plots, maps, and other visualizations.

Different types of exploratory data analysis

In every dataset, we have many variables (also called features, input-variables, or independent-variables) and target/output variables (also known as labels,

dependent-variables, classes, or class-labels). The data scientist's job is to completely understand each feature individually and the relationship between different features. The goal is to get ready the dataset for ML algorithms implementation.

We have three methods for exploratory data analysis:

Univariate analysis

In the univariate analysis, each variable is analyzed individually. It will get us to the complete statistical data for each feature. There are a variety of data visualization techniques for univariate analysis, including Box Plot, Histogram, PDF, CDF

Bivariate analysis

Bivariate analysis is performed to find the relationship between each feature with the target variable. Data visualization techniques for bivariate analysis are Scatter Plot and Heatmap

Multivariate Analysis

As the name signifies, multivariate analysis is performed to understand the relationship between different features of the dataset. One of the main multivariate analysis data visualization techniques is the Pair Plot.

We'll discuss all these visualization techniques in detail in the next section.

Data Visualization in Python

There are a wide array of libraries you can use to create Python data visualizations, including Matplotlib, seaborn, Plotly, and others. A Python data visualization helps a user understand data in a variety of ways: Distribution, mean,

median, outlier, skewness, correlation, and spread measurements. In order to see what you can do with a Python visualization, let's try some on a dataset.

Creating Python visualizations

Let's take a toy dataset featuring data on iris flowers to understand data visualizations in depth. The data set consists of 50 samples from each of the three species of Iris Flower: Setosa, Virginica, and Versicolor. Here "Species" is target variable and it has 4 features "Sepal Length," "Sepal Width," "Petal Length," and "Petal Width."

Import Libraries

First import basic libraries like numpy and pandas and Python data visualization libraries like matplotlib and seaborn.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Understanding the Dataset

Next, load the data set from sklearn libraries:

```
from sklearn.datasets import load_iris
iris = load_iris()
```

Convert this dataset into a data frame and here are the top 5 rows with 4 features (Sepal Length, Sepal Width, Petal Length, Petal width) and one target variable (Species).

	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5	3.6	1.4	0.2	setosa

Here's the code for that:

```
print(data.shape) #print number of rows and columns
>(150, 5)

print(data['Species'].value_counts()) # Counts of every unique Species value
>          virginica                  50
           versicolor                 50
           setosa                     50
 Name: Species, dtype: int64
```

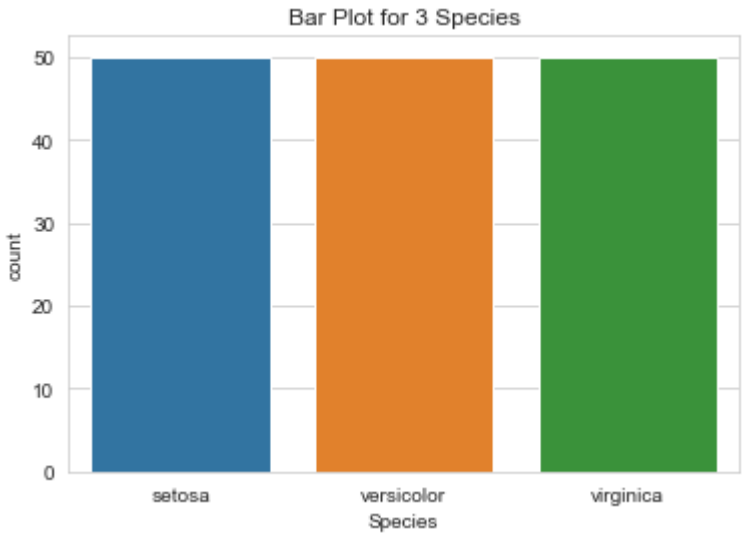
Observations: From the above outputs we can see, there are a total of 150 data points and data is distributed among 3 species equally. So, we can say this is a balanced dataset.

Bar Plot

A bar plot is a plot that presents categorical data with rectangular bars. The length or height of bars is proportional to the frequency of the category. We can count the values of various categories using bar plots.

Here, we are plotting the frequency of the three species in the Iris Dataset.

```
sns.countplot('Species',data=data
)
plt.title('Bar Plot for 3 Species')
plt.show()
```



Observations:

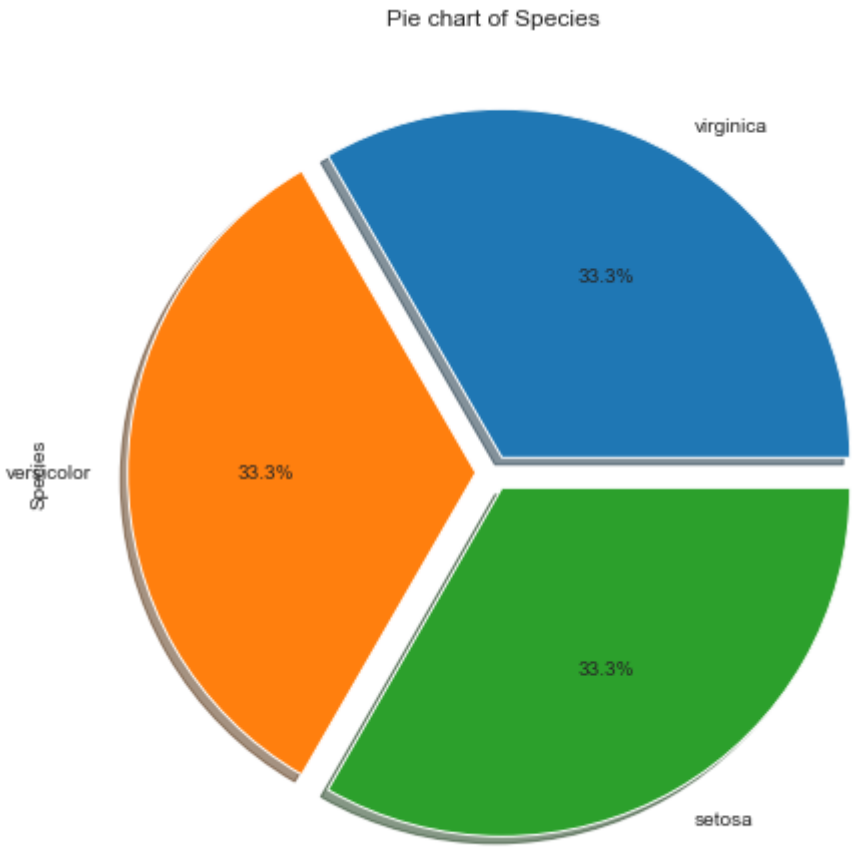
- All bars are of the same height as we know their frequencies are equal.
- Iris Dataset is a balanced dataset.

Pie Chart

Pie Chart is a circular chart that uses pie slices to show the relative size of data. The arc length of each pie slice is proportional to the quantity it represents. It works beautifully on categorical values. There are different variants of pie charts available.

We can use this code to plot a pie chart for 3 species of Iris flower:

```
data['Species'].value_counts().plot.pie(explode=[0.05,0.05,0.05],autopct='%1.1f%%',shadow=True,figsize=(8,8))
plt.title("Pie chart of Species")
plt.show()
```



Observations:

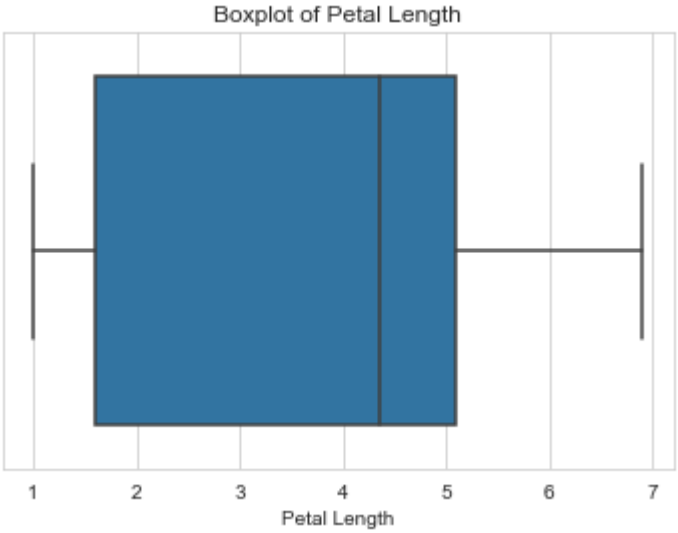
- All three flowers are equal in proportion i.e. 33% each.
- Balanced and imbalanced datasets can be easily classified using a pie chart.

Box-plot

Box-plot gives us a five-number summary of any variable: the minimum, maximum, the sample median, the first and third quartile. Box-plot helps in measuring two observations:

1. Skewness of distribution
2. Outliers (Outliers comes outside the box-plot)

```
sns.boxplot(x='Petal Length', data=data)
plt.title('Boxplot of Petal Length')
plt.show()
```



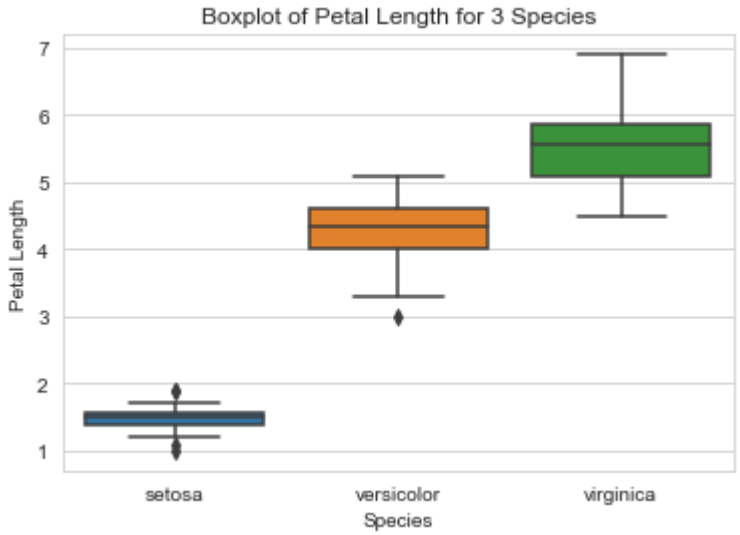
Observations: With the above box-plot visualization we can measure the following parameters:

- The minimum is 1.0
- The maximum is 6.9
- The range is Maximum – Minimum = 5.9
- The sample median is 4.3
- The first quartile Q1 is 1.6
- The third quartile Q3 is 5.1
- The IQR(Interquartile range) is Q3-Q1= 3.5
- The mean value will be between 3.5 to 4.
- There is no outlier in this box-plot

- Petal Length is left-skewed.

We can also draw a box-plot for ‘Petal Length’ for all three different species in a single plot.

```
sns.boxplot(x='Species',y='Petal Length', data=data)
plt.title('Boxplot of Petal Length for 3 Species')
plt.show()
```



Observations:

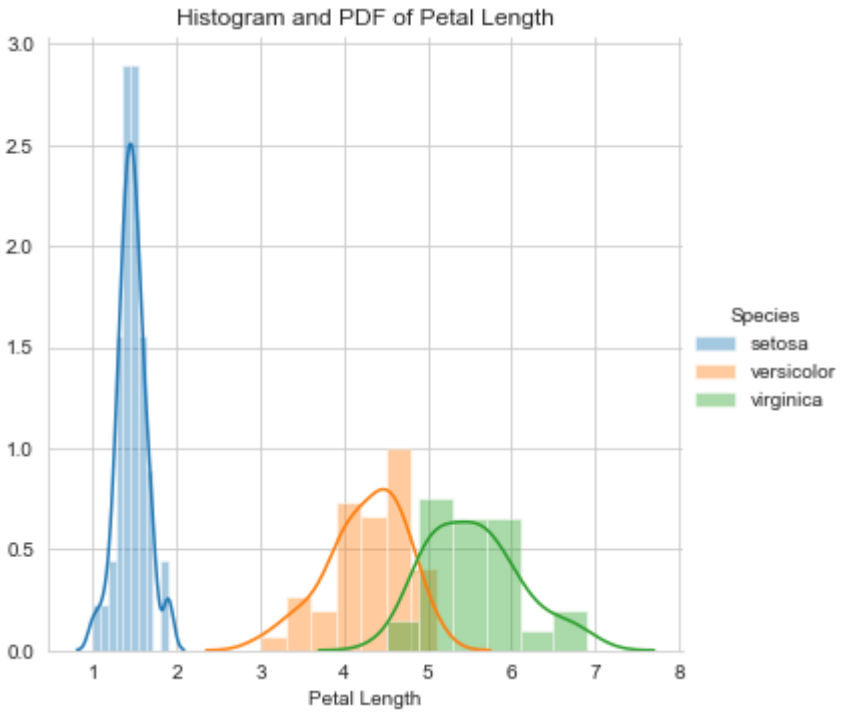
- Petal Length of Setosia is the smallest of all three.
- Virginica has the largest petal length.
- There is an outlier in Versicolor.

Similarly, we can draw box-plots for other features as well.

Histogram and PDF

A histogram is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable. Histogram basically represents the number of points that exist for each bin(range of values). PDF is a Probability Density Function which is basically smoothening of the histogram.

```
sns.FacetGrid(data, hue="Species", size=5) \
    .map(sns.distplot, "Petal Length") \
                                .add_legend();
plt.title('Histogram and PDF of Petal Length')
plt.show();
```



Observations:

In the above graph, lines which are drawn are PDF and Bars drawn is a histogram.

From the above graph, we can simply write if-else statements like:

If Petal Length < 2.3 then flower species is Setosia else-if Petal Length > 5.8 then flower species is Verginica else- if 2.3<Petal Length<3.8 then the flower is Versicolor.

So, probability can be easily calculated using these if-else statements that's why this graph is called probability density function.

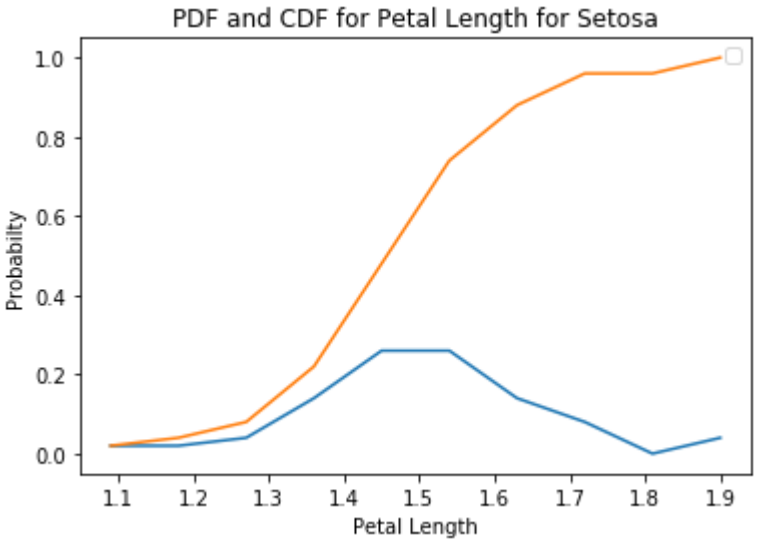
- Setosa is easily separable on the basis of Petal Length.

- There is an overlap between Versicolor and Virginia.
- Distributions are Uniform/Gaussian distribution.

CDF (Cumulative Density Function)

As the name signifies, the cumulative distribution function gives you the cumulative probability associated with a variable. It is the total count up to a certain number. CDF is always in increasing order

```
data_cdf=data[data['Species']=='setosa']
counts, bin_edges = np.histogram(data_cdf['Petal Length'], bins=10, density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)
plt.xlabel("Petal Length")
plt.ylabel("Probability")
plt.title('PDF and CDF for Petal Length for Setosa')
```



Observations:

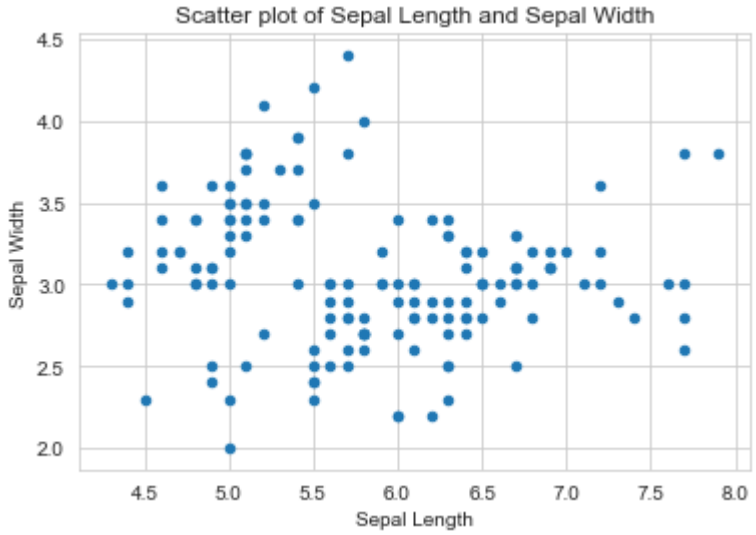
In the above graph, the Blue line is PDF and the Orange line is CDF.

- From CDF it is easy to calculate percentages like approximately 90% of Setosa flowers have Petal Length less than 1.7 which can not be calculated using PDF.
- Approx 50% of setosa flowers have Petal Length less than 1.5

Scatter Plots

A scatter plot is a plot that shows the relationship between two variables of a data set.

```
data.plot(kind='scatter', x='Sepal Length', y='Sepal Width') ;
plt.title("Scatter plot of Sepal Length and Sepal Width")
plt.show()
```

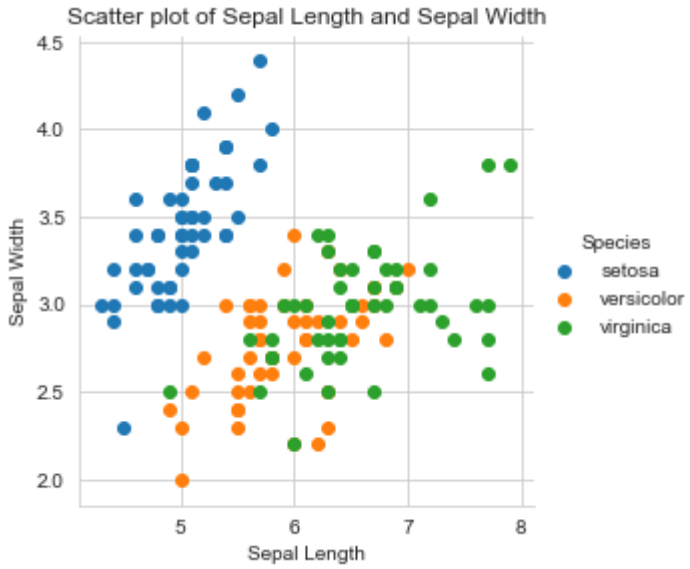


Observations:

In the above plot, we cannot differentiate different flowers, all points are in the same color.

```
sns.set_style("whitegrid");
sns.FacetGrid(data, hue="Species", size=4) \
```

```
    .map(plt.scatter, "Sepal Length", "Sepal Width") \
                                        .add_legend();
plt.title("Scatter plot of Sepal Length and Sepal Width")
plt.show();
```



Observations:

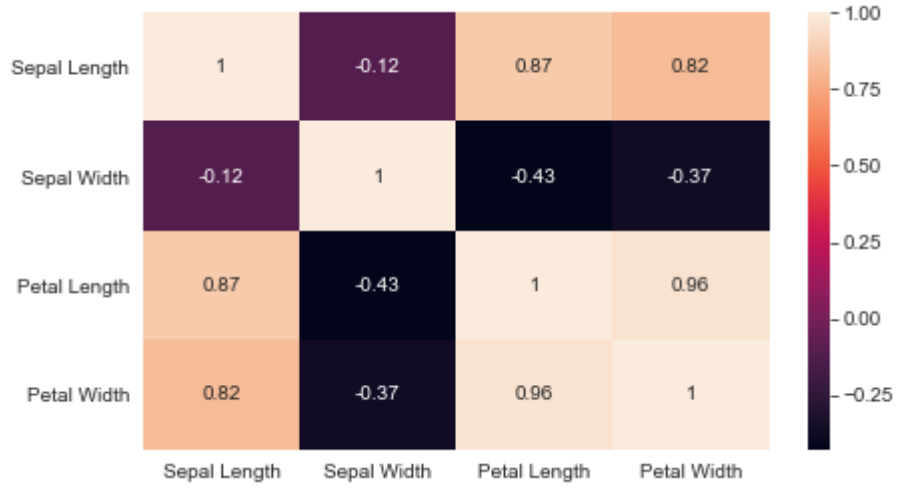
- Setosa (blue) is easily differentiable
- Versicolor and virginica overlap in Sepal Length and Sepal Width as well. They are not easily separable.

Heat Map

A heatmap is a graphical representation of data in which data values are represented as colors. It uses color in order to communicate the correlation between two variables. Values are between -1 to 1. 1 denotes perfect positive correlation. 0 means no correlation and -1 means the highest negative correlation.

Let's plot a heat map for the Iris dataset.

```
sns.heatmap(data.corr(),annot=True
)
```

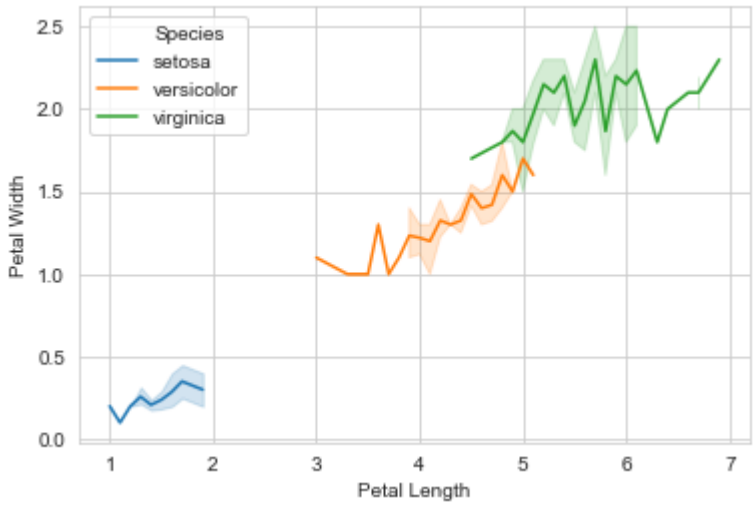


Observations:

- Petal Length and Petal Width shows highest positive correlation 0.96
- Petal Length shows a high positive correlation of 0.87 with Sepal Length as well.
- Petal Width shows a high positive correlation of 0.82 with Sepal Length as well.
- Petal Length and Sepal Width shows a negative correlation of -0.43
- Sepal Width shows a negative correlation with the other 3 features.

Line chart

The line chart represents a series of data points connected by a straight line. It is generally used to visualize data that changes over time. Here, we will draw a line chart showing how Petal Width changes with change in Petal Length.



Observations:

- Large Petal Length means Large Petal Width
- The line chart is not straight-line; it's fluctuating.
- Small Petal Length means Small Petal Width

Word Cloud

Word cloud is an image made up of words that makes a quick visualization. The size of the word shows the frequency of the word in text data. The word which is biggest in size has the highest frequency in text data.

Data Visualization in R

R is extremely easy and flexible to use with minimum code to create visualizations. R has a wide array of libraries you can use to create beautiful data visualizations,

including ggplot2, Plotly, and others. In order to see what you can do with R visualization, let's try some visualizations on the same toy dataset.

Import libraries

First, import data visualization library ggplot2 and in-built datasets library datasets.

```
library(ggplot2)
library(datasets)
```

Understanding the Dataset with R

Next, load the in-built iris data set from the library and analyze the data.

```
data(iris)
head(iris)
```

Here are the top 6 rows in the iris dataset with 4 features (Sepal.Length, Sepal.Width, Petal.Length, Petal.width) and one target variable (Species).

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa

5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Here's the code for that:

```
dim(iris) #print number of rows and columns
>(150, 5)

levels(iris$Species) # Display unique Species value
> [1] "setosa"     "versicolor" "virginica"

table(iris$Species)
> setosa versicolor  virginica
     50         50         50
```

Observations: From the above outputs, we can see there are a total of 150 data points and data is distributed among 3 species equally. So, we can say this is a balanced dataset.

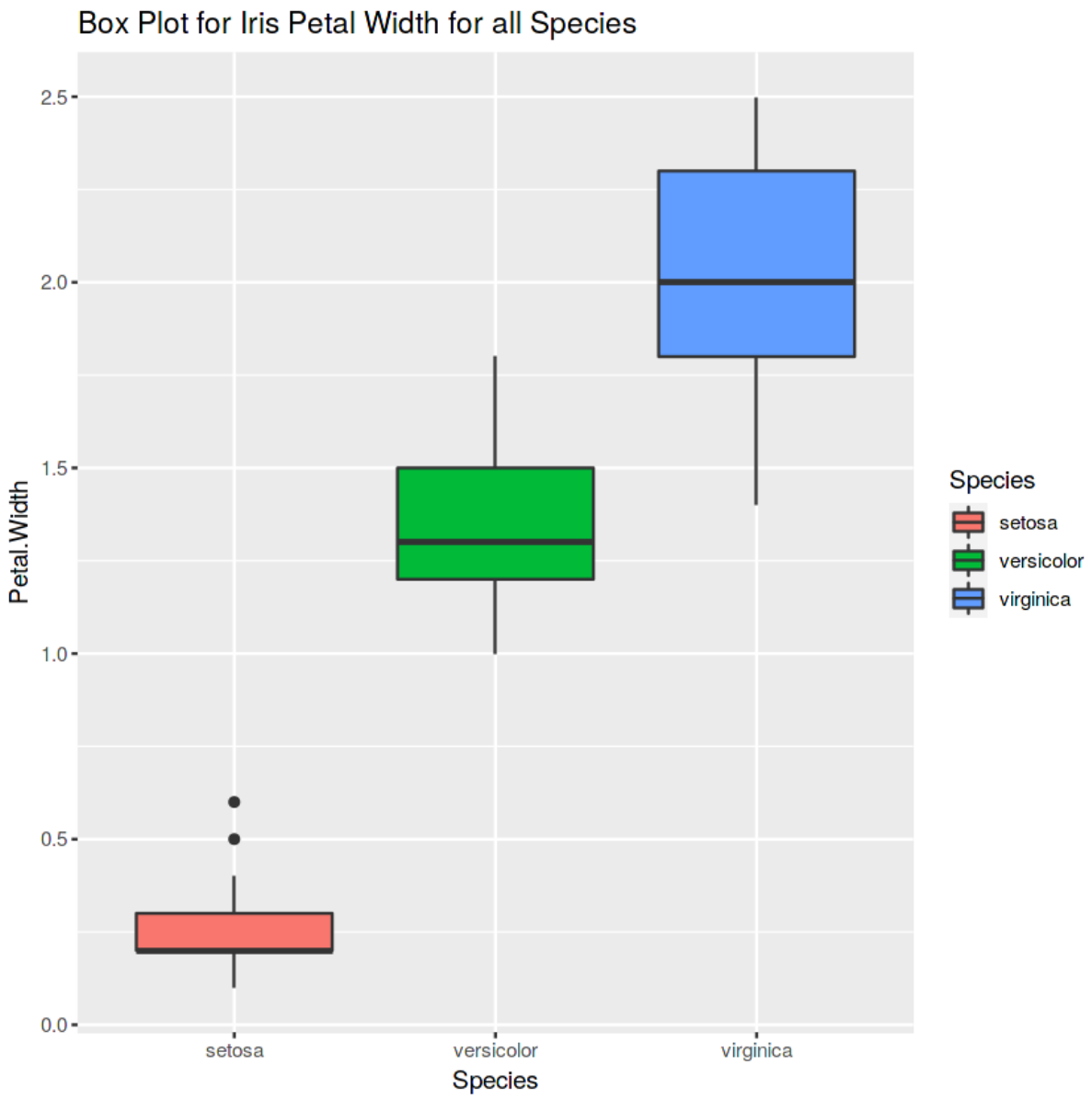
Box-plot

Like we see in Python box plots, in R as well Box-plot helps in measuring two observations:

1. Skewness of distribution
2. Outliers (outliers fall outside the box-plot)

We have drawn box-plot for 'Petal Width' for all three different species in a single plot.

```
ggplot(iris, aes(Species, Petal.Width, fill=Species)) +   geom_boxplot()+  labs(title =
"Box Plot for Iris Petal Width for all Species ", x = "Species"
```



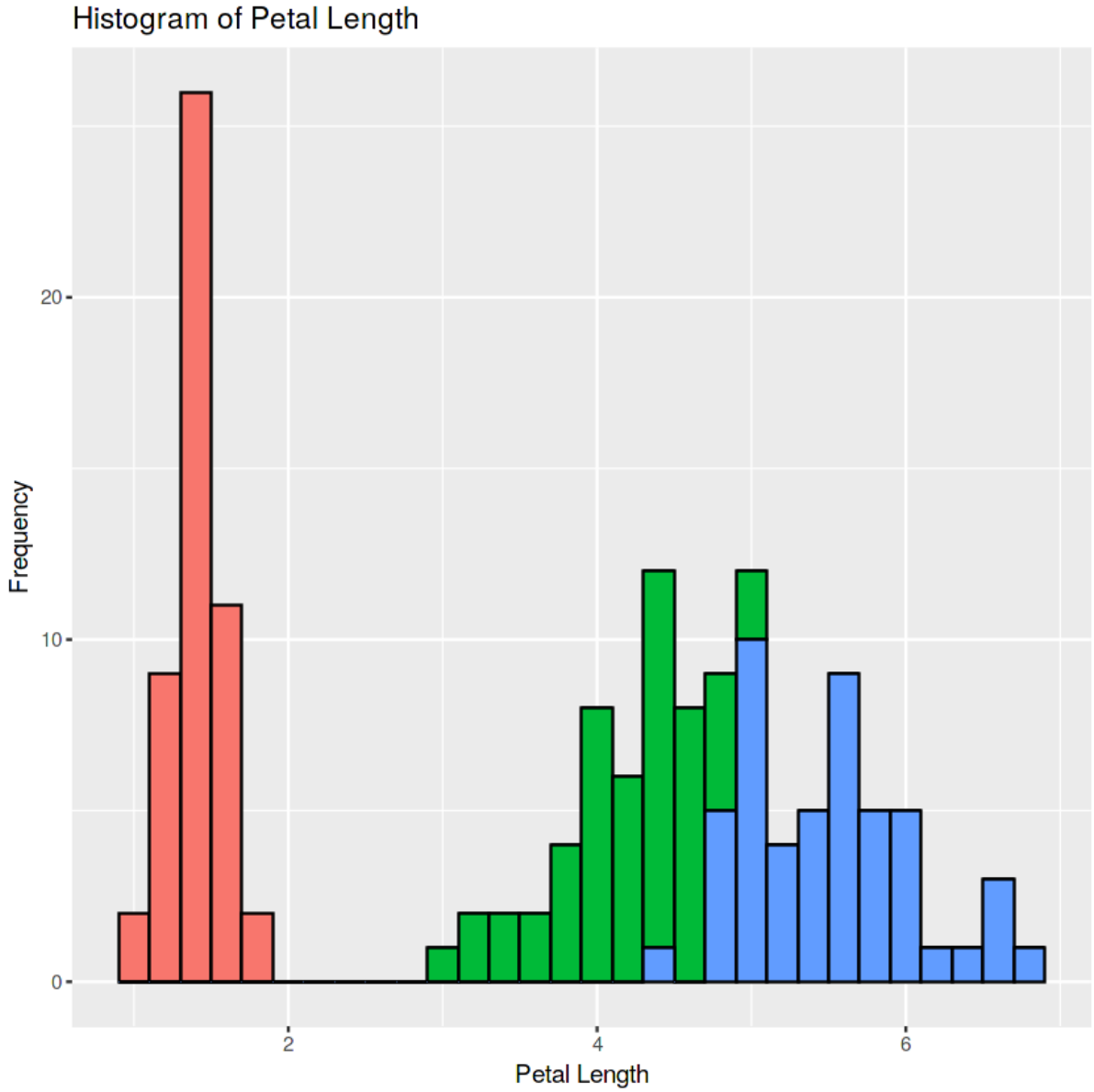
Observations:

- Petal Width of Setosa is the smallest of all three.
- Virginica has the largest petal width.
- There are outliers in Setosa.
- Petal Width is left-skewed.

Histogram

You can use the R code below to draw a histogram for Petal Length to find out the number of points that exist for each bin (range of values):

```
ggplot(data=iris, aes(x=Petal.Length))+
  geom_histogram(binwidth=0.2, color="black", aes(fill=Species))+
  xlab("Petal Length") +
  ylab("Frequency") +
  ggtitle("Histogram of Petal Length")
```



Observations:

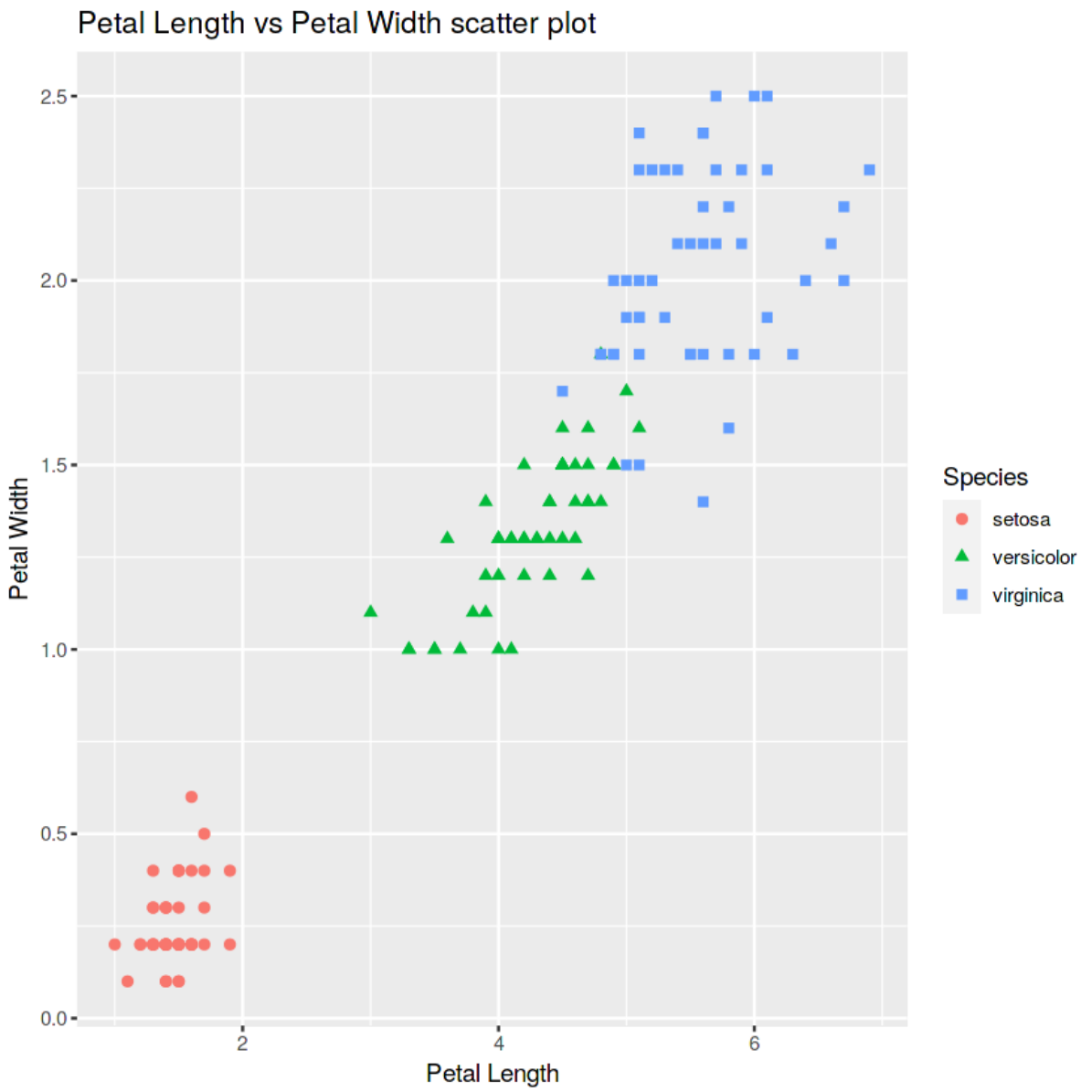
The bars in the above graph compose a histogram. The observations drawn here are the same as the ones we drew from the histogram in Python:

- Setosa is easily separable on the basis of Petal Length.
- There is an overlap between Versicolor and Virginia.
- Distributions are Uniform/Gaussian distribution.

Scatter Plots

A scatter plot is a plot that shows the relationship between two variables of a data set. You can draw a scatter plot between Petal Length and Petal Width for all three species in R with this code:

```
ggplot(data = iris, aes(x = Petal.Length, y = Petal.Width))+
  xlab("Petal Length")+
  ylab("Petal Width") +
  geom_point(aes(color = Species,shape=Species),size = 2)+
  ggtitle("Petal Length vs Petal Width scatter plot")
```



Observations:

- Setosa (red) is easily differentiable
- Versicolor and virginica overlap slightly in Petal Length and Petal Width. These two can almost distinguish using Petal Length and Petal Width

Picturing the possibilities of data visualizations

In our modern world of Big Data, data visualizations are necessary. They can literally give direction and a vision to data scientists and frontline business users alike.

Understanding the Metrics across domains

Introduction

The basic idea of building a machine learning model is to assess the relationship between the dependent and independent variables. In doing so, we need to optimize the model performance. There are two types of ML models, classification and regression; for each ML model, we need to optimize for different parameters. Evaluation metrics used for classification problems differ from regression problems. We will go through most of the classification and regression evaluation metrics with python code to implement them.

Classification Metrics

Classification models have various evaluation metrics to gauge the model's performance. Commonly used metrics are Accuracy, Precision, Recall, F1 Score, Log loss, etc. It is worth noting that not all metrics can be used for all situations. For example, Accuracy cannot be used when dealing with imbalanced classification. Before diving deep into classification metrics, it is essential to know the Confusion Matrix in detail as it is the bedrock of most of the metrics that we will discuss.

Confusion Matrix

Confusion Matrix is an (n*n) matrix that measures the predictions of the classification model against the actual values. In the case of binary classification, the confusion matrix becomes a 2*2 matrix; the size of the matrix depends on the number of classes in the dependent variable. A typical Confusion matrix looks like below,

		Actual Values	
		1	0
Predicted Values	1	True Positive (TP)	False Positive (FP)

	0	False Negative (FN)	True Negative (TN)

Some of the terms mentioned in the above confusion matrix are defined as follows,

1. **True Positives**: When the actual class is positive and the model predicts a positive course, it is termed True Positive.

2. **True Negative**: When the actual class is negative, and the model predicts a negative type, it is True Negative.

3. **False Positive:** When the actual class is negative, and the model predicts a positive course, it is False Positive. One can think of it as the model falsely indicating a positive class when it is negative. False Positives are also known as **Type 1 errors**. For example, minimizing False Positives becomes essential in the Banking industry; if a customer is falsely predicted as a loan defaulter and the customer did not default, it is a loss of revenue to the bank.

4. **False Negative:** When the actual class is positive, and the model predicts a harmful category, it is False Negative. One can think of it as the model falsely predicting a negative course when the class is positive. False Negatives are also known as **Type 2 errors**. For example, minimizing the False Negatives becomes very important in the medical field; if a cancerous patient is diagnosed as non-cancerous, it can be fatal.

One should note that the aim of the build model should be to maximize the True Positives and True Negatives and minimize the False Positives and False Negatives. Now that we know the basic terminologies of a confusion matrix, we can look at the evaluation metrics derived from the confusion matrix.

Accuracy:

Accuracy is one of the most used metrics to evaluate model performance. It describes how accurate your model is. Mathematically, it is the ratio of the sum of

True Positives and Negatives to the total number of data points. From the Confusion matrix, it can be derived as follows,

Accuracy = (TP + TN) / (TP+ TN+ FP+ FN)

There are a few things to note about Accuracy as an evaluation metric,

- Accuracy is a good metric when the classes in the dependent variable are balanced between positive and negative types.
- Accuracy is easy to calculate and easy to understand as well.
- High Accuracy in the case of imbalanced class distribution can lead to misleading results since the model might always predict the dominant class and might not predict the minor class.

Error Rate / Misclassification Rate:

Error rate or Misclassification rate is the exact opposite of Accuracy. It measures how inaccurate your model is. Mathematically, it is the ratio of the sum of False Positives and False Negatives to the total number of data points. It can also be calculated as 1-Accuracy.

Error Rate = (FP+FN) / (TP+TN+FP+FN) = 1 – Accuracy

True Positive Rate / Sensitivity / Recall:

Sensitivity measures how sensitive your model is. The model can correctly classify positive values. In simple terms, when the actual class is True or 1 or yes, how often does the model predict True or 1 or yes. Mathematically, it is the ratio of True Positives to Actual Positives. Sensitivity is an essential metric in the medical industry. If the model can predict a diseased individual as diseased, it is beneficial to the patient; the more correct predictions the model makes, the better it is.

Sensitivity = (TP) / (TP+FN)

False Positive Rate:

False Positive Rate measures the misclassifications. In simple terms, when the actual class is False or 0 or no, how often does the model predict True or 1 or yes. For example, if a patient is falsely expected as having a disease when he does not, it does not matter much as there are always False Positives that turn out in medical tests. Further tests can be conducted, and correct predictions can be obtained. Mathematically, FPR is the ratio of False Positives to the sum of True Negatives and False Positives.

FPR = (FP) / (FP+TN)

Factual Negative Rate / Specificity:

TNR or Specificity measures how specific our model is. If the model predicts all healthy individuals as not having a particular disease, the model is said to be highly specific. In simple terms, when it is No or 0 or False, how often does the model predict No or 0 or False. Mathematically, it is the ratio of True Negatives by the sum of True Negatives and False Positives.

Specificity = (TN) / (TN+FP)

Precision:

Precision measures how precise or accurate the prediction of your model is. In simple terms, when the model predicts True or Yes or 1, how often is the prediction correct? For example, when indicating fraudulent transactions, it is essential to predict trades as fraudulent correctly. If you expect non-fraudulent transactions as fraudulent, it can lead to business loss. Mathematically, it is the ratio of True Positives to the sum of True Positives and False Positives.

Precision = (TP) / (TP+FP)

F Beta Score / F1 Score:

F Beta score considers both precision and recall. There are instances where we need the model to e optimized for both precision and recall metrics. In such cases, the F Beta score is used as the metric. It is given by the equation below,

F Beta = (1+Beta^2) * ((Precision*Recall) / (Beta^2*Precision + Recall))

Another vital evaluation metric is the F1 Score. We all know it as the Harmonic mean of precision and recall metrics, and it is derived from the above equation by substituting Beta = 1. When we substitute Beta with 1, we give equal importance to both Precision and Recall metrics.

F1 Score = (2*Precision*Recall) / (Precision + Recall)

Another essential thing to note about the F1 Score is that it depends on TPR and FPR now, and these values can be altered by altering the threshold of the classifier. For example, for the default threshold of 0.5, there are specific TPR and FPR; if you alter the threshold value, the TPR and FPR change and hence the value of the F1 Score changes.

Log loss:

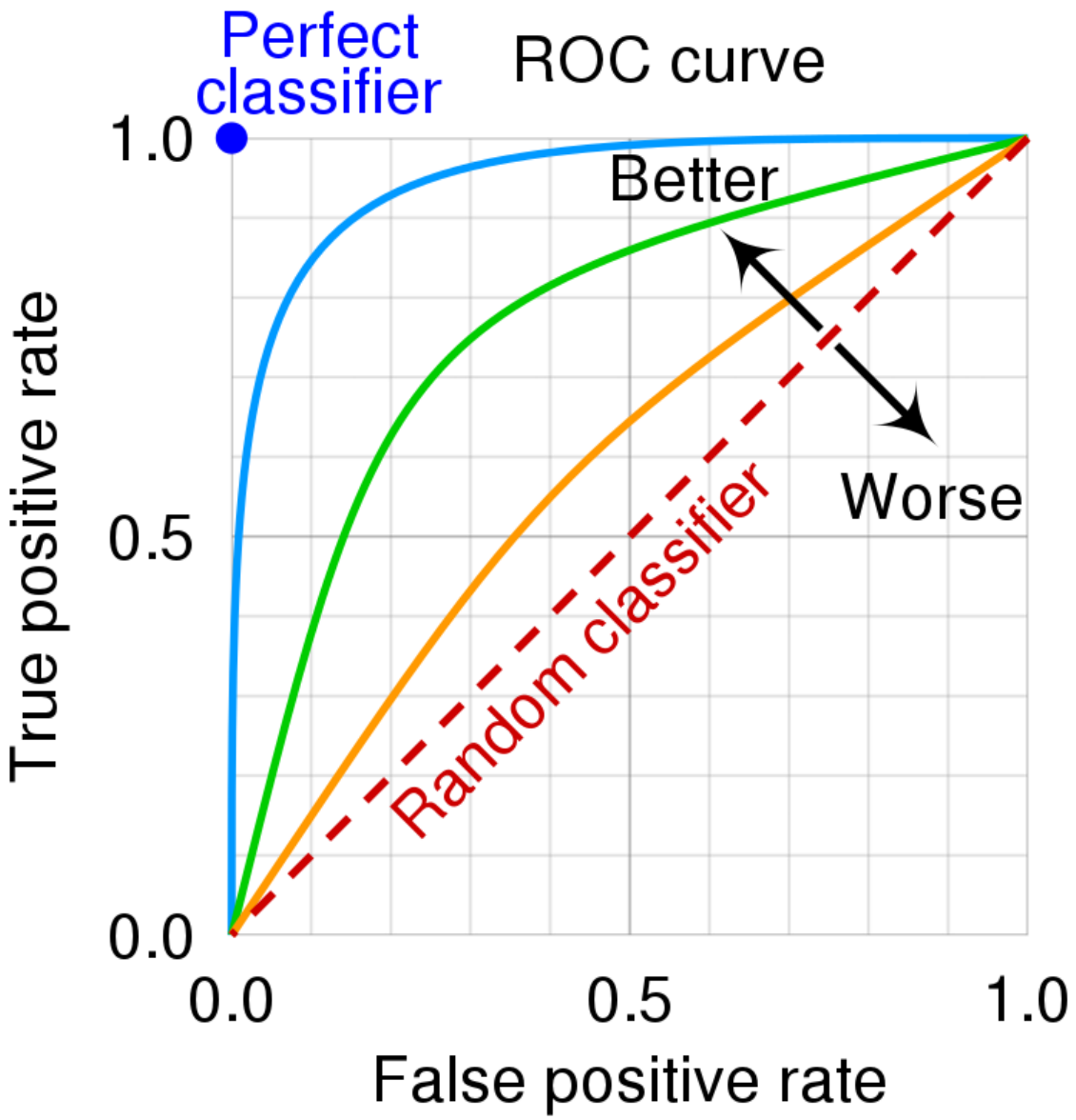
Log loss is a vital evaluation metric used to compare the performance of two classification models. Lower the log loss, better is the model in short. Log loss penalizes the false classifications. ***If the model assigns a lower probability to the correct class for a particular data point, then the log loss of the corresponding data point will be significantly significant. Similarly, if the model gives a higher probability to the incorrect class, the log loss will be higher***. So basically, the higher is the probability assigned to the correct class, the lower is the log loss. Log loss for a binary classification problem is given by the formula shown below,

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i \log, p_i + (1-y_i) \log, (1-p_i)].$$

Source: https://www.analyticsvidhya.com/blog/2020/11/binary-cross-entropy-aka-log-loss-the-cost-function-used-in-logistic-regression/

ROC Curve / Area Under the ROC Curve (ROC AUC Score):

Before diving into what the ROC curve is, it is essential to note that most of the Machine Learning models do not predict the class labels directly; they consistently indicate the probability of a data point belonging to either positive or negative class, then based on the threshold value (which is .5 by default), the labels are classified as belonging to either positive or negative type. ROC Curve stands for Receiver Operating Characteristic Curve. ***The curve checks how the observations change classes based on the variations in the threshold value; it visualizes the tradeoff between TPR and FPR rates***. It is a graph of True Positive Rate or Sensitivity on the Y-axis and False Positive Rate or (1-Specificity) on the X-axis. It is worth noting that if your model is better, the curve hugs the Y-axis as much as possible. A typical ROC curve looks like the figure below,



Source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

One can also note that practically, the curves are not as perfect as shown in the figure above. Likely, they will be very different from the above figure.

One more critical metric that can be calculated from the ROC curve is the Area Under the ROC Curve; the higher the AUC score value, the better the model performance. If the ROC curve hugs the Y-axis or is closer to the Y-axis, the AUC score will be higher. The maximum possible value of the AUC score is 1, which is practically not possible as the model cannot predict all observations correctly.

Regression Metrics:

We saw how to evaluate the performance of a classifier till now. We will now deep dive into evaluating the performance of a Regression model where we predict continuous values and not individual classes. The Regression Evaluation metrics differ from classification evaluation metrics, and the most popular ones are MAE, MSE, RMSE, R Squared, etc.

Mean Absolute Error:

The term Error in "Mean Absolute Error" stands for the difference between the actual and the predicted values of the continuous variable. When we predict a constant variable, some of the values predicted can be below the actual value, and some can be above the actual value. If you consider the sum of the differences between actual and predicted, some values may cancel out, which is why we take the absolute value of the difference between actual and predicted. This cancels out any negative values, and it is the average fundamental value of the differences between actual and predicted values.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Some important points to note about the MAE as an evaluation metric are,

- The MAE is in the same units as the continuous dependent variable.
- It is easy to interpret MAE.
- Although it takes the absolute value of the differences, it does not highlight the extreme values of the differences, i.e., it does not highlight predictions that are way off the accepted range.
- It is not sensitive to outliers.

Mean Squared Error:

Like the MAE metric, MSE measures the differences between the actual values and the predictions. It takes the square of the differences between accurate predictions instead of the absolute values. Mathematically, it is the average squared differences between actual and predictions.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Some important points to note about MSE are,

- MSE is sensitive to outliers as we take the square of the differences, which highlights extreme values.
- It is not easy to interpret as it does not have the same units as the continuous dependent variable.
- It is better than MAE.

Root Mean Squared Error:

RMSE is almost the same as MSE, except it takes the square root of the Mean Squared Error. It is the most popular evaluation metric, and it overcomes any drawbacks that MAE or MSE have. Mathematically it is the square root of the average of the squared difference between actual values and predictions.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Some important points about RMSE are,

- It is sensitive to outliers as it takes the squared differences.

- It has the same units as that of the continuous dependent variable.
- It is easier to interpret as compared to MSE.

Root Mean Squared Log Error:

RMSLE is almost the same as RMSE except that it takes the log values of the actual and predicted values instead of using them as is. It also adds 1 to the weights if the value is 0 as the log of 0 is not defined. It is also not valid as a metric when negative values are involved.

Some important points to note about RMSLE are,

- It is used when the target variable has an extensive range.
- It is also used when we look at the growth over the years or when there is exponential growth.
- It is also used to know the percentage error as the expression within evaluated as a ratio as well (applying logarithmic rules).

MAPE:

MAPE stands for Mean Absolute Percentage Error. It can be interpreted as the average of the absolute percentage of errors. It is mainly used as an evaluation metric in forecasting problems where we need to determine the values in the future. It is calculated as the absolute value of the ratio of the difference between actuals and predicted to the actual values. It is given by the formula below,

Some of the critical points to note about MAPE are,

- It cannot be used when the actual values are 0 because you cannot divide something by zero.
- A good MAPE value is always less than 10%.

R Squared:

R Squared is unlike the regression metric that we discussed above. It considers the predictions of our model and the average value of our dependent variable.

There are two ways in which R Squared can be interpreted; one ***is that it describes how much variation in the dependent variable is explained by our current set of independent variables,*** the other interpretation is that ***it tells how better our current model is as compared to just predicting the average value of the dependent variable for all data points.*** R Squared is defined by the formula given below,

As we can see above, in the numerator of the formula, we subtract the predicted values of our model from the actual values and square them. In the denominator, we remove the average value of our dependent value from the fundamental values and square it.

Some of the essential points to note about R Squared are,

- The higher the R Squared, the better the model (although the Adjusted R contests this Squared metric below).
- As we keep adding variables into the model, R Squared increases, and it does not decrease. As we add more variables, although the denominator remains the same, the numerator reduces, and hence the R squared keeps increasing.
- R Squared ranges between 0 and 1.

Adjusted R Squared:

Adjusted R Squared is the modified version of R Squared. It considers the number of data points, the number of predictors in the model, the R Squared value. Adjusted R Squared is a more reliable metric than R Squared because it mainly considers the number of predictors. ***If we keep adding variables to our model that do not contribute to the model's performance, the Adjusted R Squared does not increase; only if the variable added is contributing to improving the model, the Adjusted R Squared value increase.***

Some important points about Adjusted R Squared are,

- Adjusted R Squared might decrease, unlike R Squared.
- Adjusted R Squared is less than or equal to R Squared.
- It penalizes for adding more insignificant variables to our model.

End Notes:

Although many other evaluation metrics are used for classification and regression problems, the ones explained in the article are the most used.

Developing Metrics - Flowchart for Metric Creation

The three stages of my journey to understanding monitoring (so far) are:

- Stage 1: What? (Looks elsewhere)
- Stage 2: Without metrics, we are really flying blind.
- Stage 3: How do we keep from doing metrics wrong?

I am currently in Stage 2 and will share what I have learned so far. I'm moving gradually toward Stage 3, and I will offer some of my resources on that part of the journey at the end of this article.

Why should I monitor?

The top reasons for monitoring are:

- Understanding *normal* and *abnormal* system and service behavior
- Doing capacity planning, scaling up or down
- Assisting in performance troubleshooting
- Understanding the effect of software/hardware changes
- Changing system behavior in response to a measurement
- Alerting when a system exhibits unexpected behavior

Metrics and metric types

For our purposes, a **metric** is an *observed* value of a certain quantity at a given point in *time*. The total of number hits on a blog post, the total number of people attending a talk, the number of times the data was not found in the caching system, the number of logged-in users on your website—all are examples of metrics.

They broadly fall into three categories:

Counters

Consider your personal blog. You just published a post and want to keep an eye on how many hits it gets over time, a number that can only increase. This is an example of a **counter** metric. Its value starts at 0 and increases during the lifetime of your blog post. Graphically, a counter looks like this:

opensource.com

Gauges

Instead of the total number of hits on your blog post over time, let's say you want to track the number of hits per day or per week. This metric is called a **gauge** and its value can go up or down. Graphically, a gauge looks like this:

opensource.com

A gauge's value usually has a *ceiling* and a *floor* in a certain time window.

Histograms and timers

A **histogram** (as Prometheus calls it) or a **timer** (as StatsD calls it) is a metric to track *sampled observations*. Unlike a counter or a gauge, the value of a histogram metric doesn't necessarily show an up or down pattern. I know that doesn't make a lot of sense and may not seem different from a gauge. What's different is what you expect to *do* with histogram data compared to a gauge. Therefore, the monitoring system needs to know that a metric is a histogram type to allow you to do those things.

opensource.com

Demo 1: Calculating and reporting metrics

Demo 1 is a basic web application written using the Flask framework. It demonstrates how we can *calculate* and *report* metrics.

The src directory has the application in app.py with the src/helpers/middleware.py containing the following:

```
from flask import request
import csv
import time


def start_timer():
    request.start_time = time.time()


def stop_timer(response):
    # convert this into milliseconds for statsd
    resp_time = (time.time() - request.start_time)*1000
    with open('metrics.csv', 'a', newline='') as f:
        csvwriter = csv.writer(f)
        csvwriter.writerow([str(int(time.time())), str(resp_time)])

    return response


def setup_metrics(app):
    app.before_request(start_timer)
    app.after_request(stop_timer)
```

When setup_metrics() is called from the application, it configures the start_timer() function to be called before a request is processed and the stop_timer() function to be called after a request is processed but before the response has been sent. In the above function, we write the timestamp and the time it took (in milliseconds) for the request to be processed.

When we run docker-compose up in the demo1 directory, it starts the web application, then a client container that makes a number of requests to the web application. You will see a src/metrics.csv file that has been created with two columns: timestamp and request_latency.

Looking at this file, we can infer two things:

- There is a lot of data that has been generated
- No observation of the metric has any characteristic associated with it

Without a characteristic associated with a metric observation, we cannot say which HTTP endpoint this metric was associated with or which node of the application this metric was generated from. Hence, we need to qualify each metric observation with the appropriate metadata.

Statistics 101

If we think back to high school mathematics, there are a few statistics terms we should all recall, even if vaguely, including mean, median, percentile, and histogram. Let's briefly recap them without judging their usefulness, just like in high school.

Mean

The **mean**, or the average of a list of numbers, is the sum of the numbers divided by the cardinality of the list. The mean of 3, 2, and 10 is (3+2+10)/3 = 5.

Median

The **median** is another type of average, but it is calculated differently; it is the center numeral in a list of numbers ordered from smallest to largest (or vice

versa). In our list above (2, 3, 10), the median is 3. The calculation is not very straightforward; it depends on the number of items in the list.

Percentile

The **percentile** is a measure that gives us a measure below which a certain (k) percentage of the numbers lie. In some sense, it gives us an *idea* of how this measure is doing relative to the k percentage of our data. For example, the 95th percentile score of the above list is 9.29999. The percentile measure varies from 0 to 100 (non-inclusive). The *zeroth* percentile is the minimum score in a set of numbers. Some of you may recall that the median is the 50th percentile, which turns out to be 3.

Some monitoring systems refer to the percentile measure as upper_X where *X* is the percentile; *upper 90* refers to the value at the 90th percentile.

Quantile

The **q-Quantile** is a measure that ranks q*N* in a set of *N* numbers. The value of **q** ranges between 0 and 1 (both inclusive). When **q** is 0.5, the value is the median. The relationship between the quantile and percentile is that the measure at **q** quantile is equivalent to the measure at **100*q*** percentile.

Histogram

The metric **histogram**, which we learned about earlier, is an *implementation detail* of monitoring systems. In statistics, a histogram is a graph that groups data into *buckets*. Let's consider a different, contrived example: the ages of people reading your blog. If you got a handful of this data and wanted a rough idea of your readers' ages by group, plotting a histogram would show you a graph like this:

opensource.com

Cumulative histogram

A **cumulative histogram** is a histogram where each bucket's count includes the count of the previous bucket, hence the name *cumulative*. A cumulative histogram for the above dataset would look like this:

opensource.com

Why do we need statistics?

In Demo 1 above, we observed that there is a lot of data that is generated when we report metrics. We need statistics when working with metrics because there are just too many of them. We don't care about individual values, rather overall behavior. We expect the behavior the values exhibit is a proxy of the behavior of the system under observation.

Demo 2: Adding characteristics to metrics

In our Demo 1 application above, when we calculate and report a request latency, it refers to a specific request uniquely identified by few *characteristics*. Some of these are:

- The HTTP endpoint
- The HTTP method
- The identifier of the host/node where it's running

If we attach these characteristics to a metric observation, we have more context around each metric. Let's explore adding characteristics to our metrics in Demo 2.

The src/helpers/middleware.py file now writes multiple columns to the CSV file when writing metrics:

node_ids = ['10.0.1.1', '10.1.3.4']

def start_timer():

```
    request.start_time = time.time()



def stop_timer(response):
    # convert this into milliseconds for statsd
    resp_time = (time.time() - request.start_time)*1000
    node_id = node_ids[random.choice(range(len(node_ids)))]
    with open('metrics.csv', 'a', newline='') as f:
        csvwriter = csv.writer(f)
        csvwriter.writerow([
            str(int(time.time())), 'webapp1', node_id,
            request.endpoint, request.method, str(response.status_code),
            str(resp_time)
        ])

    return response
```

Since this is a demo, I have taken the liberty of reporting random IPs as the node IDs when reporting the metric. When we run docker-compose up in the demo2 directory, it will result in a CSV file with multiple columns.

Analyzing metrics with pandas

We'll now analyze this CSV file with pandas. Running docker-compose up will print a URL that we will use to open a Jupyter session. Once we upload the Analysis.ipynb notebook into the session, we can read the CSV file into a pandas DataFrame:

```
import pandas as pd
metrics = pd.read_csv('/data/metrics.csv', index_col=0)
```

The index_col specifies that we want to use the timestamp as the index.

Since each characteristic we add is a column in the DataFrame, we can perform grouping and aggregation based on these columns:

```
import numpy as np
metrics.groupby(['node_id', 'http_status']).latency.aggregate(np.percentile, 99.999)
```

Please refer to the Jupyter notebook for more example analysis on the data.

What should I monitor?

A software system has a number of variables whose values change during its lifetime. The software is running in some sort of an operating system, and operating system variables change as well. In my opinion, the more data you have, the better it is when something goes wrong.

Key operating system metrics I recommend monitoring are:

- CPU usage
- System memory usage
- File descriptor usage
- Disk usage

Other key metrics to monitor will vary depending on your software application.

Network applications

If your software is a network application that listens to and serves client requests, the key metrics to measure are:

- Number of requests coming in (counter)
- Unhandled errors (counter)
- Request latency (histogram/timer)
- Queued time, if there is a queue in your application (histogram/timer)
- Queue size, if there is a queue in your application (gauge)
- Worker processes/threads usage (gauge)

If your network application makes requests to other services in the context of fulfilling a client request, it should have metrics to record the behavior of communications with those services. Key metrics to monitor include number of requests, request latency, and response status.

HTTP web application backends

HTTP applications should monitor all the above. In addition, they should keep granular data about the count of non-200 HTTP statuses grouped by all the other HTTP status codes. If your web application has user signup and login functionality, it should have metrics for those as well.

Long-running processes

Long-running processes such as Rabbit MQ consumer or task-queue workers, although not network servers, work on the model of picking up a task and processing it. Hence, we should monitor the number of requests processed and the request latency for those processes.

No matter the application type, each metric should have appropriate **metadata** associated with it.

Integrating monitoring in a Python application

There are two components involved in integrating monitoring into Python applications:

- Updating your application to calculate and report metrics
- Setting up a monitoring infrastructure to house the application's metrics and allow queries to be made against them

The basic idea of recording and reporting a metric is:

```
def work():
    requests += 1
```

```
# report counter
start_time = time.time()

# < do the work >

# calculate and report latency
work_latency = time.time() - start_time
...
```

Considering the above pattern, we often take advantage of *decorators*, *context managers*, and *middleware* (for network applications) to calculate and report metrics. In Demo 1 and Demo 2, we used decorators in a Flask application.

Pull and push models for metric reporting

Essentially, there are two patterns for reporting metrics from a Python application. In the *pull* model, the monitoring system "scrapes" the application at a predefined HTTP endpoint. In the *push* model, the application sends the data to the monitoring system.

opensource.com

An example of a monitoring system working in the *pull* model is Prometheus. StatsD is an example of a monitoring system where the application *pushes* the metrics to the system.

Integrating StatsD

To integrate StatsD into a Python application, we would use the StatsD Python client, then update our metric-reporting code to push data into StatsD using the appropriate library calls.

First, we need to create a client instance:

```
statsd = statsd.StatsClient(host='statsd', port=8125, prefix='webapp1')
```

The prefix keyword argument will add the specified prefix to all the metrics reported via this client.

Once we have the client, we can report a value for a timer using:

statsd.timing(key, resp_time)

To increment a counter:

statsd.incr(key)

To associate metadata with a metric, a key is defined as metadata1.metadata2.metric, where each metadataX is a field that allows aggregation and grouping.

The demo application StatsD is a complete example of integrating a Python Flask application with statsd.

Integrating Prometheus

To use the Prometheus monitoring system, we will use the Promethius Python client. We will first create objects of the appropriate metric class:

```
REQUEST_LATENCY = Histogram('request_latency_seconds', 'Request latency',
    ['app_name', 'endpoint']
)
```

The third argument in the above statement is the labels associated with the metric. These labels are what defines the metadata associated with a single metric value.

To record a specific metric observation:

```
REQUEST_LATENCY.labels('webapp', request.path).observe(resp_time)
```

The next step is to define an HTTP endpoint in our application that Prometheus can scrape. This is usually an endpoint called /metrics:

```
@app.route('/metrics')
def metrics():
```

return Response(prometheus_client.generate_latest(), mimetype=CONTENT_TYPE_LATEST)

The demo application Prometheus is a complete example of integrating a Python Flask application with prometheus.

Which is better: StatsD or Prometheus?

The natural next question is: Should I use StatsD or Prometheus? I have written a few articles on this topic, and you may find them useful:

- Your options for monitoring multi-process Python applications with Prometheus
- Monitoring your synchronous Python web applications using Prometheus
- Monitoring your asynchronous Python web applications using Prometheus

Ways to use metrics

We've learned a bit about why we want to set up monitoring in our applications, but now let's look deeper into two of them: alerting and autoscaling.

Using metrics for alerting

A key use of metrics is creating alerts. For example, you may want to send an email or pager notification to relevant people if the number of HTTP 500s over the past five minutes increases. What we use for setting up alerts depends on our monitoring setup. For Prometheus we can use Alertmanager and for StatsD, we use Nagios.

Using metrics for autoscaling

Not only can metrics allow us to understand if our current infrastructure is over- or under-provisioned, they can also help implement autoscaling policies in a cloud infrastructure. For example, if worker process usage on our servers routinely hits 90% over the past five minutes, we may need to horizontally scale. How we would implement scaling depends on the cloud infrastructure. AWS Auto Scaling, by

default, allows scaling policies based on system CPU usage, network traffic, and other factors. However, to use application metrics for scaling up or down, we must publish custom CloudWatch metrics.

Application monitoring in a multi-service architecture

When we go beyond a single application architecture, such that a client request can trigger calls to multiple services before a response is sent back, we need more from our metrics. We need a unified view of latency metrics so we can see how much time each service took to respond to the request. This is enabled with distributed tracing.

You can see an example of distributed tracing in Python in my blog post Introducing distributed tracing in your Python application via Zipkin.

Points to remember

In summary, make sure to keep the following things in mind:

- Understand what a metric type means in your monitoring system
- Know in what unit of measurement the monitoring system wants your data
- Monitor the most critical components of your application
- Monitor the behavior of your application in its most critical stages

Process Flow Metrics- In Six Sigma, you want to define a process very precisely — down to the last detail of activity, resource, decision, dependency, and value. Sometimes, this level of definition is the only way you can sufficiently measure and analyze a process, leading to breakthrough improvements and, ultimately, effective controls.

Mapping or modelling the process is a representation of this precise process definition, and the practice of process modelling is therefore fundamental to Six Sigma. A *process map* looks like a flowchart, and, at the top level, that's exactly what it is. A process map is a picture of the activities and events in a process.

Six Sigma process mapping begins with building flowcharts. You then annotate and define the paths, encounters, decisions, and destinations on these charts in quantitative terms, including such measures as value, time, resources, yields, and the statistical distributions around each.

How to draw a Six Sigma process map?

Process mapping has been practiced for decades. The Six Sigma style of process mapping has a few different aspects, however, so you utilize a few new features from its flowchart ancestors.

- In the world of Six Sigma, you characterize the process map in mathematical terms so you can perform a plethora of statistical analyses on its various parts and pieces. You back each step, function, and activity with numerical descriptions and quantifiable attributes, enabling you to see the process in all its exacting glory.
- In Six Sigma process mapping, you characterize a practical situation in ways that let you describe it in statistical terms, allowing you to develop statistical solutions that you then apply back into your practical environment.
- As you begin, don't worry about the details of what happens inside each of these boxes. Your goal at this stage is to capture each of the steps, identify its basic function, and connect all the steps in the manner that represents the existing process.

Define and visualize the process points

After the process map is drafted, the next step is to define each of the map's objects. One must be precise and quantitative; the accuracy of your process model depends on it. If a process mapping technology tool is being used, the tool includes prompts for the numerous definitions and attributes at each *node* (step) in the map. The categories of process element definitions include the following:

- **Operation cycle time of the process element,** including its average time to complete, the variation in time called the *standard deviation,* and perhaps a distribution curve to represent all the possible completion times as well.
- **Resources used in the process element,** including human, capital, and natural resources. The better tools lets one identify resources by name and type and then later track their utilization during simulation.
- **Value added by the process step,** in the units of measure that mean the most to the organization. At a minimum, you must be able to define whether the process step is value add (VA) or non-value add (NVA).
- **Costs of the resources consumed,** including the costs of personnel, facilities, direct material, and sometimes even indirect costs.

A common practice in process modeling is to employ a visualization technique called *swim lanes.*Processes cross functional boundaries and borders, and swim lanes helps you see that movement.

In a swim lane process map, time flows from left to right; the process crosses lanes as it traverses departments on its journey from start to finish. Imagine you're the customer in the process map: You're in lane 1. As you work your way through customer service and then the banking services, you cross over into lanes 2 and 3.

Swim lanes are an effective visualization technique that lets each functional contributor to a process understand his role while giving everyone a chance to see just how complicated the process may be within your organization. Remember, each time you cross a lane, you have in essence created a supplier-customer interaction that implies needs, wants, and desires that must be met.

Acknowledge the as-is state

Process Flow Metrics- One way to think about process mapping is as an exercise in defining a better process — how you envision your process can work sometime in the future, after implementing the changes that would enable your new concepts. It's the *to-be* state of affairs. And mapping the future in this way provides you the opportunity to examine your plans in detail and consider your options before implementing the changes.

Firstly, a map of today's reality needs to be created: the *as-is* state. Many organizations skip this kind of mapping. The only excuse for not modelling the as-is process is if something brand new is being implemented. Otherwise, if a process exists today, model it first. Doing so accomplishes three important tasks:

- **Sets the baseline:** Before you can measure the effects of your sweeping changes, you must first characterize the present conditions. By using the same process mapping techniques to characterize today's as-is state as well as the future to-be state, you have the basis for measuring the effectiveness of your process improvement effort.
- **Sees the process:** Seeing the process involves recognizing that three different perspectives exist:
    - What you think is going on
    - What's really going on
    - What should be going on

These views are distinctly different, but characterizing the differences is precisely what you're trying to achieve with Six Sigma. You must replace what you think is going on with what is really going on, and only then can you understand what moving to the third view requires.

- **Stimulates closed-loop behaviour:** Your investment in mapping primes the pump for breakthrough performance improvement. To continue the

cycle of improvement, your model should be a dynamic, living entity such that your model is in sync with reality at any point in time. Modelling the as-is condition from the beginning promotes this closed-loop behaviour.

Understanding metrics across domains

Understanding metrics across domains in business analytics is crucial for evaluating performance, identifying trends, and making data-driven decisions in various areas of a business. Whether you're analyzing sales, marketing, finance, or any other aspect of an organization, the following principles can help you navigate and understand metrics effectively:

1. **Domain Expertise:**
   - Gain a deep understanding of the specific domain you're analyzing. This includes understanding industry standards, key performance indicators (KPIs), and the unique challenges and opportunities in that domain.

2. **Define Clear Objectives:**
   - Clearly define your business objectives within the domain. What are you trying to achieve? What questions are you seeking to answer through analytics?

3. **Select Relevant Metrics:**
   - Identify and select the most relevant metrics that align with your objectives. Metrics should directly measure the aspects of the business that matter most in the given domain.

4. **Data Collection and Quality:**

- Ensure that data collection methods are robust and reliable. High-quality data is essential for accurate analysis and meaningful insights.

5. **Benchmarking:**
- Benchmark your metrics against industry standards and competitors to gauge performance relative to peers.

6. **Time Frame Consideration:**
- Consider the appropriate time frame for your metrics. Some metrics may be short-term (daily or weekly), while others are more suitable for long-term tracking.

7. **Segmentation:**
- Use segmentation to break down data into smaller, more manageable groups. This can help identify trends or issues specific to certain customer segments, products, or regions.

8. **Visualization:**
- Visualize your data using charts, graphs, and dashboards. Visual representations make it easier to spot patterns and outliers.

9. **Trend Analysis:**
- Analyze trends over time to understand how metrics are changing. Are they improving, declining, or remaining stable? What factors might be influencing these trends?

10. **Correlation vs. Causation:**

- Be cautious when interpreting correlations. Just because two metrics move together does not necessarily mean one causes the other. Dig deeper to establish causation when needed.

11. **Predictive Analytics:**
- In some cases, use predictive analytics to forecast future outcomes based on historical data and trends.

12. **A/B Testing:**
- For marketing and product-related metrics, consider A/B testing to assess the impact of changes and improvements.

13. **Cost-Benefit Analysis:**
- Assess the costs associated with collecting and analyzing data against the benefits gained from insights and improvements.

14. **Ethical Considerations:**
- Be aware of ethical considerations, including data privacy, transparency, and bias. Ensure compliance with data protection regulations.

15. **Communication and Actionability:**
- Present your findings in a clear and actionable way to stakeholders. Insights should lead to decisions and actions that drive improvement.

16. **Continuous Monitoring:**
- Continuously monitor your chosen metrics to track progress toward your objectives. Be prepared to adapt your strategies based on changing data.

17. **Cross-Functional Collaboration:**
- Encourage collaboration between different departments and teams within the organization. Metrics often overlap multiple domains, and a holistic approach is valuable.

18. **Education and Training:**
- Ensure that employees involved in analytics have the necessary skills and training to understand and work with metrics effectively.

19. **Feedback and Iteration:**
- Encourage feedback from stakeholders and analysts to iterate and refine the metrics and analysis methodologies as needed.

Understanding metrics across domains in business analytics requires a combination of analytical skills, domain-specific knowledge, and a commitment to using data to drive informed decisions. By following these principles, you can navigate the complex world of business analytics and extract valuable insights that can help your organization succeed.

Developing metrics across domains

Developing metrics across domains in business analytics involves creating a set of key performance indicators (KPIs) and measurement tools that are tailored to the specific needs and objectives of your organization, regardless of the domain. Here's a step-by-step guide on how to develop metrics effectively:

1. **Understand Business Goals:**

- Start by gaining a clear understanding of your organization's overarching business goals and objectives. These should guide the development of metrics in any domain.

2. **Identify Domains:**
   - Determine the specific domains within your organization that require metrics. This could include sales, marketing, finance, operations, customer service, and more.

3. **Engage Stakeholders:**
   - Involve relevant stakeholders from each domain in the metric development process. This ensures that you capture the unique perspectives and priorities of each area.

4. **Set SMART Objectives:**
   - Define Specific, Measurable, Achievable, Relevant, and Time-bound (SMART) objectives for each domain. What do you want to achieve? What are the desired outcomes?

5. **Map Key Processes:**
   - Understand the key processes and workflows within each domain. This helps identify points where metrics can provide valuable insights.

6. **Brainstorm Metrics:**
   - Conduct brainstorming sessions with domain experts to generate a list of potential metrics. Encourage creativity and diverse perspectives during this phase.

7. **Prioritize Metrics:**
   - Evaluate the list of potential metrics and prioritize them based on their alignment with objectives and their potential impact on the business.

8. **Quantify Metrics:**
   - Ensure that each metric can be quantified and measured. Metrics should involve numerical values or ratios that reflect performance.

9. **Data Sources and Collection:**
   - Determine where and how you'll collect data for each metric. This may involve existing data sources, surveys, sensors, or new data collection methods.

10. **Baseline Measurement:**
    - Establish a baseline measurement for each metric to provide a starting point for performance evaluation.

11. **Normalization and Weighting:**
    - Normalize metrics if necessary to allow for meaningful comparisons across different domains. Assign weights to metrics to reflect their relative importance.

12. **Test Metrics:**
    - Pilot-test the selected metrics to ensure they are practical, reliable, and capable of delivering valuable insights.

13. **Data Quality Assurance:**

- Implement data quality checks and validation processes to maintain the accuracy and integrity of the data.

14. **Dashboard and Reporting:**
- Develop a reporting system or dashboard that presents the metrics in a user-friendly and visually appealing format for stakeholders.

15. **Feedback and Iteration:**
- Continuously gather feedback from stakeholders and domain experts to iterate and refine the metrics as needed.

16. **Alignment with Strategy:**
- Ensure that the selected metrics align with your organization's overall strategic goals and that they are consistent with the direction in which the business wants to move.

17. **Monitoring and Analysis:**
- Regularly monitor and analyze the metrics to track progress, identify trends, and assess performance.

18. **Cross-Domain Integration:**
- Consider how metrics from different domains may interact and influence one another. Cross-domain insights can lead to more informed decisions.

19. **Training and Education:**
- Provide training to individuals responsible for collecting, analyzing, and interpreting data. Ensure they understand the metrics and their significance.

20. **Ethical Considerations:**
- Address ethical considerations, such as data privacy, consent, and fairness, in the metric development and data collection process.

21. **Communication and Actionability:**
- Communicate the findings and insights derived from the metrics to relevant stakeholders and ensure that the information is actionable.

22. **Review and Adapt:**
- Periodically review and adapt the metrics to reflect changing business conditions, goals, and industry standards.

Developing metrics across domains in business analytics is an ongoing process that requires collaboration, adaptability, and a commitment to data-driven decision-making. By following these steps, you can create a robust set of metrics that help your organization measure and improve performance in various domains.

Flowchart for metric creation

Creating a flowchart for metric creation in business analytics is a useful visual representation of the steps involved in the process. Here's a simplified flowchart outlining the key steps in metric creation:

```mermaid
graph TD
  A[Start] --> B[Understand Business Goals]
  B --> C[Identify Domains]
```

```
C --> D[Engage Stakeholders]
D --> E[Set SMART Objectives]
E --> F[Map Key Processes]
F --> G[Brainstorm Metrics]
G --> H[Prioritize Metrics]
H --> I[Quantify Metrics]
I --> J[Data Sources and Collection]
J --> K[Baseline Measurement]
K --> L[Normalization and Weighting]
L --> M[Test Metrics]
M --> N[Data Quality Assurance]
N --> O[Dashboard and Reporting]
O --> P[Feedback and Iteration]
P --> Q[Alignment with Strategy]
Q --> R[Monitoring and Analysis]
R --> S[Cross-Domain Integration]
S --> T[Training and Education]
T --> U[Ethical Considerations]
U --> V[Communication and Actionability]
V --> W[Review and Adapt]
W --> X[End]
```

Let's briefly explain each step in the flowchart:

1. **Understand Business Goals:** Begin by gaining a clear understanding of your organization's overall business goals and objectives.

2. **Identify Domains:** Determine the specific areas or domains within your organization that require metrics.

3. **Engage Stakeholders:** Involve relevant stakeholders from each domain in the metric creation process.

4. **Set SMART Objectives:** Define Specific, Measurable, Achievable, Relevant, and Time-bound (SMART) objectives for each domain.

5. **Map Key Processes:** Understand the critical processes and workflows within each domain.

6. **Brainstorm Metrics:** Conduct brainstorming sessions to generate a list of potential metrics for each domain.

7. **Prioritize Metrics:** Evaluate and prioritize the metrics based on alignment with objectives and potential impact.

8. **Quantify Metrics:** Ensure that each metric can be quantified and measured with numerical values.

9. **Data Sources and Collection:** Determine how and where data will be collected for each metric.

10. **Baseline Measurement:** Establish a baseline measurement to provide a starting point for performance evaluation.

11. **Normalization and Weighting:** Normalize metrics if necessary and assign weights to reflect their importance.

12. **Test Metrics:** Pilot-test selected metrics to ensure practicality and reliability.

13. **Data Quality Assurance:** Implement data quality checks to maintain data accuracy.

14. **Dashboard and Reporting:** Develop reporting systems or dashboards to present metrics visually.

15. **Feedback and Iteration:** Continuously gather feedback to refine metrics as needed.

16. **Alignment with Strategy:** Ensure metrics align with the organization's strategic goals.

17. **Monitoring and Analysis:** Regularly monitor and analyze metrics to track progress.

18. **Cross-Domain Integration:** Consider how metrics from different domains interact.

19. **Training and Education:** Provide training for data analysts and stakeholders.

20. **Ethical Considerations:** Address ethical considerations in data collection.

21. **Communication and Actionability:** Communicate insights derived from metrics and ensure they are actionable.

22. **Review and Adapt:** Periodically review and adapt metrics to changing conditions.

23. **End:** Conclude the metric creation process.

This flowchart provides a structured approach to developing metrics for business analytics, ensuring that they are aligned with organizational goals and effectively measure performance in various domains.

UNIT 3

Every online store has a business model of its own. Many earn by attracting visitors to the website. Selecting ecommerce business model is a challenge, especially for beginners who have little to no experience in the industry.
Selecting the right model for your ecommerce venture is essential for keeping the store afloat and bringing in sustainable profits. However, when planning the ecommerce venture, many people make the mistake of jumping straight to the fine details and forget that all this depends upon what you plan to sell and what model you adopt for selling your inventory. If successfully executed, an ecommerce venture can become a significant source of income.
Table of Contents

Managed Cloud Hosting for Your Startup | Starts at $11

Focus on growing your business without the stress of managing website operations.

Clain Your 15% Discount

What is an Ecommerce Business Model?

An ecommerce business model refers to how a business operates to sell goods and services online. There are 6 main types of ecommerce business models, namely Business-to-Government (B2G), Business-to-Business (B2B), Business-to-Consumer (B2C), Consumer-to-Consumer (C2C), Consumer-to-Business (C2B), and Business-to-Business-to-Consumer (B2B2C).

In order to find the right ecommerce model for your business, you need to define two things. Firstly, you will have to define who you will sell to, and then define how you will position what you have to sell. Then, figure out your ecommerce business plan. This will define how you will attract customers and how they will engage with

your product. Lastly, figure out your delivery framework, by assessing what will work best for your ecommerce business.

What do You Want to Sell?

The beauty of online commerce is that you can sell pretty much anything. However, it is always a good idea to start with a small range of products. Your store can sell physical products (clothing or shoes), digital products (ebooks are a good place to start), or services such as babysitting.

Let's see what type of products are currently being sold online and how you can tap their market.

Physical Products

This is the most commonly sold commodity on ecommerce stores. Physical products (pretty much anything that requires packing, shipping and delivery) often achieve the highest sales.

But, how do you decide which products to sell?

Discover what you are passionate about. Do you love cars? How about selling car parts and accessories then? Do you love books? Why not start an online book store? Online commerce gives you the perfect opportunity for converting your passion into a viable business.

Analyze your chosen niche and find the opportunity gaps. This covers all the aspects of the industry that are underserved. Similarly, try to analyze the pain points of the target customers.

Next, conduct keyword research on the product you wish to sell. This way, you can pinpoint the demand for your product that will help you plan your inventory and order placements.

Digital Products

There are many products that can be delivered to a customer online. Are you a web designer, content writer, or drawing artist? You can create an ecommerce

store around digital products. Piracy and Copyrights infringements is a serious challenge for such stores. Another important requirement is the FAQ and Legal sections that cover the mechanism of product delivery and the copyright status of your offerings.

Services

If you have a crew of skilled carpenters, or house cleaners, or you are an expert hair stylist who offers to visit the customer's residence, why not create a website to sell these services online? You can significantly increase the demand for your services by creating a comprehensive FAQ section and a Legal section detailing exactly what you are offering and what the customers can expect.

6 Types of Ecommerce Business Models

Ecommerce is a global phenomenon and as such support several models. The good thing about ecommerce is that you could choose one or more models for your venture.

1. Business-to-Business (B2B)

If the nature of your products or services is geared towards meeting the needs of businesses, setting up a B2B strategy is your best bet. Networking and reaching out is a bigger part of this strategy. A big advertising budget is not of much help. The most important challenge you would face is convincing established businesses that your products/services are a great fit for their processes.

The advantage of this business model is that order sizes are usually large, and repeat orders are very common, if you maintain the quality of your products and services. An example of a great B2B model is Media Lounge.

2. Business-to-Consumer (B2C)

This is the model you should adopt if your products/services are targeted primarily towards individuals. The potential customer finds your website and determines whether your product could address their pain points.

After browsing the store, the customer may decide to place an order. An example of a successful B2C business is Portugal Footwear.

3. Consumer-to-Consumer (C2C)

While B2B and B2B business concepts are familiar, Customer-to-Customer (C2C) is a concept unique to ecommerce. This is mainly due to the sheer demand of the platforms such as Craigslist, OLX, and eBay.

These platforms allow their users to trade, buy, sell, and rent products and services. In all transactions, the platforms receives a small commission. This business model is complex and requires careful planning to operate. Many platforms have failed, generally due to legal issues.

4. Consumer-to-Business (C2B)

Customer-to-Business (C2B) business model is another great concept that is popular mainly due to platforms that cater to freelancers. In C2B, freelance workers work on tasks provided by clients. Most of these clients are commercial entities and freelancers are often individuals. In simpler terms, consider C2B is a sole proprietorship serving larger businesses.

Reverse auction websites, freelance marketplaces, affiliate marketing all form part of this business model. Again, this model requires planning due to the legal complexities involved.

5. Business to Government (B2G)

Business to Government (B2G) is an ecommerce business model where a business markets its products to government agencies. If you want to choose this ecommerce business model, you will have to bid on government contracts. Governments usually put up requests for proposals and ecommerce businesses then have to bid on government projects. In most cases, a government agency would not come to place an order on your ecommerce website. However, some local government agencies are exceptions to the rule, depending on their needs.

6. Business to Business to Consumer (B2B2C)

When a business sells products to another business, and then that business sells to the consumers online, this is what is defined as B2B2C ecommerce.

There are three parties involved in this type of ecommerce business model. For example, if you choose to go with it, you will have to partner with another business, and only then can you sell its products and offer the partner a commission for each sale.

Ecommerce store owners choose this business model mainly for new customer acquisition. This happens because even though customers are already familiar with the partner's products, they can't order from them online, due to obstacles such as geographical location, hefty shipping costs, and others.

Hence, this ecommerce business model is most suitable for new ecommerce store owners who want to expand their customer base.

Have You Selected An Ecommerce Business Model?

9 Innovative Value Delivery Frameworks for Your Ecommerce Store with Examples

Once you have chosen the ecommerce business model, the next step is the selection of an appropriate value delivery framework. Let's discuss the most innovative and profitable value delivery frameworks for ecommerce businesses.

Ecommerce Business Plans

1. Just-in-Time Purchasing
2. Dropshipping
3. Wholesaling
4. Warehousing
5. White-labeling
6. Outsourced Fulfillment
7. Subscriptions
8. Rent and Loan Model

9. Freemium Model

1. Just-in-Time Purchasing

Just-in-time purchasing is a popular business plan in which an ecommerce store put up products on the store. Whenever a user orders an item, the store gets the item from the supplier and ships it to customer. The plan is ideal for people with low budget or no warehousing space.

Just-in-Time Purchasing Example: Both Apple and McDonald's follow the just-in-time delivery framework. A case study on Apple suggests how this value delivery framework helped it streamline the waiting time and a number of steps in the delivery of its tailor-made iPods. From 90 days the delivery time was reduced to 90 hours as the JIT framework helped Apple produce tailor-made products when customers had placed orders.

2. Dropshipping

In this plan, an ecommerce store gets the products from a wholesale or manufacturer and sells to the visitors at a commission. For example, you have an ecommerce store where you add products from AliExpress and set the prices at a higher level. Once the store is up, the store targets potential customers through ads and other digital marketing channels.

Dropshipping became very popular when ecommerce dropshipping platforms like WooCommerce, PrestaShop, and Shopify went mainstream.

Dropshipping Example: Daily Steals is a successful ecommerce business that follows the dropshipping value delivery framework. It works in the niche of technology, home, and office, with a peak traffic of 1000000.

They are a successful dropshipping store because they consistently present the best deals for the products in their niche. They also make sure they show the discount badges on all products throughout every page of their site. Daily Steals

also uses PPC ads to their advantage and strategically place display ads showcasing their flashy premium deals and discounts.

3. Wholesaling

Wholesaling is a business plan where an ecommerce store sells products in bulk and at a lower price than the general market prices. The biggest example of this model is Alibaba, a very popular platform for small and large wholesalers that trade with the businesses all over the world.

4. Warehousing

Many ecommerce stores have warehouses where they keep products. These are then put as listings on the ecommerce stores and when a person buys them, they are shipped directly from the warehouse.

5. White-labeling

White-label branding is a business plan in which one company produces the product, and another company rebrands and distributes it. An example of this plan is of influencers that sell white-label products through their social media accounts.

White-labeling Example: Seed Beauty, a private label company, produces Kylie Jenner's products and white label cosmetics for ColourPop. White-label products are generic products that are mass-produced. For example, if you want to sell white label cosmetics, you can focus on one product, like lip balm.

6. Outsourced Fulfillment

Outsource fulfillment is a business model in which the shipping is outsourced to a third party. This model is mostly used by ecommerce stores that are too busy running the operations or too understaffed to ship the products themselves.

Fulfillment by Amazon (FBA) and 3PL services for ecommerce stores fall under this category.

Outsourced Fulfillment Example: A hand sanitizer brand, Touch Land, was growing fast during covid, selling high-quality moisturizing hand sanitizers. But soon, they were sold out, and had almost 34000 customers on their waiting list. They even did pre-orders to meet the demand as they had up to 700 orders per day and sold 10,000 dispensers to industry-leading brands in those three months. This is when using 3PL services helped them.

7. Subscriptions

A subscription based e-business model allows users to purchase and then subscribe to a service for a set period (usually monthly or annually). Once the product subscription expires, the users could either terminate the contract or renew it. Ecommerce stores such as Tie Bar and Five Four Club work on this subscription-based business model.

Subscription Model Example: An American meal kit service, Blue Apron, provides high-quality food ingredients. It allows its customers to set their food preferences and then takes care of everything after the customers have subscribed to receiving their meal kits. It's a great example of how the subscription business model can work with an ecommerce store.

8. Rent and Loan Model

With better digital payment models, it is now possible to set up rent and loan business plans. Under this plan, users or companies can rent out physical or digital products to others for a monthly cost. In several cass, this model also include lending money for earning interest.

Websites such as Loan Now and Lending Club work on this model.

Rent and Loan Model Example: Lending Luxury is an ecommerce store that is successfully using the rent and loan ecommerce value delivery framework to make luxury apparel affordable for couture-hungry people.

9. Freemium Model

Freemium is a pricing model in which some features of a product are provided to the users for free, with the rest behind a paywall. Hootsuite uses this strategy of its social media scheduling service. It provides a limited number of posts' scheduling for free. Users have to pay for unlimited scheduling.

Freemium Model Example: Spotify is a music streaming platform that uses this strategy. Users can access basic, limited, ad-supported service for free. However, they have to upgrade to the premium account to access unlimited service for a subscription fee.

Ecommerce Promotion Options

Once you have the model and the plan, the next step is choosing the right promotion options.

Ecommerce Promotion Options

1. Affiliate Marketing
2. Pay Per Click (PPC)
3. Pay Per Sale (PPS)
4. Pay per Lead (PPL)
5. Pay Per Action (PPA)
6. Pay Per Mile (PPM)
7. Pay Per View (PPV)
8. Native Advertising
9. Sponsored Posts

Affiliate Marketing

Affiliate marketing is when you promote a product by another producer/supplier on your website or blog. Top ecommerce websites provide affiliate programs where content producers can sign up to become an affiliate. For example, you can sign up for Cloudways web hosting affiliate program. When you bring leads to Cloudways, you will earn a commission. Check the complete details of how an affiliate program works on the affiliate page.

Pay Per Click (PPC)

Pay Per Click (PPC) is a e-business model in which the advertisers will pay for every click that leads to their products page. The business model is offered by affiliate marketing programs such as Viral9 and Max Bounty.

Pay Per Sale (PPS)

Similar to the PPC business model, Pay Per Sale (PPS) lets publishers or promoters earn commission whenever they drive a sale for the advertiser. This model is practiced by affiliate marketers such as Shareasale, Clickbank and Commission junction.

Pay per Lead (PPL)

Pay per Lead (PPL) works in the same way as the above to work. The only difference is that in a PPL plan, promoters receive commission for the leads. Lead generation programs such as Facebook Ads, Google Adwords, and Maxbounty offer this plan.

Pay Per Action (PPA)

Pay Per Action (PPA) is a generic term in marketing. The model applies to all types of affiliate marketing tactics in which any type of action is expected of the visitors. Usually, leads, sales, and clicks all are considered valid outcomes for PPA programs. Most affiliate programs use these actions as the performance measurement unit for their campaigns. Maxbounty uses PPA or (CPA Marketing) for all types of lead generation methods.

Pay Per View (PPV)

Pay Per View (PPV) is used for video marketing. Youtube, Dailymotion, and Facebook videos use this as a unit of measurement for paying to their content producers. A 'view' is usually ten seconds long because that is the duration after which Google shows an ad.

PPM

Pay Per Mile (PPM) is a unit of measurement used for display advertisements. It is used by Google Ads for paying the users for every 1000 views. Every 1000 views are considered a 'mile' in the marketing terminology. Rates for per mile marketing varies from country to country. In the USA, the PPM cost would be higher while in India it would be a lot lower.

Native Advertising

Native advertisements is a recent addition to online marketing. It came to light when Buzzfeed started adding promotional content in native or natural articles. These articles were about regular topics but promoted brand products by mentioning them somewhere within the article without breaking the flow. Readers would not think that these are promotional articles as nothing of the sort is mentioned in the content. Native advertising costs vary on various factors such as the content websites rankings. While going this route, make sure what Federal Trade Commission (FTC) says about them.

Sponsored Posts

Sponsored posts is a practice in which one brand buys an article of a third-party website. In sponsored posts, the terms 'Sponsored,' 'Paid,' or 'Promoted' is mentioned at the top so readers know that this is not a regular publication of the website.

How to Choose Your Ecommerce Business Model?

It is very important to come to the right decision when choosing your ecommerce business model. Why? Because once you have selected one, your finances will be involved in your ecommerce business and so will your time and efforts. Therefore, it is important that you ask yourself some primary questions before you choose one. We have a few examples.

- What will you sell and how much will you sell it for?
- Will you sell a single product or want to sell a range of products?
- Who is your product for? Who is the audience you want to sell to?
- What would your audience/potential customers want? What will be their expectations?
- Which factors will you compete on? (price, quality, selection, service, the value you add, or something else)

Answering the above set of questions and being clear on them will give you a clear idea of what ecommerce business model would work best for your online store.

Which Type of Ecommerce Business Model should I choose?

As you must have observed by now that there are several ecommerce models available, each with its own set of benefits. The right thing to do is to analyze your business model and then pick the right model.

What is Market Engineering?

Market engineering pursues the goal of making a market work as efficiently as possible. From the design to transaction process, this concept of addressing a market is useful for implementing new policies and proposing principles that keep company goals and approaches aligned.

Since the term "market" has been used in many different contexts that ranges from a store, to a chain, to a stock exchange, to even broader activity, a business

must adopt a stricter definition. In order to understand market engineering, it helps to start with a solid definition of a market, which is as follows:

A market is a set of rules that structure how costs are set and information is exchanged in order for transactions to generate a desired profit.

The foundation for determining costs and pricing is deeply rooted in supply and demand conditions. Other factors that shape a market include the set of constraints determined by the marketer, survival strategies, an information processing system, business activity and a marketable service.

Market engineering involves creating a conscious and well structured system of procedures for analyzing and re-engineering an efficient market that facilitates transactions.

Elements to analyze

- market lifecycle
- market engineering object
- market engineering processes

Various activity that occurs within a market can be closely associated with lifestyle. The type of people who are attracted to a particular product or service may be connected based on their lifecycle choices. They may like adventure or convenience, which can be reflected in the product's benefits. The four phases of a product lifecycle are:

1. emerging
2. growth
3. maturity
4. decline

Market Lifecycle

New markets emerge outside of existing traditional markets, as a result of imaginative solutions to market problems not being served. Usually R&D and engineering are the focal points of this phase, as designing and testing shape the product and the marketing strategy. The product may not be profitable yet, regardless of high expectations.

Growth occurs from promotion and responding to market feedback about quality. It may attract new competitors who see openings in the market. During growth periods, a top concern may be refining IT systems and personnel.

When competition escalates in the growth phase, it gives market participants many advantages. This stage attracts competitors on eBay, adding B2C and C2C players who affect the marketplace. Buyers now have online auction options to bid for the lowest prices. Meanwhile, sellers use the same platforms to try to sell at the highest prices. Such platforms help expand market choices.

Maturity represents a streamlined product in its most efficient form. At this phase operators pay closer attention to costs and how to gain an edge over competitors. It's the period in which prices may fall due to increased competition trying to offer the best deals. It's at this stage that the market engineer must decide if the product needs to incorporate new features for survival in a crowded market.

Decline is a phase in which the product has run its course in serving a purpose. It may be surpassed in quality by other products or people have moved on to other concerns. The product may have only been envisioned to make a disruptive impact for a certain length of time. When the goal is to release a series of products, a short-term product can set up a long-term product.

Some markets are regulated in a manner that only so many players are allowed to compete, as in the licensing of radio and television broadcast stations. Other markets occur "naturally" with wide open opportunities and are not shaped or controlled as much by regulators.

Market Engineering Object

The most crucial elements of a market that can determine success are collectively known as the market engineering object. Viewing a static diagram that lists the most important components of a market is another way of understanding this model. The main components include:

- Market Outcome Agent Behavior
- Market Structure: microstructure, IT infrastructure, business structure
- Transaction Elements
- Socio-economic and Legal Conditions

A market engineer must design a transaction object that fits the expectations of the desired market outcome or performance. The engineer also designs the market structure, which includes all the details of how the business and technology work together. But the transaction object and the market structure are only instrumental for success and do not have a direct impact on the market outcome. Behavior of the market participants will determine what actually happens, making it challenging for the market engineer to see ahead of the curve.

The socio-economic and legal environment is outside the control of the market engineer. Applicable laws determine much of the framework and context in which business planners must operate and plan for the future.

Market Engineering Process

A more detailed model of the market lifecycle is needed to map out the market engineering process of a market institution. This process begins with an environmental analysis, which looks closer at transactions, participants, customers and market requirements. At this phase the market engineer must identify tools and methods that will allow the business to reach its goals using the most cost efficient methods. At this stage, tools and methods listed in the

preliminary design are modified or omitted to reflect the most useful choices to help increase productivity and efficiency.

Deciding on Methodologies

Methodologies are the part of market structure that determines systems that speed up productivity, leading to a more efficient operation. Choosing consistent methods helps save time in determining costs, work schedules and product quality. Rules and guidelines need to be set so that managers can train employees to perform with self-discipline, which yields better results.

Conclusion

The reason businesses should explore market engineering before rushing into a launch is that it will help them structure the business so that it fits into the market. This process answers questions about how transactions will be made and how other market components will work together. The key is to integrate multiple marketing concepts with attention paid to detail as far as research and design.

Customer Segmentation in Python: A Practical Approach

Customer segmentation can help businesses tailor their marketing efforts and improve customer satisfaction. Here's how.

Functionally, customer segmentation involves dividing a customer base into distinct *groups* or *segments*—based on shared characteristics and behaviors. By understanding the needs and preferences of each segment, businesses can deliver more personalized and effective marketing campaigns, leading to increased customer retention and revenue.

In this tutorial, we'll explore customer segmentation in Python by combining two fundamental techniques: **RFM (Recency, Frequency, Monetary) analysis** and **K-Means clustering**. RFM analysis provides a structured framework for evaluating customer behavior, while K-means clustering offers a data-driven approach to

group customers into meaningful segments. We'll work with a real-world dataset from the retail industry: the Online Retail dataset from UCI machine learning repository.

From data preprocessing to cluster analysis and visualization, we'll code our way through each step. So let's dive in!

Our Approach: RFM Analysis and K-Means Clustering

Let's start by stating our goal: By applying RFM analysis and K-means clustering to this dataset, we'd like to gain insights into customer behavior and preferences.

RFM Analysis is a simple yet powerful method to quantify customer behavior. It evaluates customers based on three key dimensions:

- **Recency (R)**: How recently did a particular customer make a purchase?
- **Frequency (F)**: How often do they make purchases?
- **Monetary Value (M)**: How much money do they spend?

We'll use the information in the dataset to compute the recency, frequency, and monetary values. Then, we'll map these values to the generally used RFM score scale of 1 - 5.

If you'd like, you can explore and analyze further using these RFM scores. But we'll try to identify customer segments with similar RFM characteristics. And for this, we'll use K-Means clustering, an unsupervised machine learning algorithm that groups similar data points into clusters.

So let's start coding!

· Link to Google Colab notebook.

Step 1 – Import Necessary Libraries and Modules

First, let's import the necessary libraries and the specific modules as needed:

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
```

We need pandas and matplotlib for data exploration and visualization, and the KMeans class from scikit-learn's cluster module to perform K-Means clustering.

Step 2 – Load the Dataset

As mentioned, we'll use the Online Retail dataset. The dataset contains customer records: transactional information, including purchase dates, quantities, prices, and customer IDs.

Let's read in the data that's originally in an excel file from its URL into a pandas dataframe.

```
# Load the dataset from UCI repository
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/00352/Online%20Retail.xlsx"
data = pd.read_excel(url)
```

Alternatively, you can download the dataset and read the excel file into a pandas dataframe.

Step 3 – Explore and Clean the Dataset

Now let's start exploring the dataset. Look at the first few rows of the dataset:

data.head()

Output of data.head()

Now call the describe() method on the dataframe to understand the numerical features better:

data.describe()

We see that the "CustomerID" column is currently a floating point value. When we clean the data, we'll cast it into an integer:

Output of data.describe()

Also note that the dataset is quite noisy. The "Quantity" and "UnitPrice" columns contain negative values:

Output of data.describe()

Let's take a closer look at the columns and their data types:

data.info()

We see that the dataset has over 541K records and the “Description” and “CustomerID” columns contain missing values:

Let’s get the count of missing values in each column:

```
# Check for missing values in each column
missing_values = data.isnull().sum()
print(missing_values)
```

As expected, the “CustomerID” and “Description” columns contain missing values:

For our analysis, we don’t need the product description contained in the “Description” column. However, we need the “CustomerID” for the next steps in our analysis. So let’s drop the records with missing “CustomerID”:

```
# Drop rows with missing CustomerID
data.dropna(subset=['CustomerID'], inplace=True)
```

Also recall that the values “Quantity” and “UnitPrice” columns should be strictly non-negative. But they contain negative values. So let's also drop the records with negative values for “Quantity” and “UnitPrice”:

```
# Remove rows with negative Quantity and Price
data = data[(data['Quantity'] > 0) & (data['UnitPrice'] > 0)]
```

Let's also convert the "CustomerID" to an integer:

```
data['CustomerID'] = data['CustomerID'].astype(int)

# Verify the data type conversion
print(data.dtypes)
```

Step 4 – Compute Recency, Frequency, and Monetary Value

Let's start out by defining a reference date snapshot_date that's a day later than the most recent date in the "InvoiceDate" column:

```
snapshot_date = max(data['InvoiceDate']) + pd.DateOffset(days=1)
```

Next, create a "Total" column that contains Quantity*UnitPrice for all the records:

```
data['Total'] = data['Quantity'] * data['UnitPrice']
```

To calculate the Recency, Frequency, and MonetaryValue, we calculate the following—**grouped by CustomerID**:

- For **recency**, we'll calculate the difference between the most recent purchase date and a reference date (snapshot_date). This gives the **number of days since the customer's last purchase**. So *smaller values* indicate that a customer has made a purchase *more recently*. But when we talk about *recency scores*, we'd want customers who bought recently to have a higher recency score, yes? We'll handle this in the next step.

- Because **frequency** measures how often a customer makes purchases, we'll calculate it as the total **number of unique invoices** or transactions made by each customer.
- **Monetary value** quantifies how much money a customer spends. So we'll find the average of the total monetary value across transactions.

```
rfm = data.groupby('CustomerID').agg({
    'InvoiceDate': lambda x: (snapshot_date - x.max()).days,
    'InvoiceNo': 'nunique',
    'Total': 'sum'
})
```

Let's rename the columns for readability:

```
rfm.rename(columns={'InvoiceDate': 'Recency', 'InvoiceNo': 'Frequency', 'Total': 'MonetaryValue'}, inplace=True)
rfm.head()
```

Step 5 – Map RFM Values onto a 1-5 Scale

Now let's map the "Recency", "Frequency", and "MonetaryValue" columns to take on values in a scale of 1-5; one of {1,2,3,4,5}.

We'll essentially assign the values to five different bins, and map each bin to a value. To help us fix the bin edges, let's use the quantile values of the "Recency", "Frequency", and "MonetaryValue" columns:

```
rfm.describe()
```

Here's how we define the custom bin edges:

```
# Calculate custom bin edges for Recency, Frequency, and Monetary scores
recency_bins = [rfm['Recency'].min()-1, 20, 50, 150, 250, rfm['Recency'].max()]
frequency_bins = [rfm['Frequency'].min() - 1, 2, 3, 10, 100, rfm['Frequency'].max()]
monetary_bins = [rfm['MonetaryValue'].min() - 3, 300, 600, 2000, 5000, rfm['MonetaryValue'].max()]
```

Now that we've defined the bin edges, let's map the scores to corresponding labels between 1 and 5 (both inclusive):

```
# Calculate Recency score based on custom bins
rfm['R_Score'] = pd.cut(rfm['Recency'], bins=recency_bins, labels=range(1, 6), include_lowest=True)

# Reverse the Recency scores so that higher values indicate more recent purchases
rfm['R_Score'] = 5 - rfm['R_Score'].astype(int) + 1

# Calculate Frequency and Monetary scores based on custom bins
rfm['F_Score'] = pd.cut(rfm['Frequency'], bins=frequency_bins, labels=range(1, 6), include_lowest=True).astype(int)
rfm['M_Score'] = pd.cut(rfm['MonetaryValue'], bins=monetary_bins, labels=range(1, 6), include_lowest=True).astype(int)
```

Notice that the R_Score, based on the bins, is 1 for recent purchases 5 for all purchases made over 250 days ago. But we'd like the most recent purchases to

have an R_Score of 5 and purchases made over 250 days ago to have an R_Score of 1.

To achieve the desired mapping, we do: 5 - rfm['R_Score'].astype(int) + 1.

Let's look at the first few rows of the R_Score, F_Score, and M_Score columns:

```
# Print the first few rows of the RFM DataFrame to verify the scores
print(rfm[['R_Score', 'F_Score', 'M_Score']].head(10))
```

If you'd like, you can use these R, F, and M scores to carry out an in-depth analysis. Or use clustering to identify segments with similar RFM characteristics. We'll choose the latter!

Step 6 – Perform K-Means Clustering

K-Means clustering is sensitive to the scale of features. Because the R, F, and M values are all on the same scale, we can proceed to perform clustering without further scaling the features.

Let's extract the R, F, and M scores to perform K-Means clustering:

```
# Extract RFM scores for K-means clustering
X = rfm[['R_Score', 'F_Score', 'M_Score']]
```

Next, we need to find the *optimal* number of clusters. For this let's run the K-Means algorithm for a range of K values and use the *elbow method* to pick the optimal K:

```
# Calculate inertia (sum of squared distances) for different values of k
inertia = []
```

```
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, n_init= 10, random_state=42)
    kmeans.fit(X)
    inertia.append(kmeans.inertia_)

# Plot the elbow curve
plt.figure(figsize=(8, 6),dpi=150)
plt.plot(range(2, 11), inertia, marker='o')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')
plt.title('Elbow Curve for K-means Clustering')
plt.grid(True)
plt.show()
```

We see that the curve elbows out at 4 clusters. So let's divide the customer base into four segments.

We've fixed K to 4. So let's run the K-Means algorithm to get the cluster assignments for all points in the dataset:

```
# Perform K-means clustering with best K
best_kmeans = KMeans(n_clusters=4, n_init=10, random_state=42)
rfm['Cluster'] = best_kmeans.fit_predict(X)
```

Step 7 – Interpret the Clusters to Identify Customer Segments

Now that we have the clusters, let's try to characterize them based on the RFM scores.

```
# Group by cluster and calculate mean values
cluster_summary = rfm.groupby('Cluster').agg({
    'R_Score': 'mean',
    'F_Score': 'mean',
    'M_Score': 'mean'
}).reset_index()
```

The average R, F, and M scores for each cluster should already give you an idea of the characteristics.

```
print(cluster_summary)
```

But let's visualize the average R, F, and M scores for the clusters so it's easy to interpret:

```
colors = ['#3498db', '#2ecc71', '#f39c12','#C9B1BD']

# Plot the average RFM scores for each cluster
plt.figure(figsize=(10, 8),dpi=150)

# Plot Avg Recency
plt.subplot(3, 1, 1)
bars = plt.bar(cluster_summary.index, cluster_summary['R_Score'], color=colors)
plt.xlabel('Cluster')
plt.ylabel('Avg Recency')
```

```
plt.title('Average Recency for Each Cluster')

plt.grid(True, linestyle='--', alpha=0.5)
plt.legend(bars, cluster_summary.index, title='Clusters')

# Plot Avg Frequency
plt.subplot(3, 1, 2)
bars = plt.bar(cluster_summary.index, cluster_summary['F_Score'], color=colors)
plt.xlabel('Cluster')
plt.ylabel('Avg Frequency')
plt.title('Average Frequency for Each Cluster')
plt.grid(True, linestyle='--', alpha=0.5)
plt.legend(bars, cluster_summary.index, title='Clusters')

# Plot Avg Monetary
plt.subplot(3, 1, 3)
bars = plt.bar(cluster_summary.index, cluster_summary['M_Score'], color=colors)
plt.xlabel('Cluster')
plt.ylabel('Avg Monetary')
plt.title('Average Monetary Value for Each Cluster')
plt.grid(True, linestyle='--', alpha=0.5)
plt.legend(bars, cluster_summary.index, title='Clusters')

plt.tight_layout()
plt.show()
```

Notice how the customers in each of the segments can be characterized based on the recency, frequency, and monetary values:

- **Cluster 0**: Of all the four clusters, this cluster has the *highest* recency, frequency, and monetary values. Let's call the customers in this cluster **champions (or power shoppers)**.
- **Cluster 1**: This cluster is characterized by *moderate* recency, frequency, and monetary values. These customers still spend more and purchase more frequently than clusters 2 and 3. Let's call them **loyal customers**.
- **Cluster 2**: Customers in this cluster tend to spend less. They don't buy often, and haven't made a purchase recently either. These are likely *inactive* or **at-risk customers**.
- **Cluster 3**: This cluster is characterized by *high recency* and relatively lower frequency and moderate monetary values. So these are **recent customers** who can potentially become long-term customers.

Here are some examples of how you can tailor marketing efforts—to target customers in each segment—to enhance customer engagement and retention:

- **For Champions/Power Shoppers**: Offer personalized special discounts, early access, and other premium perks to make them feel valued and appreciated.
- **For Loyal Customers**: Appreciation campaigns, referral bonuses, and rewards for loyalty.
- **For At-Risk Customers**: Re-engagement efforts that include running discounts or promotions to encourage buying.

- **For Recent Customers**: Targeted campaigns educating them about the brand and discounts on subsequent purchases.

It's also helpful to understand what percentage of customers are in the different segments. This will further help streamline marketing efforts and grow your business.

Let's visualize the distribution of the different clusters using a pie chart:

```
cluster_counts = rfm['Cluster'].value_counts()

colors = ['#3498db', '#2ecc71', '#f39c12','#C9B1BD']
# Calculate the total number of customers
total_customers = cluster_counts.sum()

# Calculate the percentage of customers in each cluster
percentage_customers = (cluster_counts / total_customers) * 100

labels = ['Champions(Power Shoppers)','Loyal Customers','At-risk Customers','Recent Customers']

# Create a pie chart
plt.figure(figsize=(8, 8),dpi=200)
plt.pie(percentage_customers, labels=labels, autopct='%1.1f%%', startangle=90, colors=colors)
plt.title('Percentage of Customers in Each Cluster')
plt.legend(cluster_summary['Cluster'], title='Cluster', loc='upper left')

plt.show()
```

For this example, we have quite an even distribution of customers across segments. So we can invest time and effort in retaining existing customers, re-engaging with at-risk customers, and educating recent customers.

Wrapping Up

And that's a wrap! We went from over *154K customer records to 4 clusters in 7 easy steps*. I hope you understand how customer segmentation allows you to make data-driven decisions that influence business growth and customer satisfaction by allowing for:

- **Personalization**: Segmentation allows businesses to tailor their marketing messages, product recommendations, and promotions to each customer group's specific needs and interests.
- **Improved Targeting**: By identifying *high-value* and *at-risk* customers, businesses can allocate resources more efficiently, focusing efforts where they are most likely to yield results.
- **Customer Retention**: Segmentation helps businesses create retention strategies by understanding what keeps customers engaged and satisfied.

As a next step, try applying this approach to another dataset, document your journey, and share with the community! But remember, effective customer segmentation and running targeted campaigns requires a good understanding of your customer base—and how the customer base evolves. So it requires periodic analysis to refine your strategies over time.

Dataset Credits

The Online Retail Dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license:

10 Clustering Algorithms With Python

Clustering or **cluster analysis** is an unsupervised learning problem.

It is often used as a data analysis technique for discovering interesting patterns in data, such as groups of customers based on their behavior.

There are many clustering algorithms to choose from and no single best clustering algorithm for all cases. Instead, it is a good idea to explore a range of clustering algorithms and different configurations for each algorithm.

- Clustering is an unsupervised problem of finding natural groups in the feature space of input data.
- There are many different clustering algorithms and no single best method for all datasets.
- How to implement, fit, and use top clustering algorithms in Python with the scikit-learn machine learning library.

Tutorial Overview

This tutorial is divided into three parts; they are:

1. Clustering
2. Clustering Algorithms
3. Examples of Clustering Algorithms
    1. Library Installation
    2. Clustering Dataset
    3. Affinity Propagation

4. Agglomerative Clustering
5. BIRCH
6. DBSCAN
7. K-Means
8. Mini-Batch K-Means
9. Mean Shift
10. OPTICS
11. Spectral Clustering
12. Gaussian Mixture Model

Clustering

Cluster analysis, or clustering, is an unsupervised machine learning task.

It involves automatically discovering natural grouping in data. Unlike supervised learning (like predictive modeling), clustering algorithms only interpret the input data and find natural groups or clusters in feature space.

Clustering techniques apply when there is no class to be predicted but rather when the instances are to be divided into natural groups.

— Page 141, Data Mining: Practical Machine Learning Tools and Techniques, 2016.

A cluster is often an area of density in the feature space where examples from the domain (observations or rows of data) are closer to the cluster than other clusters. The cluster may have a center (the centroid) that is a sample or a point feature space and may have a boundary or extent.

These clusters presumably reflect some mechanism at work in the domain from which instances are drawn, a mechanism that causes some instances to bear a stronger resemblance to each other than they do to the remaining instances.

— Pages 141-142, Data Mining: Practical Machine Learning Tools and Techniques, 2016.

Clustering can be helpful as a data analysis activity in order to learn more about the problem domain, so-called pattern discovery or knowledge discovery.

For example:

- The phylogenetic tree could be considered the result of a manual clustering analysis.
- Separating normal data from outliers or anomalies may be considered a clustering problem.
- Separating clusters based on their natural behavior is a clustering problem, referred to as market segmentation.

Clustering can also be useful as a type of feature engineering, where existing and new examples can be mapped and labeled as belonging to one of the identified clusters in the data.

Evaluation of identified clusters is subjective and may require a domain expert, although many clustering-specific quantitative measures do exist. Typically, clustering algorithms are compared academically on synthetic datasets with pre-defined clusters, which an algorithm is expected to discover.

Clustering is an unsupervised learning technique, so it is hard to evaluate the quality of the output of any given method.

— Page 534, Machine Learning: A Probabilistic Perspective, 2012.

Clustering Algorithms

There are many types of clustering algorithms.

Many algorithms use similarity or distance measures between examples in the feature space in an effort to discover dense regions of observations. As such, it is often good practice to scale data prior to using clustering algorithms.

Central to all of the goals of cluster analysis is the notion of the degree of similarity (or dissimilarity) between the individual objects being clustered. A clustering method attempts to group the objects based on the definition of similarity supplied to it.

— Page 502, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2016.

Some clustering algorithms require you to specify or guess at the number of clusters to discover in the data, whereas others require the specification of some minimum distance between observations in which examples may be considered "*close*" or "*connected*."

As such, cluster analysis is an iterative process where subjective evaluation of the identified clusters is fed back into changes to algorithm configuration until a desired or appropriate result is achieved.

The scikit-learn library provides a suite of different clustering algorithms to choose from.

A list of 10 of the more popular algorithms is as follows:

- Affinity Propagation
- Agglomerative Clustering
- BIRCH
- DBSCAN
- K-Means
- Mini-Batch K-Means
- Mean Shift
- OPTICS
- Spectral Clustering
- Mixture of Gaussians

Each algorithm offers a different approach to the challenge of discovering natural groups in data.

There is no best clustering algorithm, and no easy way to find the best algorithm for your data without using controlled experiments.

In this tutorial, we will review how to use each of these 10 popular clustering algorithms from the scikit-learn library.

The examples will provide the basis for you to copy-paste the examples and test the methods on your own data.

We will not dive into the theory behind how the algorithms work or compare them directly. For a good starting point on this topic, see:

- Clustering, scikit-learn API.

Examples of Clustering Algorithms

In this section, we will review how to use 10 popular clustering algorithms in scikit-learn.

This includes an example of fitting the model and an example of visualizing the result.

The examples are designed for you to copy-paste into your own project and apply the methods to your own data.

Library Installation

First, let's install the library.

Don't skip this step as you will need to ensure you have the latest version installed.

You can install the scikit-learn library using the pip Python installer, as follows:

```
1 sudo pip install scikit-learn
```

For additional installation instructions specific to your platform, see:

- Installing scikit-learn

Next, let's confirm that the library is installed and you are using a modern version.

Run the following script to print the library version number.

```
# check scikit-learn version
import sklearn
print(sklearn.__version__)
```

Running the example, you should see the following version number or higher.

```
0.22.
 1
```

Clustering Dataset

We will use the make_classification() function to create a test binary classification dataset.

The dataset will have 1,000 examples, with two input features and one cluster per class. The clusters are visually obvious in two dimensions so that we can plot the data with a scatter plot and color the points in the plot by the assigned cluster. This will help to see, at least on the test problem, how "well" the clusters were identified.

The clusters in this test problem are based on a multivariate Gaussian, and not all clustering algorithms will be effective at identifying these types of clusters. As such, the results in this tutorial should not be used as the basis for comparing the methods generally.

An example of creating and summarizing the synthetic clustering dataset is listed below.

```
# synthetic classification dataset
from numpy import where
from sklearn.datasets import make_classification
from matplotlib import pyplot
# define dataset
```

```
X, y = make_classification(n_samples=1000, n_features=2, n_informative=2,
n_redundant=0, n_clusters_per_class=1, random_state=4)
# create scatter plot for samples from each class
for class_value in range(2):
# get row indexes for samples with this class
 row_ix = where(y == class_value)
# create scatter of these samples
 pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
# show the plot
pyplot.show()
```

Running the example creates the synthetic clustering dataset, then creates a scatter plot of the input data with points colored by class label (idealized clusters). We can clearly see two distinct groups of data in two dimensions and the hope would be that an automatic clustering algorithm can detect these groupings.

Scatter Plot of Synthetic Clustering Dataset With Points Colored by Known Cluster

Next, we can start looking at examples of clustering algorithms applied to this dataset.

I have made some minimal attempts to tune each method to the dataset.

Can you get a better result for one of the algorithms?

Affinity Propagation

Affinity Propagation involves finding a set of exemplars that best summarize the data.

We devised a method called "affinity propagation," which takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges

— Clustering by Passing Messages Between Data Points, 2007.

The technique is described in the paper:

- Clustering by Passing Messages Between Data Points, 2007.

It is implemented via the AffinityPropagation class and the main configuration to tune is the "*damping*" set between 0.5 and 1, and perhaps "preference."

The complete example is listed below.

```
# affinity propagation clustering
from numpy import unique
from numpy import where
from sklearn.datasets import make_classification
from sklearn.cluster import AffinityPropagation
from matplotlib import pyplot
# define dataset
X, _ = make_classification(n_samples=1000, n_features=2, n_informative=2,
n_redundant=0, n_clusters_per_class=1, random_state=4)
# define the model
model = AffinityPropagation(damping=0.9)
# fit the model
model.fit(X)
# assign a cluster to each example
yhat = model.predict(X)
# retrieve unique clusters
clusters = unique(yhat)
# create scatter plot for samples from each cluster
for cluster in clusters:
 # get row indexes for samples with this cluster
 row_ix = where(yhat == cluster)
```

```
# create scatter of these samples
pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
# show the plot
pyplot.show()
```

Running the example fits the model on the training dataset and predicts a cluster for each example in the dataset. A scatter plot is then created with points colored by their assigned cluster.

In this case, I could not achieve a good result.

Scatter Plot of Dataset With Clusters Identified Using Affinity Propagation

Agglomerative Clustering

Agglomerative clustering involves merging examples until the desired number of clusters is achieved.

It is a part of a broader class of hierarchical clustering methods and you can learn more here:

- Hierarchical clustering, Wikipedia.

It is implemented via the AgglomerativeClustering class and the main configuration to tune is the "*n_clusters*" set, an estimate of the number of clusters in the data, e.g. 2.

The complete example is listed below.

```
# agglomerative clustering
from numpy import unique
from numpy import where
from sklearn.datasets import make_classification
from sklearn.cluster import AgglomerativeClustering
from matplotlib import pyplot
# define dataset
```

```
X, _ = make_classification(n_samples=1000, n_features=2, n_informative=2,
n_redundant=0, n_clusters_per_class=1, random_state=4)
# define the model
model = AgglomerativeClustering(n_clusters=2)
# fit model and predict clusters
yhat = model.fit_predict(X)
# retrieve unique clusters
clusters = unique(yhat)
# create scatter plot for samples from each cluster
for cluster in clusters:
 # get row indexes for samples with this cluster
 row_ix = where(yhat == cluster)
 # create scatter of these samples
 pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
# show the plot
pyplot.show()
```

Running the example fits the model on the training dataset and predicts a cluster for each example in the dataset. A scatter plot is then created with points colored by their assigned cluster.

In this case, a reasonable grouping is found.

Scatter Plot of Dataset With Clusters Identified Using Agglomerative Clustering

BIRCH

BIRCH Clustering (BIRCH is short for Balanced Iterative Reducing and Clustering using

Hierarchies) involves constructing a tree structure from which cluster centroids are extracted.

BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality clustering with the available resources (i. e., available memory and time constraints).

— BIRCH: An efficient data clustering method for large databases, 1996.

The technique is described in the paper:

- BIRCH: An efficient data clustering method for large databases, 1996.

It is implemented via the Birch class and the main configuration to tune is the "*threshold*" and "*n_clusters*" hyperparameters, the latter of which provides an estimate of the number of clusters.

The complete example is listed below.

```
# birch clustering
from numpy import unique
from numpy import where
from sklearn.datasets import make_classification
from sklearn.cluster import Birch
from matplotlib import pyplot
# define dataset
X, _ = make_classification(n_samples=1000, n_features=2, n_informative=2,
n_redundant=0, n_clusters_per_class=1, random_state=4)
# define the model
model = Birch(threshold=0.01, n_clusters=2)
# fit the model
model.fit(X)
# assign a cluster to each example
yhat = model.predict(X)
# retrieve unique clusters
```

```
clusters = unique(yhat)
# create scatter plot for samples from each cluster
for cluster in clusters:
 # get row indexes for samples with this cluster
 row_ix = where(yhat == cluster)
 # create scatter of these samples
 pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
# show the plot
pyplot.show()
```

Running the example fits the model on the training dataset and predicts a cluster for each example in the dataset. A scatter plot is then created with points colored by their assigned cluster.

In this case, an excellent grouping is found.

Scatter Plot of Dataset With Clusters Identified Using BIRCH Clustering

DBSCAN

DBSCAN Clustering (where DBSCAN is short for Density-Based Spatial Clustering of Applications with Noise) involves finding high-density areas in the domain and expanding those areas of the feature space around them as clusters.

... we present the new clustering algorithm DBSCAN relying on a density-based notion of clusters which is designed to discover clusters of arbitrary shape. DBSCAN requires only one input parameter and supports the user in determining an appropriate value for it

— A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, 1996.

The technique is described in the paper:

- A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, 1996.

It is implemented via the DBSCAN class and the main configuration to tune is the "*eps*" and "*min_samples*" hyperparameters.

The complete example is listed below.

```
# dbscan clustering
from numpy import unique
from numpy import where
from sklearn.datasets import make_classification
from sklearn.cluster import DBSCAN
from matplotlib import pyplot
# define dataset
X, _ = make_classification(n_samples=1000, n_features=2, n_informative=2,
n_redundant=0, n_clusters_per_class=1, random_state=4)
# define the model
model = DBSCAN(eps=0.30, min_samples=9)
# fit model and predict clusters
yhat = model.fit_predict(X)
# retrieve unique clusters
clusters = unique(yhat)
# create scatter plot for samples from each cluster
for cluster in clusters:
 # get row indexes for samples with this cluster
 row_ix = where(yhat == cluster)
 # create scatter of these samples
 pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
# show the plot
pyplot.show()
```

Running the example fits the model on the training dataset and predicts a cluster for each example in the dataset. A scatter plot is then created with points colored by their assigned cluster.

In this case, a reasonable grouping is found, although more tuning is required.

Scatter Plot of Dataset With Clusters Identified Using DBSCAN Clustering

K-Means

K-Means Clustering may be the most widely known clustering algorithm and involves assigning examples to clusters in an effort to minimize the variance within each cluster.

The main purpose of this paper is to describe a process for partitioning an N-dimensional population into k sets on the basis of a sample. The process, which is called 'k-means,' appears to give partitions which are reasonably efficient in the sense of within-class variance.

— Some methods for classification and analysis of multivariate observations, 1967.

The technique is described here:

- k-means clustering, Wikipedia.

It is implemented via the KMeans class and the main configuration to tune is the "*n_clusters*" hyperparameter set to the estimated number of clusters in the data.

The complete example is listed below.

```
1 # k-means clustering
2 from numpy import unique
3 from numpy import where
4 from sklearn.datasets import make_classification
5 from sklearn.cluster import KMeans
6 from matplotlib import pyplot
```

```
# define dataset
X, _ = make_classification(n_samples=1000, n_features=2, n_informative=2,
n_redundant=0, n_clusters_per_class=1, random_state=4)
# define the model
model = KMeans(n_clusters=2)
# fit the model
model.fit(X)
# assign a cluster to each example
yhat = model.predict(X)
# retrieve unique clusters
clusters = unique(yhat)
# create scatter plot for samples from each cluster
for cluster in clusters:
 # get row indexes for samples with this cluster
 row_ix = where(yhat == cluster)
 # create scatter of these samples
 pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
# show the plot
pyplot.show()
```

Running the example fits the model on the training dataset and predicts a cluster for each example in the dataset. A scatter plot is then created with points colored by their assigned cluster.

In this case, a reasonable grouping is found, although the unequal equal variance in each dimension makes the method less suited to this dataset.

Scatter Plot of Dataset With Clusters Identified Using K-Means Clustering

Mini-Batch K-Means

Mini-Batch K-Means is a modified version of k-means that makes updates to the cluster centroids using mini-batches of samples rather than the entire dataset, which can make it faster for large datasets, and perhaps more robust to statistical noise.

... we propose the use of mini-batch optimization for k-means clustering. This reduces computation cost by orders of magnitude compared to the classic batch algorithm while yielding significantly better solutions than online stochastic gradient descent.

— Web-Scale K-Means Clustering, 2010.

The technique is described in the paper:

- Web-Scale K-Means Clustering, 2010.

It is implemented via the MiniBatchKMeans class and the main configuration to tune is the "*n_clusters*" hyperparameter set to the estimated number of clusters in the data.

The complete example is listed below.

```
# mini-batch k-means clustering
from numpy import unique
from numpy import where
from sklearn.datasets import make_classification
from sklearn.cluster import MiniBatchKMeans
from matplotlib import pyplot
# define dataset
X, _ = make_classification(n_samples=1000, n_features=2, n_informative=2,
n_redundant=0, n_clusters_per_class=1, random_state=4)
# define the model
model = MiniBatchKMeans(n_clusters=2)
# fit the model
```

```
model.fit(X)
# assign a cluster to each example
yhat = model.predict(X)
# retrieve unique clusters
clusters = unique(yhat)
# create scatter plot for samples from each cluster
for cluster in clusters:
 # get row indexes for samples with this cluster
 row_ix = where(yhat == cluster)
 # create scatter of these samples
 pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
# show the plot
pyplot.show()
```

Running the example fits the model on the training dataset and predicts a cluster for each example in the dataset. A scatter plot is then created with points colored by their assigned cluster.

In this case, a result equivalent to the standard k-means algorithm is found.

Scatter Plot of Dataset With Clusters Identified Using Mini-Batch K-Means Clustering

Mean Shift

Mean shift clustering involves finding and adapting centroids based on the density of examples in the feature space.

We prove for discrete data the convergence of a recursive mean shift procedure to the nearest stationary point of the underlying density function and thus its utility in detecting the modes of the density.

— Mean Shift: A robust approach toward feature space analysis, 2002.

The technique is described in the paper:

- Mean Shift: A robust approach toward feature space analysis, 2002.

It is implemented via the MeanShift class and the main configuration to tune is the "*bandwidth*" hyperparameter.

The complete example is listed below.

```
# mean shift clustering
from numpy import unique
from numpy import where
from sklearn.datasets import make_classification
from sklearn.cluster import MeanShift
from matplotlib import pyplot
# define dataset
X, _ = make_classification(n_samples=1000, n_features=2, n_informative=2,
n_redundant=0, n_clusters_per_class=1, random_state=4)
# define the model
model = MeanShift()
# fit model and predict clusters
yhat = model.fit_predict(X)
# retrieve unique clusters
clusters = unique(yhat)
# create scatter plot for samples from each cluster
for cluster in clusters:
 # get row indexes for samples with this cluster
 row_ix = where(yhat == cluster)
 # create scatter of these samples
 pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
# show the plot
```

```
pyplot.show()
```

Running the example fits the model on the training dataset and predicts a cluster for each example in the dataset. A scatter plot is then created with points colored by their assigned cluster.

In this case, a reasonable set of clusters are found in the data.

Scatter Plot of Dataset With Clusters Identified Using Mean Shift Clustering

OPTICS

OPTICS clustering (where OPTICS is short for Ordering Points To Identify the Clustering Structure) is a modified version of DBSCAN described above.

We introduce a new algorithm for the purpose of cluster analysis which does not produce a clustering of a data set explicitly; but instead creates an augmented ordering of the database representing its density-based clustering structure. This cluster-ordering contains information which is equivalent to the density-based clusterings corresponding to a broad range of parameter settings.

— OPTICS: ordering points to identify the clustering structure, 1999.

The technique is described in the paper:

- OPTICS: ordering points to identify the clustering structure, 1999.

It is implemented via the OPTICS class and the main configuration to tune is the "*eps*" and "*min_samples*" hyperparameters.

The complete example is listed below.

```
# optics clustering
from numpy import unique
from numpy import where
from sklearn.datasets import make_classification
from sklearn.cluster import OPTICS
from matplotlib import pyplot
```

```
# define dataset
X, _ = make_classification(n_samples=1000, n_features=2, n_informative=2,
n_redundant=0, n_clusters_per_class=1, random_state=4)
# define the model
model = OPTICS(eps=0.8, min_samples=10)
# fit model and predict clusters
yhat = model.fit_predict(X)
# retrieve unique clusters
clusters = unique(yhat)
# create scatter plot for samples from each cluster
for cluster in clusters:
 # get row indexes for samples with this cluster
 row_ix = where(yhat == cluster)
 # create scatter of these samples
 pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
# show the plot
pyplot.show()
```

Running the example fits the model on the training dataset and predicts a cluster for each example in the dataset. A scatter plot is then created with points colored by their assigned cluster.

In this case, I could not achieve a reasonable result on this dataset.

Scatter Plot of Dataset With Clusters Identified Using OPTICS Clustering

Spectral Clustering

Spectral Clustering is a general class of clustering methods, drawn from linear algebra.

A promising alternative that has recently emerged in a number of fields is to use spectral methods for clustering. Here, one uses the top eigenvectors of a matrix derived from the distance between points.

— On Spectral Clustering: Analysis and an algorithm, 2002.

The technique is described in the paper:

- On Spectral Clustering: Analysis and an algorithm, 2002.

It is implemented via the SpectralClustering class and the main Spectral Clustering is a general class of clustering methods, drawn from linear algebra. to tune is the "*n_clusters*" hyperparameter used to specify the estimated number of clusters in the data.

The complete example is listed below.

```
# spectral clustering
from numpy import unique
from numpy import where
from sklearn.datasets import make_classification
from sklearn.cluster import SpectralClustering
from matplotlib import pyplot
# define dataset
X, _ = make_classification(n_samples=1000, n_features=2, n_informative=2,
n_redundant=0, n_clusters_per_class=1, random_state=4)
# define the model
model = SpectralClustering(n_clusters=2)
# fit model and predict clusters
yhat = model.fit_predict(X)
# retrieve unique clusters
clusters = unique(yhat)
```

```
# create scatter plot for samples from each cluster
for cluster in clusters:
 # get row indexes for samples with this cluster
 row_ix = where(yhat == cluster)
 # create scatter of these samples
 pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
# show the plot
pyplot.show()
```

Running the example fits the model on the training dataset and predicts a cluster for each example in the dataset. A scatter plot is then created with points colored by their assigned cluster.

In this case, reasonable clusters were found.

Scatter Plot of Dataset With Clusters Identified Using Spectra Clustering Clustering

Gaussian Mixture Model

A Gaussian mixture model summarizes a multivariate probability density function with a mixture of Gaussian probability distributions as its name suggests.

For more on the model, see:

- Mixture model, Wikipedia.

It is implemented via the GaussianMixture class and the main configuration to tune is the "*n_clusters*" hyperparameter used to specify the estimated number of clusters in the data.

The complete example is listed below.

```
# gaussian mixture clustering
from numpy import unique
from numpy import where
from sklearn.datasets import make_classification
```

```
from sklearn.mixture import GaussianMixture
from matplotlib import pyplot
# define dataset
X, _ = make_classification(n_samples=1000, n_features=2, n_informative=2,
n_redundant=0, n_clusters_per_class=1, random_state=4)
# define the model
model = GaussianMixture(n_components=2)
# fit the model
model.fit(X)
# assign a cluster to each example
yhat = model.predict(X)
# retrieve unique clusters
clusters = unique(yhat)
# create scatter plot for samples from each cluster
for cluster in clusters:
 # get row indexes for samples with this cluster
 row_ix = where(yhat == cluster)
 # create scatter of these samples
 pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
# show the plot
pyplot.show()
```

Running the example fits the model on the training dataset and predicts a cluster for each example in the dataset. A scatter plot is then created with points colored by their assigned cluster.

In this case, we can see that the clusters were identified perfectly. This is not surprising given that the dataset was generated as a mixture of Gaussians.

Scatter Plot of Dataset With Clusters Identified Using Gaussian Mixture Clustering

Market Positioning

Influencing consumer perception of a brand or product in relation to rival brands

Over 1.8 million professionals use CFI to learn accounting, financial analysis, modeling and more. Start with a free account to explore 20+ always-free courses and hundreds of finance templates and cheat sheets.

What is Market Positioning?

Market Positioning refers to the ability to influence consumer perception regarding a brand or product relative to competitors. The objective of market positioning is to establish the image or identity of a brand or product so that consumers perceive it in a certain way.

For example:

- A handbag maker may position itself as a luxury status symbol
- A TV maker may position its TV as the most innovative and cutting-edge
- A fast-food restaurant chain may position itself as the provider of cheap meals

Types of Positioning Strategies

There are several types of positioning strategies. A few examples are positioning by:

- **Product attributes and benefits:** Associating your brand/product with certain characteristics or with certain beneficial value
- **Product price:** Associating your brand/product with competitive pricing
- **Product quality:** Associating your brand/product with high quality
- **Product use and application:** Associating your brand/product with a specific use

- **Competitors:** Making consumers think that your brand/product is better than that of your competitors

A Perceptual Map in Market Positioning

A perceptual map is used to show consumer perception of certain brands. The map allows you to identify how competitors are positioned relative to you and to identify opportunities in the marketplace.

An example of consumers perception of price and quality of brands in the automobile industry are mapped below:

This map is for illustrative educational purposes only.

How to Create an Effective Market Positioning Strategy?

Create a positioning statement that will serve to identify your business and how you want the brand to be perceived by consumers.

For example, the positioning statement of Volvo: "For upscale American families, Volvo is the family automobile that offers maximum safety."

1. Determine company uniqueness by comparing to competitors

Compare and contrast differences between your company and competitors to identify opportunities. Focus on your strengths and how they can exploit these opportunities.

2. Identify current market position

Identify your existing market position and how the new positioning will be beneficial in setting you apart from competitors.

3. Competitor positioning analysis

Identify the conditions of the marketplace and the amount of influence each competitor can have on each other.

4. Develop a positioning strategy

Through the preceding steps, you should achieve an understanding of what your company is, how your company is different from competitors, the conditions of

the marketplace, opportunities in the marketplace, and how your company can position itself.

What is Market Repositioning?

Market repositioning is when a company changes its existing brand or product status in the marketplace. Repositioning is usually done due to declining performance or major shifts in the environment.

Many companies, instead of repositioning, choose to launch a new product or brand because of the high cost and effort required to successfully reposition a brand or product.

Example of Market Repositioning

The example below describes Coca-Cola's repositioning of Mother Energy Drinks:

The Coca-Cola Company launched Mother Energy Drinks in 2006 into the Australian market. The launch campaign was professionally executed, and Coca-Cola was able to leverage its distribution channels to get the product into major retailers. However, the taste of Mother Energy Drink was subpar and repeat purchases were very low.

Coca-Cola was faced with a decision: to improve and reposition the product or withdraw it and introduce a new brand and product. The company ultimately decided to reposition the product due to already high brand awareness.

The biggest challenge faced by Coca-Cola was to persuade consumers to try the product again. The company changed the packaging, increased the size of the can, and improved the taste of the product. The relaunch of the product featured a new phrase – "New Mother, tastes nothing like the old one."

Ultimately, Coca-Cola was able to successfully reposition Mother Energy Drinks and the brand today competes with the two leading energy drinks in the market – V and Red Bull.

Applications of Data Mining

Data is a set of discrete objective facts about an event or a process that have little use by themselves unless converted into information. We have been collecting numerous data, from simple numerical measurements and text documents to more complex information such as spatial data, multimedia channels, and hypertext documents.

Nowadays, large quantities of data are being accumulated. The amount of data collected is said to be almost doubled every year. An extracting data or seeking knowledge from this massive data, data mining techniques are used. Data mining is used in almost all places where a large amount of data is stored and processed. For example, banks typically use 'data mining' to find out their prospective customers who could be interested in credit cards, personal loans, or insurance as well. Since banks have the transaction details and detailed profiles of their customers, they analyze all this data and try to find out patterns that help them predict that certain customers could be interested in personal loans, etc.

Basically, the motive behind mining data, whether commercial or scientific, is the same – the need to find useful information in data to enable better decision-making or a better understanding of the world around us.

"Extraction of interesting information or patterns from data in large databases is known as data mining."

According to William J.Frawley "Data mining or KDD(Knowledge Discovery in Databases) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data."

Technically, data mining is the computational process of analyzing data from different perspectives, dimensions, angles and categorizing/summarizing it into meaningful information. Data Mining can be applied to any type of data e.g. Data Warehouses, Transactional Databases, Relational Databases, Multimedia Databases, Spatial Databases, Time-series Databases, World Wide Web.

Data mining provides competitive advantages in the knowledge economy. It does this by providing the maximum knowledge needed to rapidly make valuable business decisions despite the enormous amounts of available data.

There are many measurable benefits that have been achieved in different application areas from data mining. So, let's discuss different applications of Data Mining:

Scientific Analysis: Scientific simulations are generating bulks of data every day. This includes data collected from nuclear laboratories, data about human psychology, etc. Data mining techniques are capable of the analysis of these data. Now we can capture and store more new data faster than we can analyze the old data already accumulated. Example of scientific analysis:

- Sequence analysis in bioinformatics
- Classification of astronomical objects
- Medical decision support.

Intrusion Detection: A network intrusion refers to any unauthorized activity on a digital network. Network intrusions often involve stealing valuable network resources. Data mining technique plays a vital role in searching intrusion detection, network attacks, and anomalies. These techniques help in selecting and refining useful and relevant information from large data sets. Data mining technique helps in classify relevant data for Intrusion Detection System. Intrusion Detection system generates alarms for the network traffic about the foreign invasions in the system. For example:

- Detect security violations
- Misuse Detection
- Anomaly Detection

Business Transactions: Every business industry is memorized for perpetuity. Such transactions are usually time-related and can be inter-business deals or intra-business operations. The effective and in-time use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world. Data mining helps to analyze these business transactions and identify marketing approaches and decision-making. Example :

- Direct mail targeting
- Stock trading
- Customer segmentation
- Churn prediction (Churn prediction is one of the most popular Big Data use cases in business)

Market Basket Analysis: Market Basket Analysis is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies and data mining techniques helps to achieve this analysis task. Example:

- Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
- Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Education: For analyzing the education sector, data mining uses Educational Data Mining (EDM) method. This method generates patterns that can be used both by

learners and educators. By using data mining EDM we can perform some educational task:

- Predicting students admission in higher education
- Predicting students profiling
- Predicting student performance
- Teachers teaching performance
- Curriculum development
- Predicting student placement opportunities

Research: A data mining technique can perform predictions, classification, clustering, associations, and grouping of data with perfection in the research area. Rules generated by data mining are unique to find results. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model. Example:

- Classification of uncertain data.
- Information-based clustering.
- Decision support system
- Web Mining
- Domain-driven data mining
- IoT (Internet of Things)and Cybersecurity
- Smart farming IoT(Internet of Things)

Healthcare and Insurance: A Pharmaceutical sector can examine its new deals force activity and their outcomes to improve the focusing of high-value physicians and figure out which promoting activities will have the best effect in the following

upcoming months, Whereas the Insurance sector, data mining can help to predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior of customers.

- Claims analysis i.e which medical procedures are claimed together.
- Identify successful medical therapies for different illnesses.
- Characterizes patient behavior to predict office visits.

Transportation: A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.

- Determine the distribution schedules among outlets.
- Analyze loading patterns.

Financial/Banking Sector: A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product.

- Credit card fraud detection.
- Identify 'Loyal' customers.
- Extraction of information related to customers.
- Determine credit card spending by customer groups.

UNIT 4

New Product Development (NPD) is the a set of design, engineering, and research processes which combine to create and launch a new product to market. Unlike regular product development, NPD is specifically about developing a brand new idea and seeing it through the entire product development process.

In today's competitive market, the ability to offer products that meet customers' needs and expectations has never been more important.

Customer requirements and behaviors, technology, and competition are changing rapidly, and businesses can not rely on existing products to stay ahead of the market. They need to innovate, and that means to develop and successfully launch new products.

But how do you go about finding a great, original idea and turning it into a marketable product?

In this article, we explain what new product development is and break it down into seven stages. We also discuss the best practices for developing your own process, along with some tips from product and marketing experts at Booking.com, Bumble, Typeform, EduMe, and Slite.

What is new product development?

New Product Development refers to the *complete* process of bringing a new product to market. This can apply to developing an entirely new product, improving an existing one to keep it attractive and competitive, or introducing an old product to a new market.

The emergence of new product development can be attributed to the needs of companies to maintain a competitive advantage in the market by introducing new products or innovating existing ones. While regular product development refers to building a product that already has a proof of concept, new product development focuses on developing an entirely new idea—from idea generation to development to launch.

The 7 stages of new product development

When it comes to new product development, each journey to a finished product is different. Although the product development process can vary from company

to company, it's possible to break it down into seven main stages. Let's have a look at them one by one.

1. Idea generation

Idea generation involves brainstorming for new product ideas or ways to improve an existing product. During product discovery, companies examine market trends, conduct product research, and dig deep into users' wants and needs to identify a problem and propose innovative solutions.

A SWOT Analysis is a framework for evaluating your Strengths, Weaknesses, Opportunities, and Threats. It can be a very effective way to identify the problematic areas of your product and understand where the greatest opportunities lie.

There are two primary sources of generating new ideas. Internal ideas come from different areas within the company—such as marketing, customer support, the sales team, or the technical department. External ideas come from outside sources, such as studying your competitors and, most importantly, feedback from your target audience.

Some methods you can use are:

- Conducting market analysis
- Working with product marketing and sales to check if your product's value is being positioned correctly
- Collecting user feedback with interviews, focus groups, surveys, and data analytics
- Running user tests to see how people are using your product and identify gaps and room for improvement

Ultimately, the goal of the idea generation stage is to come up with as many ideas as possible while focusing on delivering value to your customers.

2. Idea screening

This second step of new product development revolves around screening all your generated ideas and picking only the ones with the highest chance of success. Deciding which ideas to pursue and discard depends on many factors, including the expected benefits to your consumers, product improvements most needed, technical feasibility, or marketing potential.

The idea screening stage is best carried out within the company. Experts from different teams can help you check aspects such as the technical requirements, resources needed, and marketability of your idea.

Logic trees, like Teresa Torres' Opportunity Solution Tree, can help you visualize and chart the best path to your desired outcome. Ben Zacharias, Senior Product Manager at Booking.com, explains:

Logic trees are a valuable tool to try and make sure I have a structured understanding of the space I'm working in and making good decisions when choosing problem spaces to work on.

Ben Zacharias, Senior Product Manager at Booking.com

3. Concept development and testing

All ideas passing the screening stage are developed into concepts. A product concept is a detailed description or blueprint of your idea. It should indicate the target market for your product, the features and benefits of your solution that may appeal to your customers, and the proposed price for the product. A concept should also contain the estimated cost of designing, developing, and launching the product.

Developing alternative product concepts will help you determine how attractive each concept is to customers and select the one that would provide them the highest value.

Once you've developed your concepts, test each of them with a select group of consumers. Concept testing is a great way to validate product ideas with users before investing time and resources into building them.

Concepts are also often used for market validation. Before committing to developing a new product, share your concept with your prospective buyers to collect insights and gauge how viable the product idea would be in the target market.

4. Marketing strategy and business analysis

Now that you've selected the concept, it's time to put together an initial marketing strategy to introduce the product to the market and analyze the value of your solution from a business perspective.

- **The marketing strategy** serves to guide the positioning, pricing, and promotion of your new product. Once the marketing strategy is planned, product management can evaluate the business attractiveness of the product idea.
- **The business analysis** comprises a review of the sales forecasts, expected costs, and profit projections. If they satisfy the company's objectives, the product can move to the product development stage.

5. Product development

The product development stage consists of developing the product concept into a finished, marketable product. Your product development process and the stages you'll go through will depend on your company's preference for development, whether it's agile product development, waterfall, or another viable alternative.

This stage usually involves creating the prototype and testing it with users to see how they interact with it and collect feedback. Prototype testing allows product

teams to validate design decisions and uncover any flaws or usability issues before handing the designs to the development team.

We always test the main features with usability testing, first, to choose the best flow, and second, to iterate on the flow and make sure it's clear for the users. After usability testing, we can finalize the flow and prepare it for the developer handoff.

Regina Smirnova, Senior Product Designer at Bumble

Regina Smirnova, Senior Product Designer at Bumble, uses the IDEO Design Thinking approach when working on a new product. Design thinking brings together "what is desirable from a human point of view with what is technologically feasible and economically viable." As Regina explains, a successful Minimum Viable Product (MVP) lives at the intersection of desirability, feasibility, and viability.

6. Test marketing

At this stage, it's essential to stay in touch with customers and gather research data to understand what works and resonates with the target audience and what doesn't. Results can also be used to write the copy and the messaging around the launch.

Laure Albouy, Product Marketing Manager at Slite

Test marketing involves releasing the finished product to a sample market to evaluate its performance under the predetermined marketing strategy.

There are two testing methods you can employ:

- **Alpha testing** is software testing used to identify bugs before releasing the product to the public
- **Beta testing** is an opportunity for actual users to use the product and give their feedback about it

The goal of the test marketing stage is to validate the entire concept behind the new product and get ready to launch the product.

7. Product launch

A successful product launch is about setting your key results as early as possible, understanding how to track them, and then figuring out how to use the learnings to make changes or adapt.

Ian Booth, Senior Product Manager at EduMe

At this point, you're ready to introduce your new product to the market. Ensure your product, marketing, sales, and customer support teams are in place to guarantee a successful launch and monitor its performance.

To better understand how to prepare a go-to-market strategy, we spoke to Ganna Kryklii, Senior Product Marketing Manager at Typeform. Here are some essential elements to consider.

- **Customers**: Understand who will be making the final purchasing decisions and why they will be purchasing your product. Create buyer personas and identify their roles, objectives, and pain points.
- **Value proposition**: Identify what makes you different from the competition and why people should choose to buy your product
- **Messaging**: Determine how you will communicate your product's value to potential customers
- **Channels**: Pick the right marketing channels to promote your products, such as email marketing, social media, SEO, and more

You will need to constantly track and measure the success of your product launch and make adjustments if it doesn't achieve the desired goals.

Expert tips for creating a product development process

Looking to deliver successful products? Here are some tips from product and marketing experts at Bumble, Booking.com, EduMe, and Typeform on creating an effective product development process.

Align around the same vision

“I think the most important part is to align on the product vision and the company goal. Everyone in the team should understand where we are moving and what principles we follow during the product development process,” says Regina Smirnova, Senior Product Designer at Bumble.

Yet Laure Alboy, Product Marketing Manager at Slite, points out: "Sometimes it’s hard to be in the right conversation at the right time, and there are so many conversations to be a part of. I think one way to get visibility is to be part of strategic conversations and being the person who’s leading the questions around 'why are we doing this?'"

Ben Zacharias, Senior Product Manager at Booking.com, explains that having a clear understanding of the product development strategy and company goals makes it much easier to make good decisions and trade-offs along the way.

As a product team, you should focus on what you can work on to deliver the most impact. So the critical question you have to start with is: do you really understand what impact means for your team? Do you know your overarching goal and how you're contributing to a broader business/product strategy?

Ben Zacharias, Senior Product Manager at Booking.com

Use roadmaps, backlogs, recurring meetings, and syncs, but keep communicating with your team. “At EduMe, we communicate our vision all the time,” explains Ian Booth, Senior Product Manager at EduMe. "Always focus on the value that you're bringing and communicate it constantly."

Understand your customers’ needs

At every stage of the product development process, there is one critical driving factor: the customer. Identify what your customers need, which features would help them the most, and how to make your product appealing to them.

The user voice and the collaboration with customers is something that's really part of my routine. We're not talking about the feature or the product. We're talking about the solution for the problem that the customer has. So it's less feature-oriented and more benefit-oriented.

Ganna Kryklii, Senior Product Marketing Manager at Typeform

Collecting product feedback and insights helps you ensure that the end product meets their expectations, solves their problems, and fulfills their needs.

When it comes to making and validating decisions, Ian points out that it's always best to have qualitative data alongside quantitative information. You can use product surveys, customer interviews, market research, but make sure you back up those insights with behavioral data on how users use the product.

Build a strong team

Product development is a creative process at its core. Better results often come from teams being able to create a process together that works for that specific group.

Ben Zacharias, Senior Product Manager at Booking.com

"I think the foundational parts of a great product development process are often the intangible, human elements that help create motivation, focus, and impact," says Ben. Supportive leadership, clear direction, an open and high empathy culture, and a learning mindset are crucial to building productive teams and great products.

Also, each team is different. So, it's essential to create a supportive and flexible environment that allows you to identify which product development process works best for you and your organization.

How long does it take to develop a new product and get it to market?

Developing new products is a time-consuming activity, especially if you want to deliver a high-quality product. How much time you need will depend on several

factors, including the complexity of the product, the industry, the company stage, and the resources available.

Whether you're developing a new product or improving an existing one, having a well-defined process will enable you and your team to achieve greater speed and efficiency and increase the likelihood of success. You can now choose from a wide array of product management tools to streamline your product development process and achieve better team collaboration.

Final notes

Having an efficient new product development process is essential to bringing your final product to the market. Hopefully, by following these steps and expert tips and adapting them to your business strategy, you can build a successful product.

Using Python for Product Development

Python is also a major tool for mobile application development, frequently selected to develop APIs to ensure compatibility with diverse operating systems. There are also a variety of libraries used in creating the GUI of an application or game. Popular domains such as Instagram, Uber, Dropbox, and Pinterest use Python's features available through open source libraries to create their apps.

Any company that's looking for the best way to bring a smart app or game to life that will remain engaging and relevant in an ever-changing digital landscape is sure to find Python to be the answer to their question. Easily one of the most robust, versatile, and easy-to-learn programming languages ever developed, developers can create virtually anything with Python, from sophisticated role-playing games to web apps and mobile apps.

Advantages and Disadvantages of Python

- **Simplicity & Rapid Development**

Developing applications with Python can be done five to ten times quicker than with C/C++ as well as three to five times faster than using Java. Well-structured,

direct, easy to learn, and make use of, concise yet expressive, versatile, and uncluttered are what characterize the Python programming language. Python focuses entirely on code readability and visibility. This means developers can easily read, understand, and modify existing code and spend less time and effort coding. These benefits make Python one of the best languages for startups because getting to market quickly often translates into a competitive advantage and faster ROI.

- **Feature-Packed & Versatile**

Python can be made use of to build almost anything, from websites to AI-based solutions to digital and scientific applications. The language provides many standard libraries and features that meet almost any programming need, which again results in rapid development.

- **Security-Driven**

Cybersecurity threats are increasing rapidly, which is why companies are looking for ways to ensure maximum security. Python proves to be an excellent choice for those who focus on data security. It provides authentication and authorization, email verification, and password reset features. Additionally, Python and its frameworks provide different mechanisms to address and eliminate security-related challenges such as –

- CSRF (cross-site request forgery),
    - XSS (cross-site scripting),
    - SQL injection, and
    - clickjacking.
- **No Dearth Of Python Talent & Support**

Having a robust, supportive community is crucial for any language, as community members share experiences and help each other, create and upgrade features, update documentation, and resolve issues. A large community, in short, makes the language grow and develop faster. Additionally, a large community also implies that talented developers are easy to come by.

- **Handles AI & ML Operations Seamlessly**

How do popular eCommerce platforms manage to offer you the right product recommendations? And how do OTT platforms know what you want to watch/listen to even before you provide any inputs? It's all through machine learning. It just so happens that Python happens to be the best for machine learning and artificial intelligence programming. Among its key machine learning use cases are –

- content customization,
    - consumer recommendations,
    - image recognition,
    - machine translation,
    - speech recognition,
    - fraud detection, and
    - user behavior analysis.

Python's Drawbacks

Despite all these advantages that the Python programming language offers developers and companies, it might not be the best choice for a project. There are times when Python simply isn't the language for the job. There are specific applications where languages such as C, C++, and Java prove to be better options. Highlighted below are the drawbacks of the Python programming language.

- **Python Isn't Necessarily Recognized For Its Speed**

Python is an interpreted language. What this means is that everything you do goes through an extra layer so that the target machine can interpret and execute the code. It's akin to speaking talking to someone who speaks a different language through a translator.

When comparing Python with C (which is a compiled programming language), it's true that Python code runs slower if the execution time is measured. Still, Python's flexibility serves as a counterbalance. The language provides practical ways to solve challenges and offers dynamic typing, which assists with rapid development. AS a result, Python beats C in terms of development time, which is often far more crucial than runtime performance, as less development time translates into lower costs and faster time to market.

When dealing with algorithms that need to run fast (e.g., sorting, searching, or just doing something on embedded hardware that needs to meet a specific speed requirement), it's best to avoid Python altogether.

- **Not The Best For Mobile App Development**

Although Python is powerful for both simple and complex web and desktop applications, it does not win any medals for mobile application development. There are several ways to make Python more mobile-friendly. For instance, using modules such as Kiwi. However, when it comes to deploying, using, and updating Python mobile apps, they normally fail because Python is not made for mobile app development.

Building Python apps for mobile phones is a daunting task, which is why Python is rarely made use of for mobile app development. Mobile app developers should consider using Swift for iOS app development and either Kotlin or Java for Android app development.

- **Not The Best For Game Development & 3D rendering**

3D rendering is computationally heavy. That's why if game developers use a relatively slower language like Python for 3D rendering, their rendering won't be as efficient and fast as most gamers expect. Also, developers cannot use Python with Unity, a popular game development engine.

Graphics require multithreading, movement, and raw machine power, which is why games developed with Python won't run as fast as expected. Going for C++ or C# for game development over Python is always recommended.

- **High Memory Usage**

Owing to its structure, Python requires a lot of memory. Python is not a developer's best choice if his/her project is full of memory-intensive tasks and they have tight memory and performance constraints.

Popular Use Cases

- **Instagram**

Instagram is an application that's inconceivable without the use of Python. It used to be a simple site using Django. Django is essentially a high-level Python web framework. Therefore, in a nutshell, Instagram is a great example of one of the best Python sites. It has 400 million daily active users. Instagram allows users to take photos, edit them and share them. This shows that Python is essentially scalable.

- **Spotify**

Spotify, the world's leading music streaming platform, is entirely Python-based. It completely transformed the way music is listened to. Nowadays, there is no need for downloading mp3s or surfing Torrent sites to enjoy the latest tunes. Spotify, which uses Python, is perhaps the best platform for these purposes. Through the

use of Python, Spotify can not only handle features like Radio and Discover but also run powerful ML-based music recommendation engines.

- **Dropbox**

Dropbox is perhaps the most popular of all Python-based web platforms. With a valuation of $10 billion back in 2014, which includes their polished, user-friendly desktop client, the creators of Dropbox have created a product that is recognized for its incredible user-friendliness. Dropbox has opened up much of its code, and it's written primarily in Python. Dropbox can be installed on the Windows, Mac, and Linux operating systems, which is why it's easy for programmers to see how the usage of the Python programming language has made it incredibly portable. Many third-party open source libraries are written in Python, and many of Dropbox's projects are hosted on the Github repository.

Product Pricing

Table of Contents

After developing a product, companies need to figure out how much to charge customers before executing their go-to-market (GTM) strategy.

Product pricing is much more complicated than it looks—price optimization involves various internal and external factors, such as setting a price that maximizes profits, taking into account customer demand, market and competition data, and development costs.

What is Product Pricing?

Product pricing is the process of setting a selling price for a product or service that considers all costs associated with producing and selling it, as well as what customers are willing to pay.

The goal of product pricing should be to match the value of the product or service with its cost and customer demand so that the company can maximize profits while providing competitive prices.

Several factors work together when setting a price for a product or service:

- **Overhead Costs:** This includes all the expenses related to producing a product or providing a service, such as labor costs, marketing/advertising costs, and shipping costs.
- **Competition:** What are other businesses charging for similar products or services in the same market? Companies should consider their competitors' pricing structures when setting prices for their own products and services.
- **Price Sensitivity (Demand Elasticity):** How sensitive are customers to changes in prices? If demand for a product or service decreases too much with an increase in price, it may not be profitable to charge more.
- **Value Proposition:** The value of a product or service should reflect its price. Companies should consider the extra features, quality, customer service, and brand value that customers gain when they purchase their products or services.

- **Pricing Strategy:** Loss leaders, price skimming, penetration pricing, premium pricing—there are many different strategies to consider when setting prices for products and services. Companies should determine the most appropriate strategy based on their goals and market situations.
- **Product Complexity:** Businesses with complex products (e.g., software with multiple features) may need to consider different pricing models depending on the complexity of their products. For example, subscription-based pricing or pay-as-you-go plans may be suitable for software products.

On the surface, product pricing seems simple—just set a price. But in practice, the difference of even a few dollars can sway customers. This is especially true when companies are competing against each other, as customers will look for the lowest price available.

Synonyms

- **Pricing Model** – The strategy or structure used to set prices for products.
- **Product Pricing Method** – A systematic approach to setting prices for products.
- **Product Pricing Strategy** – A set of tactics and strategies used to optimize prices for products.

Top Product Pricing Methods

The exact method a business uses to find a product price will vary depending on the factors listed above.

Some organizations use a combination of pricing models. But there are some common pricing models that companies use.

Here are a few:

Value-Based Pricing

Value-based price is a pricing strategy that bases prices on the value customers receive from products or services rather than their production costs.

Rather than focusing on competitors, companies that use value-based pricing largely base their retail price on the value customers attribute to the product or service (i.e., what they are willing to pay).

In theory, money can be left on the table when setting prices too low—value-based pricing seeks to capture that value. When correctly executed, this approach can drastically improve profitability by allowing for higher prices without sacrificing sales volumes.

The main drawback to the value-based method is that it requires a lot of market research to understand the customer's value perception, which may be too low to justify the internal costs.

Value-based pricing works best for companies that:

- Sell unique products with a high perceived value.
- Sell lightweight and efficient products that exponentially boost revenue growth for their clients.
- Have a limited number of competitors in the market.
- Have substantial market research data about customer perceptions.

Competitor-Based Pricing

Competitor-based pricing is the opposite of value-based pricing. It's a pricing method that bases prices on those of competitors in the same market.

Companies should consider their competitors' pricing strategies when setting prices for their own products and services.

It's important to note that competitor-based pricing isn't about having the lowest price, but rather finding a balance between charging what customers are willing to pay and what the company needs in order to make a profit.

To be truly successful with this pricing model, the organization can theoretically charge the same amount as its competitors but sell a more efficient product.

Competitive pricing is well suited to companies that:

- Sell products with similar features across a competitive market.
- Have a limited number of customers and aren't able to charge higher prices due to competition.
- Recently entered the market and don't have much customer data.
- Have the resources needed to track competitor pricing strategies.
- Are comfortable charging the same price as competitors, but offering a better product or service.

Cost-Plus Pricing

The cost-plus pricing strategy is a product pricing method that uses the production costs of a product or service as the baseline and adds an additional percentage (the "plus") to determine the final price.

In other words, companies figure out their costs for producing a product, then add a profit margin on top of that cost. This helps them cover overhead expenses, account for risk, and gain a profit.

Cost-based pricing strategies work best when the product or service has:

- High production costs (e.g., industrial goods).
- Few competitors in the market.
- A large customer base that can absorb price increases.

Market-Oriented Pricing

Market-oriented pricing is a strategy where companies set their prices based on the current market trends and customer preferences.

This method is closely tied to competitor-based pricing, but it focuses more on understanding customers' needs and behaviors than tracking competitors.

Companies that use this approach try to identify what customers are willing to pay for products or services and set prices accordingly. It's important to note that this method isn't solely about setting prices low or high—it's also about understanding how customers respond to different pricing strategies and using that information to make informed decisions.

Market-oriented pricing works well when businesses:

- Have a large customer base with diverse needs and preferences.
- Have a good understanding of customer needs and behaviors.
- Are able to respond quickly to changes in the market.
- Are comfortable with risks such as pricing experiments and variable pricing.

Dynamic Pricing

Dynamic pricing is an approach where companies adjust their prices in real time based on market demand, customer behavior, and other factors. This type of pricing strategy relies heavily on analytics and data to make sure prices are set correctly.

In terms of revenue optimization, dynamic pricing is one of the most effective strategies. It allows companies to adjust their prices in real time and maximize profits while still offering customers a good value.

Airlines, for example, use dynamic pricing to adjust their fares according to customer demand. By doing this, they can make sure that the prices are in line with market conditions and maximize profit potential.

Still, it is important to note that airlines can only get away with demand pricing because they sell a commodity product that customers have few—if any—lower-cost alternatives.

Dynamic pricing is most suitable for companies that:

- Have access to large amounts of customer data.

- Are able to respond quickly to changes in the market.
- Can create models or algorithms to analyze customer data and adjust prices accordingly.
- Are comfortable with risk.
- Have enough leverage to guarantee that customers will accept the pricing changes.

Factors to Consider in Product Pricing

To develop a competitive price that is also profitable, companies need to consider several factors, including costs, demand, and their target customer.

Costs

For a business to stay alive, it needs to continuously generate revenue. And that revenue needs to be greater than the cost of making and selling a product or service.

Depending on the product and company structure, there are several costs they may incur:

- Product research and development (R&D)
- Continued maintenance (for software products)
- Production costs (raw materials, labor, utilities)
- Shipping and distribution
- Marketing and advertising
- Sales and customer success
- Rent and utilities

Plenty of businesses operate remotely, using a contract workforce, and outsource production and other tasks to save on costs. Some Software-as-a-Service (SaaS) and ecommerce companies even operate as one or two people.

Either way, it's important to calculate all costs before setting a price that meets business goals.

Market Demand

Figuring out how much to charge is the easy part. Determining how much customers are willing to pay and whether there is enough demand for the product is a much more complicated equation.

There are a few best practices to keep in mind when evaluating market conditions:

- Understand the value that the product or service provides to customers.
- Analyze customer demographics, needs, and behaviors.
- Research competitors' pricing strategies.
- Take into account any seasonal changes in demand.

If the product or service is new, it's also critical to test different pricing models with a select group of customers to understand the market better.

Target Audience

An organization's ideal customer profile (ICP) is a key factor in product pricing. It is a detailed description of an ideal customer, based on customer data and market research, that helps businesses target the right demographic and provide more tailored prices to maximize profits.

Once the target audience is identified, companies can tailor their prices accordingly. They may offer discounts for larger orders or product bundles, volume-based pricing (i.e., enterprise plans), or recurring revenue models with different tiers (i.e., SaaS companies).

It is also important to note that customers may have different price points depending on their location, age, income level, and other demographic details.

Market Prices

Unless a business is the first of its kind (which is almost never the case), there will always be competitors vying for market share. Once businesses understand the value they offer to customers, they can set prices that are competitive within their industry.

Companies should also keep in mind that the price of a product or service doesn't have to be fixed. They can experiment with different pricing models and sales discounts as well as increase or decrease prices as needed.

Ideal Profit Margin

The profit margin is the amount of profit a business makes after subtracting all expenses from revenue. A healthy profit margin ensures that the business can stay in operation, pay its employees, and generate returns for its investors.

There are two profit margins businesses need to consider: gross and net profit margin.

- **Gross Profit Margin:** The amount of revenue left after subtracting production costs from sales revenue.
- **Net Profit Margin:** The amount of revenue left after subtracting all expenses (including overhead, marketing, and other operating costs).

The ideal level of profitability depends on factors like the industry, company size, and type of product or service being offered. For example, ecommerce businesses typically operate at a profit margin of around 10%.

SaaS companies typically operate using the Rule of 40, meaning the growth rate and profit should add up to about 40%.

Distribution Channels

When it comes to product pricing, businesses need to consider how they will distribute the product or service. Distribution channels include retail stores, ecommerce websites, digital retailers, direct-to-consumer (D2C) platforms, and more.

Each channel has different costs associated with it that must be taken into account when setting prices.

For instance, a business selling on Amazon will need to pay the retail giant's commission fee (usually around 15%).

Software companies have fewer required distribution channels, but they often need sales reps and a robust sales stack to manage the process.

Technology to Manage Product Pricing

There are numerous software tools available to help businesses manage their product pricing. These solutions typically offer features like price tracking, optimization, and analytics so businesses can make smarter pricing decisions.

Pricing Engine

A pricing engine is a powerful accounting software tool that helps companies identify profitable prices for their products and services.

This system works by taking into account various conditions to determine the best cost for each item, ultimately aiming to help businesses maximize earnings through price optimization.

Businesses use pricing engines to automate pricing so they can focus more on customer engagement, product innovation, and growth.

Pricing Software

Pricing software helps businesses track and analyze pricing data to ensure that the prices they set are competitive. It also provides insights into customer behavior, enabling companies to make better decisions about their product pricing strategies.

ERP

Enterprise resource planning (ERP) isn't exactly product pricing software, but it helps businesses manage their operations efficiently. This includes automated processes like inventory tracking, sales order processing, and financial reporting.

All of these features can help inform product pricing decisions and optimize profitability.

Organizations use ERP systems to get the clearest possible view of their variable and fixed costs, and one of the residual benefits is they can also determine the optimal pricing strategies.

CPQ

CPQ software (configure, price, and quote) helps streamline the sales process from start to finish. It enables businesses to configure their products or services, generate quotes quickly, and identify the best prices for certain customers based on their needs.

In terms of product pricing, CPQ helps businesses identify the right price for their offerings so they can increase sales while still generating a healthy profit. Then, it enables them to generate proposals and quotes quickly so they don't miss out on potential sales opportunities Retail Management - Space

Space management is one of the crucial challenges faced by today's retail managers. A well-organized shopping place increases productivity of inventory, enhances customers' shopping experience, reduces operating costs, and increases financial performance of the retail store. It also elevates the chances of customer loyalty.

Let us see, how space management is important and how retailers manage it.

What is Space Management?

It is the process of managing the floor space adequately to facilitate the customers and to increase the sale. Since store space is a limited resource, it needs to be used wisely.

Space management is very crucial in retail as the sales volume and gross profitability depends on the amount of space used to generate those sales.

Optimum Space Use

While allocating the space to various products, the managers need to consider the following points −

- Product Category −
    - **Profit builders** − High profit margins-low sales products. Allocate quality space rather than quantity.
    - **Star performers** − Products exceeding sales and profit margins. Allocate large amount of quality space.
    - **Space wasters** − Low sales-low profit margins products. Put them at the top or bottom of shelves.
    - **Traffic builders** − High sales-low profit margins products. These products need to be displayed close to impulse products.
- Size, shape, and weight of the product.
- Product adjacencies − It means which products can coexist on display?
- Product life on the shelf.

Retail Floor Space

Here are the steps to take into consideration for using floor space effectively −

- Measure the total area of space available.
- Divide this area into selling and non-selling areas such as aisle, storage, promotional displays, customer support cell, (trial rooms in case of clothing retail) and billing counters.
- Create a **Planogram**, a pictorial diagram that depicts how and where to place specific retail products on shelves or displays in order to increase customer purchases.
- Allocate the selling space to each product category. Determine the amount of space for a particular category by considering historical and forecasted sales data. Determine the space for billing counter by referring historical

customer volume data. In case of clothing retail, allocate a separate space for trial rooms that is near the product display but away from the billing area.

- Determine the location of the product categories within the space. This helps the customers to locate the required product easily.
- Decide product adjacencies logically. This facilitates multiple product purchase. For example, pasta sauces and spices are kept near raw pasta packets.
- Make use of irregular shaped corner space wisely. Some products such as domestic cleaning devices or garden furniture can stand in a corner.
- Allocate space for promotional displays and schemes facing towards road to notify and attract the customers. Use glass walls or doors wisely for promotion.

Store Layout and Design

Customer buying behavior is an important point of consideration while designing store layout. The objectives of store layout and design are −

- It should attract customers.
- It should help the customers to locate the products effortlessly.
- It should help the customers spend longer time in the store.
- It should motivate customers to make unplanned, impulsive purchases.
- It should influence the customers' buying behavior.

Store Layout Formats

The retail store layouts are designed in way to use the space efficiently. There are broadly three popular layouts for retail stores −

Grid Layout − Mainly used in grocery stores.

Loop Layout − Used in malls and departmental stores.

Free Layout − Followed mainly in luxury retail or fashion stores.

AD

Store Design

Both internal and external factors matter when it comes to store design.

Interior Design

The store interior is the area where customers actually look for products and make purchases. It directly contributes to influence customer decision making. In includes the following −

- Clear and adequate walking space, separate from product display area.
- Free standing displays: Fixtures, rotary displays, or mannequins installed to attract customers' attention and bring them to the store.
- End caps: These displays at the end of the aisles can be used to display promotional offers.
- Windows and doors can provide visual messages about merchandise on sale.
- Proper lighting at the product display. For example, jewelry retail needs more acute lighting.
- Relevant signage with readable typefaces and limited text for product categories, for promotional schemes, and at Point of Sale (POS) that guides customers' decision-making process. It can also include hanging signage for enhancing visibility.
- Sitting area for a few differently abled people or senior citizens.

Exterior Design

This area outside the store is as much important as the interior of the store. It communicates with the customer on who the retailer is and what it stands for. The exterior includes −

- Name of the store, which tells the world that it exists. It can be a plain painted board or as fancy as an aesthetically designed digital board of the outlet.
- The store entrance: Standard or automatic, glass, wood, or metal? Width of the entrance.
- The cleanliness of the area around the store.
- The aesthetics used to draw the customers inside the store.

Top 7 Types of Attribution Models for You to Try in 2023

Understanding the impact of marketing efforts on business outcomes has been an important focus area for marketing ever since the beginning. Marketing attribution models have been a lifesaver for businesses in this respect.

With attribution models, it is now possible to identify touchpoints in the customer journey and attribute the credit for the conversion to appropriate marketing channels. Marketers can optimize their approach and focus on the channels and tactics that drive maximum ROI.

However, there are several types of attribution models available. These attribution models provide you with answers to different questions, and depending on what type of question you want to answer, the models change. So you need to know each to understand what works best for your business and use case.

In this blog, **you will learn the different types of attribution models** with examples and graphs to understand when and how to use each one.

TL;DR:

- Attribution models help businesses understand the impact of their marketing efforts by assigning credit to different touchpoints in the customer journey.
- There are two main types of attribution models: single-touch attribution models and multi-touch attribution models.
- Single-touch attribution models assign 100% credit to a single touchpoint, such as first-touch, last-touch, or last non-direct touch.
- Multi-touch attribution models consider all/ most touchpoints. Multitouch attribution models include linear, U-shaped, time decay, or W-shaped models.
- Each attribution model has its own strengths and limitations, and the choice depends on the specific use case. For instance, first-touch models are effective for assessing initial brand awareness, while U-shaped models assess the initial engagement and final conversion stages.

What are attribution models?

Attribution models are frameworks that help analyze the customer journey and assign credit to the various touchpoints prior to the conversion. The method for assigning the credit is different for each attribution model depending on either the position of the touchpoint in the customer journey or a data-driven estimation of the significance of that touchpoint.

Additionally, businesses may need to configure these attribution models to suit their unique circumstances - such as considering an attribution window of, say, 60 days or 365 days depending on their sales cycle or performing the attribution analysis at a contact or opportunity, or account level depending on their sales motion.

With the help of these models, marketers are able to identify channels and tactics that drive more conversions and revenue, driving higher ROI for the business.

The following are some of the main reasons why attribution modeling is important.

- They provide insight into channels and campaigns that drive conversions and revenue
- They help plan and distribute spending to the right marketing channels
- Also, they help us identify the most influential channels and campaigns for each stage of the marketing and sales funnel.

There are different types of attribution model available for marketers, and we will dive into each in the coming sections.

Categories of Attribution Models

Before delving into how some of the most popular attribution models work, it's worth understanding the mechanics of attribution modeling. A general categorization of attribution models would include two types. They are -

1. **Rule-based attribution models**
2. **Data-driven attribution models**.

1. Rule-based attribution models

These models use predetermined rules for assigning attribution credits to touchpoints. These pre-defined rules determine the weightage or credit for a touchpoint primarily based on its position in the customer journey. Hence, these models are also called Position based Attribution Models.

In addition to the position, you can also define custom logic to assign differential weights based on the seniority of the customer representative involved in the

touchpoint (say Director and above gets higher weight) as well as the amount of effort expended by the buyer in that interaction (attending a webinar required higher effort from a buyer than clicking on a paid search ad).

2. Data-driven attribution models

These models assign attribution credits to touchpoints based on an algorithmic estimation of the significance of that touchpoint in converting the customer. Some of the popular algorithmic techniques are Markov Chain models and Shapely value-based models. Whilst data-driven attribution is seen as the north star of Multi-Touch Attribution, they are also more expensive to compute, require a large volume of conversions and touchpoints not to be biased, and are harder to debug.

Whilst each approach has its own pros and cons, a combination of these models may be leveraged to identify marketing leakage and improve ROI.

What are the different types of attribution models?

Single-Touch attribution models

Single-touch attribution, also known as single source attribution, assigns 100% of attribution credit to a single touchpoint (or a single source). While they're easy to use and interpret, single-touch attribution models may skew your results and affect ROI if used in all situations.

Some of the most common types of single-touch attribution models include:

1. First-Touch Attribution

In this type of attribution model, your customer's first touch-point — whether that be an ad campaign impression, content interaction, or anything else— is deemed the most critical touchpoint in their journey. Hence, this interaction is assigned 100% of the attribution credit.

For instance, let's assume that you're in the market for a project management software and come across an advert for one that catches your attention. The ad prompts you to visit the company's website.

After landing on their "features" page, you follow through with more research and come across the company's weekly blog — before finally signing up for a demo. In this case, the advert you clicked on is your first impression of the brand and product. Hence, a first-touch attribution model would reward the advert with 100% of the attribution credits.

The First Touch attribution model is most effective in identifying the channels or campaigns that drove your brand's initial awareness amongst your prospects. This would work best for businesses with low sales cycles or a PLG flow, or if you are trying to assess the effectiveness of only Branding Campaigns and Top of the funnel content.

In a normal B2B sales process that stretches over weeks and months, it would be presumptuous to assign 100% of the credit to the very first touchpoint.

2. Last-Touch Attribution

In a similar vein to first-touch attribution, a last-touch attribution model assigns 100% of the attribution credits to the touchpoint closest to a customer's conversion milestone.

This would imply that the last impression made on the customer before their decision to convert was the most prominent in their journey.

Continuing with the previous example of the PMS, the blog piece you come across before scheduling a demo would be the last touch. And so, out of all the touch points that influenced your decision to sign up for a demo, attribution credits will be solely assigned to the final one.

3. Last Non-Direct Touch Attribution:

This model assigns 100% attribution credit to the last non-direct touchpoint. A non-direct touchpoint is an interaction that is guided by a specific source the business sets up (like an ad, email campaign, newsletter, etc.).

When your website traffic doesn't come from a known source, they are considered direct traffic (traffic that came from prospects directly entering the company URL into the browser, for example).

Let's assume that a lead interacted with your brand 5 times, each touchpoint is as given below.

- Touchpoint 1 - Prospect clicks on a PPC ad
- Touchpoint 2 - Prospect arrives at your site's landing page
- Touchpoint 3 - Prospect subscribes to your newsletter
- Touchpoint 4 - A week later, your prospect clicks on a newsletter campaign
- Touchpoint 5 - Prospect directly visits the website and initiates a free trial before purchasing a subscription

Touchpoints 1, 2, 3, 4, and 5 constitute all the prospect's interactions with your brand that led to them purchasing your product. Keep in mind that, in reality, businesses deal with numerous prospects interacting with several touchpoints, making the process of mapping the customer journey far more convoluted.

So if we consider the above-given example, this model would assign 100% sales credit to touchpoint 4 or the newsletter campaign clicked on, as that was the last

non-direct source before the sale. This model assumes that every interaction is a consequence of the non-direct campaign, hence making it the most influential.

Is Single-Touch attribution an INEFFECTIVE model?

Many businesses and marketing aficionados are of the opinion that single-touch attribution is not an effective model on its own. It is often considered to be a one-dimensional approach that fails to faithfully represent a customer's conversion journey down the funnel.

As we have discussed, while single-touch models may have their own relevant use cases (like for products with shorter sales cycles), it may not be as effective in identifying the most influential touch-point in a B2B customer journey.

If big data in marketing has proved anything, it's that customer journeys can be non-linear, sophisticated paths spanning several channels and mediums. Assigning 100% of the credit to a single touchpoint will rarely be sufficient.

Multi-Touch attribution models

Multi-touch attribution modeling is the holy grail of marketing attribution. As customers' buying patterns evolve and become increasingly scattered, a model that can track and account for all these interactions would be more representative of the buying journey.

A multi-touch attribution model accounts for all the touchpoints encountered in a customer's conversion journey. It's a holistic view that helps paint a substantially better picture of patterns and behavior than single-touch models.

Remember to keep in mind that the goal of multi-touch attribution isn't just to map out customers' interactions. It is also employed to understand which touchpoints influence a customer the most, which touchpoints work in

conjunction with each other, and the relative probabilities of channel interactions among different customer paths.

With this established, there is still the issue of assigning credits to many touchpoints. To help illustrate multi-touch attribution better, here are a few of the most commonly used models:

4. Linear Attribution

A linear attribution model assigns attribution credits evenly among all touchpoints. While this model is far more illustrative than any of our single-touch attribution options, it's a relatively simplistic approach when compared to its nonlinear variants.

Let's assume that the total number of touchpoints in our PMS example is four: An advert, a blog, a review, and a retargeting campaign. Linear attribution would reward 25% of attribution credits to each of these touchpoints.

Of course, in reality, the number of touchpoints a B2B customer goes through is significantly higher — so the weights for each one are likely to be far smaller.

5. U-Shaped Attribution

The U-shaped model assigns attribution credits to all touchpoints — but assigns higher credits specifically to the first and last touchpoints. This would imply that your customer's first and last interactions prior to the conversion milestone are the two most valuable touch-points in their journey.

Consider the same four touch points as with the previous example (Ad, Blog, Review, and Retargeting campaign). This time, maybe 40% of the credits will be assigned to the first and last touch points each. The two touchpoints in-between will receive only 10% each as they are deemed less influential to the conversion decision.

The model laid out in a bar graph takes the shape of the letter 'U', hence the name.

6. Time Decay Attribution

Time decay attribution assigns attribution credits in an ascending cascade.

What this means is that each touchpoint is given progressively higher credit, with the first touchpoint having the least credit and the last touchpoint having the most. This is an effective tool in mapping out a customer's conversion journey.

The model works on the assumption that touchpoints closer to the conversion were far more influential than touchpoints further away from the conversion. Again, using our handy four touchpoint PMS example, a time decay model would assign attribution credits in this manner: 5% for the advert, 15% for the blog, 20% for the reviews page, and 60% for the retargeting campaign.

7. W-shaped attribution

This type of attribution model is similar to the U-shaped model we discussed earlier.

The first and the last touchpoints are also given importance in this model, just as in the U-shaped model. But during the middle of the sales funnel, if you generate quality leads, then that touchpoint is also considered influential. And therefore is given equal importance as that of the first and last touchpoint.

So, if there are 5 first touchpoints in total, the first, middle, and last touchpoints will be given 30% each and the rest only 5%.

To give you a clear-cut idea, take five touchpoints. For example, an advert, a blog, a case study, reviews, and finally, retargeting campaign.

A prospect got in touch with your business through an advertisement, prompted to read your blogs, where they decide to subscribe to your business's newsletter. Thereby generating a lead towards the middle of the process. The lead then

continued to follow up on their research by constantly staying in touch with the business through newsletters. And finally, the lead converts by signing up for a free trial. Following is an example of a graphical representation of the W-shaped attribution model for the given example.

Limitations of Attribution Models

Single-touch attribution models (like first-touch, last-touch, and list non-direct touch) are simple to implement but have several disadvantages. They oversimplify the customer journey by assigning credit to a single touchpoint, ignoring the contributions of other touchpoints. Similarly, these models also neglect the aggregate effect of multiple touchpoints over time. What results is inaccurate credit allocation, because the model disregards individual customer behavior and other factors.

On the other hand, multi-touch attribution models are definitely more complex because they work with complicated algorithms and technology. This often requires expert knowledge and pro- marketing knowledge of marketing software. The impressions from data can be misleading because of shortcomings like wrong assumptions and wrong weightage assigned to each marketing activity. To add on, while multi-touch attribution models are efficient for data- rich digital marketing campaigns, they are not equipped to measure external factors like word-of-mouth, seasonality or pricing.

Like single touch attribution models, multi-touch attribution models can also miss out on giving the full picture. Linear attribution models assume that all touchpoints have equal influence on customer behavior, which is not always the case. U-shaped, W-shaped and Time-Decay models run the risk of oversimplifying the customer journey since they assign more credit only to some touchpoints, while neglecting others. This could cost the model some valuable insights and

paint an incomplete picture. The time-decay attribution model considers the recency of the customers close to the conversion event, but it can still overlook the significance of earlier touchpoints.

Takeaway

Needless to say, all attribution models are not appropriate for all use cases. Different attribution models aid different types of marketing campaigns and can reveal different insights into the customer journey.

Attribution Model	How It Works	Use-cases
First-touch	The first touchpoint is assigned 100% of the attribution credit	First-touch attribution is most effective in identifying the channels or campaigns that drove your brand's initial awareness amongst your prospects. This model can help assess the impact of initial brand awareness efforts and gauge the success of activities like advertising campaigns.
Last-touch	The last touchpoint is assigned 100% of the conversion credit	This attribution mode is useful in cases where the final touchpoint is the most influential in improving

Attribution Model	How It Works	Use-cases
		conversion. For instance, you can use last-touch attribution in cases where customer journeys are short, when the customer's path to conversion is straightforward and quick, or when you need to get a clear understanding of the touchpoint responsible for the final conversion.
Last-touch non-direct	The last non-direct touchpoint is assigned 100% of the attribution credit. A non-direct touchpoint refers to customer interactions that occur outside of direct company communication channels and can influence customer decisions and brand perceptions (like word of mouth or online reviews)	This model helps understand the role of nurturing touchpoints. In customer journeys, there are often touchpoints that play a crucial role in guiding leads towards conversion. This model helps us identify and acknowledge their contribution to the conversion.
Linear	All touchpoints are evenly assigned attribution credit.	By assigning equal credit to all touchpoints, you can identify the strengths and weaknesses

Attribution Model	How It Works	Use-cases
		of each channel and make data-driven decisions on budget allocation and campaign optimization.
U-shaped	All touchpoints are assigned attribution credits– but higher credits are assigned specifically to the first and last touchpoints	The U-shaped attribution model considers the impact of branding and remarketing touchpoints throughout the customer journey. It recognizes the role of initial brand awareness and subsequent remarketing efforts in driving conversions. With this model, one can assess the effectiveness of your branding and remarketing strategies in nurturing leads and increasing conversion rates.
W-shaped	Like the U-shaped attribution model, the first and the last touchpoints are also given importance in the W-shaped attribution model. However, if	It helps identify touchpoints that contribute to initial awareness, consideration, and final conversion. This attribution model is beneficial for analyzing

Attribution Model	How It Works	Use-cases
	you generate quality leads in the middle of the sales funnel, then that touchpoint is also considered influential And is, therefore, given equal importance as that of the first and last touchpoint.	the effectiveness of campaigns across various channels, evaluating mid-funnel touchpoints, and optimizing lead nurturing efforts. It helps you identify touchpoints that contribute to building trust, addressing customer concerns, and influencing the decision-making process.
Time-decay	Each touchpoint is given progressively higher credit, with the first touchpoint having the least credit and the last touchpoint having the most.	Time decay attribution considers the cumulative effect of touchpoints over time. It recognizes the value of consistent and continuous engagement with customers throughout their journey. This attribution model can be valuable for assessing the impact of ongoing nurturing activities, such as email marketing campaigns or drip campaigns, in driving

Attribution Model	How It Works	Use-cases
		conversions and maintaining customer engagement.

In the end, a lot of the use cases for these types of attribution models are subjective. The decision to opt for a specific model can be based on several reasons spanning from the nature of your product to the extent of your brand equity. It may also vary based on the specific kind of insight you want to achieve.

More often than not, you will find yourself using more than just one model with several stipulations and custom values for each variant. Fortunately, the progressive ingenuity of AI and constant innovations around attribution modeling will render your experience less of a trial by fire and more of an intuitive, insightful practice.

Leveraging the right marketing analytics platform will be the first step in deciding the attribution model required for your company/business. As we said, it's best to rely on more than one model to improve your desired results. And for that, you will need an expert team, like Factors, that understands your requirements and guides you in leveraging the right techniques.

With Factors.ai, you can easily track the effectiveness of your campaigns and content, identify which channels are driving the most conversions, and optimize your marketing efforts for maximum results. The tool also offers a user-friendly interface and customizable dashboards, making it easy for you to access and interpret your data.

UNIT 5

Customer Churn Modelling

Churn (aka customer attrition) is a scourge on subscription businesses. When your revenue is based on recurring monthly or annual contracts, every customer who leaves puts a dent in your cash flow. High retention rates are vital for your survival.

So what if we told you there was a way to predict, at least to some degree, how and when your customers will cancel?

That's exactly what a churn model can do.

Building a predictive churn model helps you make proactive changes to your retention efforts that drive down churn rates. Understanding how churn impacts your current revenue goals and making predictions about how to manage those issues in the future also helps you stem the flow of churned customers. If you don't take action against your churn now, any company growth you experience simply won't be sustainable.

In the following article, we'll talk through everything that goes into building your own churn prediction model and how to go about starting the process.

What is a churn model?

A churn model is a mathematical representation of how churn impacts your business. Churn calculations are built on existing data – the number of customers who left your service during a given time period. A predictive churn model extrapolates on this data to show future potential churn rates. This helps you predict your revenue and avoid risks like overspending.

Types of churn: voluntary vs. involuntary

Churn traditionally falls into two buckets: voluntary or involuntary. Analyzing how these types of churn impact your business provides a comprehensive picture of how and why customers cancel their accounts. With that knowledge, it's easy to formulate a plan for proactively addressing the issue.

Voluntary churn

Voluntary churn is characterized by customers actively choosing to cancel their service. This can happen for any number of reasons, including:

- Switching to a competitor
- Closing down a business venture
- Negative customer experiences

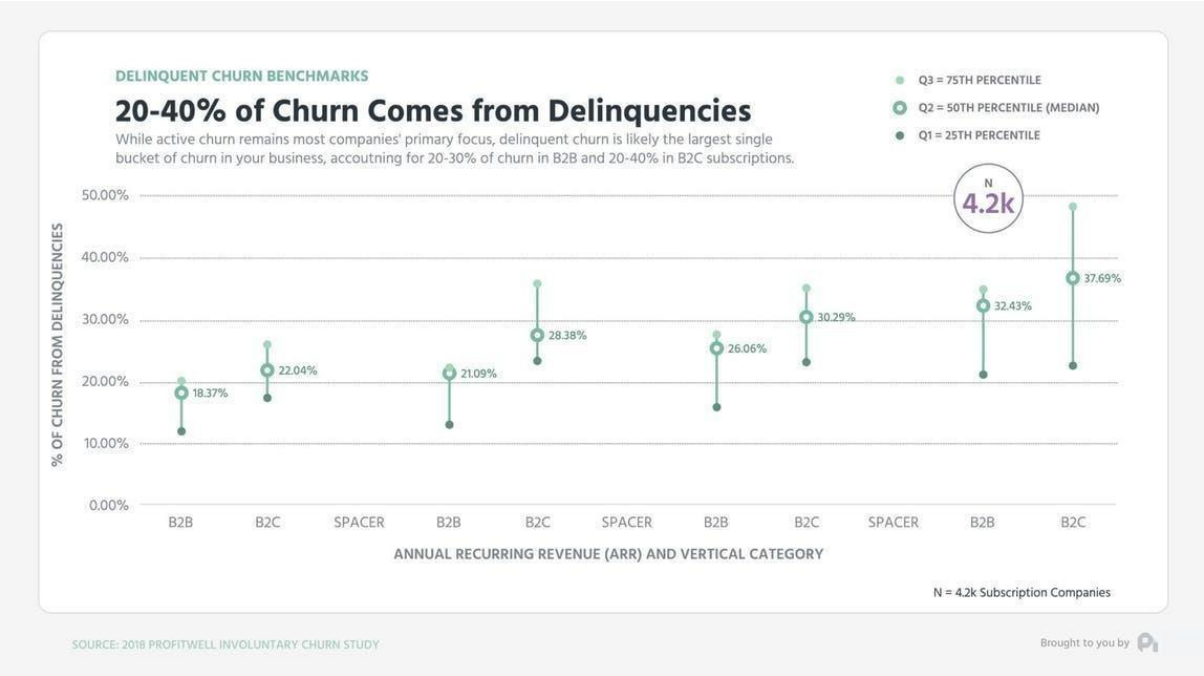
This type of churn is preventable. A customer who's thinking about leaving you for a competitor might be convinced to stay if they're reminded of the value your service provides. However, it is tricky to dissuade customers from voluntary churn since the reasons are often complicated to solve.

Involuntary churn

Involuntary churn occurs when a customer's account is canceled when they didn't intend it to be. Some examples of involuntary churn include:

- Expired credit cards
- Reaching the limit of available funds
- Failed mechanical payment processing
- Fraud protection on recurring payments

Research by ProftiWell (now part of Paddle) found that 20-40% of total churn comes from involuntary churn.



Involuntary churn is more easily prevented than voluntary churn because it is almost always the fault of a mechanical failure in your internal processes. In most instances, a proper system of dunning emails and notifications can take care of delinquencies at the company level.

Understanding how voluntary and involuntary churn differ helps you segment your customer base and create a more comprehensive model of churn. Use this data to analyze the impact that each type of churn has on your overall churn rate.

The data you need to build a churn model

Building a predictive churn model for your business starts with categorizing everything you know about your customers. Most businesses are already tracking this data, so you just have to know how to use it. Every customer data point you have helps build a more targeted churn model.

Customer information

The first step is building comprehensive customer profiles. At their core, these profiles should include the customer's name and address but can be expanded to include job title, employment status, team size, and much more.

With this data, you can easily spot patterns in churned customers related to their demographics and segment them into cohorts for more granular analysis. Different customer types will churn in significantly different ways.

With Retain, businesses can track customer cohorts over time:

Purchase Information

Expand on your customer profiles by including information about their purchase and billing history. Knowing when a customer signed up, when they canceled your service, their payment history, and overall lifetime value (LTV) helps you build a clear picture of how billing processes impact your churn.

As a SaaS company, it's important to include a customer's chosen pricing tier in your churn data as well. This information helps you see how your pricing decisions affect the way customers churn from your service.

Interaction Information

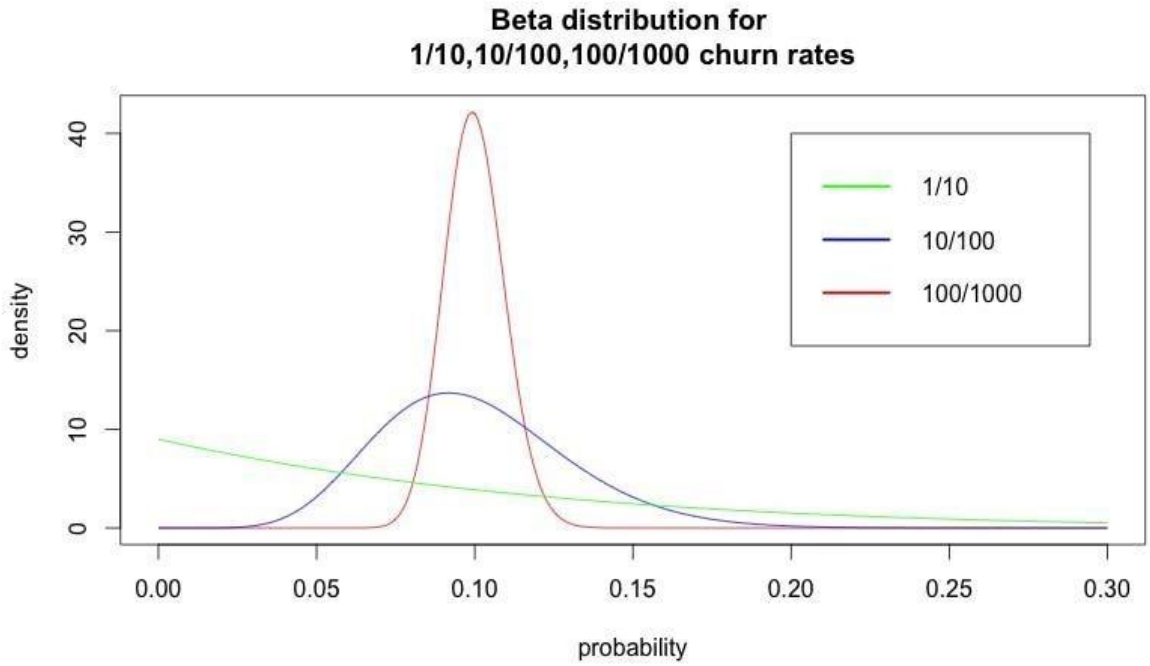
One of the biggest contributors to voluntary churn is the customer experience. Make sure you're tracking every interaction a customer has with your team as well as your product. Including this information in your customer profiles helps you see the impact of your product and the customer experience on churn rates.

Tracking past interactions can also be valuable for surfacing points along the customer journey where churn is more likely to occur.

Customer profiles are the basis for more in-depth churn analysis. With this data, you can start looking for patterns in how and why different types of customers leave your service.

How to build a churn model manually

With your customer profiles created and analyzed, it's time to talk through how to create an actual churn model. As this is a mathematical process, you'll need a strong understanding of statistical concepts and data science to move forward. You'll also need a significant amount of data to perform these calculations.



Example beta distribution of churn via Neil Patel.

If you only rely on a small sample size, like the green line in the above graph, you'll never be able to create an accurate picture of your overall churn. The more data you have at your disposal, the more specific your churn model will be.

Let's get started.

1. Gather and review your data

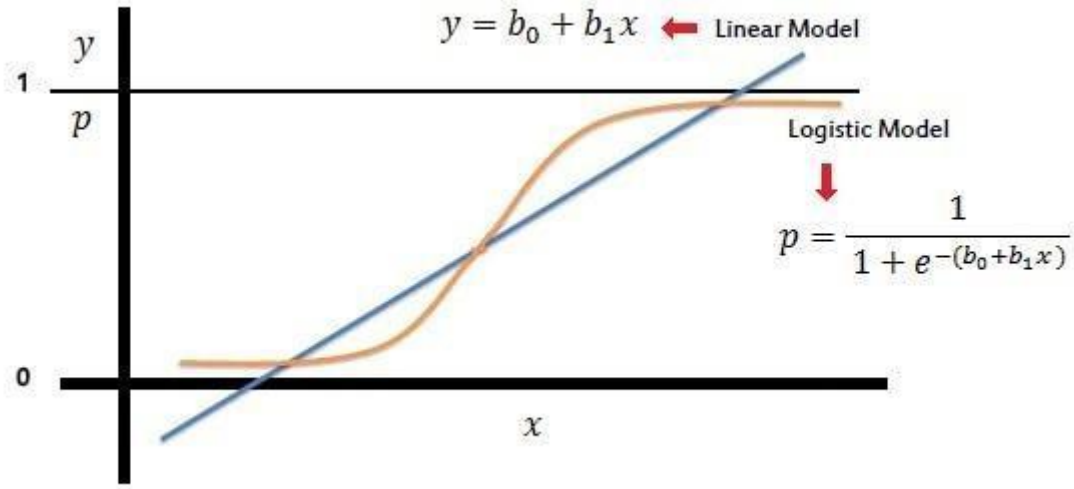
You've spent all this time building up a data set—every bit of customer information you have is a valuable data point in the upcoming churn calculations. Make sure you review all of your data for accuracy and validity before moving on to the math.

2. Set up a regression formula

Mathematical modeling for churn is built on a statistical process called logistic regression. This process determines the relationships between points in your data

set based on a formula and limits the outcome to between 0 and 1. You'll take all the customer information, purchase history, SaaS metrics, and prior churn data and turn it into a statistical prediction of when certain types of customers might churn in the future.

Your formula and potential outcome will look like this:



Logistic model vs. Linear model via Saed Sayad.

You'll likely need to employ a data scientist to determine the correct variables and constants required to build a formula. Learn more about this process in The Complete Guide to Statistics for the SaaS Executive.

3. Come up with a retention plan

Once you've modeled churn through the logistic regression formula, you'll be able to more clearly analyze retention and see the probability of certain customer segments churning.

To help maximize retention, use this information to formulate a plan, based on these findings, that targets each of your cohorts directly. The probability of certain customers churning your service earlier than others will make it easy to prioritize your actions.

4. Implement and track your results

With a plan in place, it's time to implement your retention strategy. As you do so, keep track of how it impacts your churn rate over the next few months. Gather enough data to see the real impact of your efforts before making additional changes to your plan. This data might look something like this:

	August	September	October
Existing Customers	10,000	9,598	13,993
Existing Churn	-500	-480	-700
New Customers	100	5,000	5,000
New Churn	-2	-125	-125
Total Customers	9,598	13,993	18,168
Adjusted Churn Rate	5.12%	5.13%	5.13%
Quarterly Churn Rate	4.57%	4.57%	4.57%

Example churn data from The Best SaaS Churn Formula.

5. Test retention strategies

Your churn model will provide probabilities for a number of different cohorts in your customer base. Make sure you're always testing out new strategies and recording the impact on these customer segments. Each subsequent test can help you create a better model for the future.

There's a lot more that goes into creating a mathematical model of your churn than you think. Analyzing this information takes time and resources from across your team.

Easy and accurate churn models with ProfitWell Retain

Creating a predictive churn model for your business is a lot of work and requires considerable expertise and mathematical knowledge. Fortunately, there is an easier way to build a churn model—ProfitWell Retain.

1. Get ProfitWell Retain for your business

ProfitWell Retain, a Paddle product, has one of the largest databases of subscription data on the planet. Using this information, we've built a tool to help model churn for you. Our retention experts have years of experience working with hundreds of SaaS companies to help you make sense of your data and make smart decisions for the future of your business.

2. Minimize churn and keep more customers

ProfitWell Retain is easy to set up, secure, and offers industry-leading retention rates for customers. Using our tool, we'll demystify the modeling and retention process and give you back time that you can spend on keeping customers happy and growing your business.

Take the headache out of growing your software business

We handle your payments, tax, subscription management and more, so you can focus on growing your software and subscription business.

Modeling churn helps you understand it

A predictive churn model is one of the best tools you have for deciding where to focus your retention efforts. It helps you weed out both types of churn and focus

on where your team can make the most impact. That focus lets you spend your time looking at new ways to keep more customers and grow your company. The less customers you lose to churn, the more revenue you'll be able to capture from them.

Purchase Behaviour Prediction Modelling

When it comes to human behavior and understanding the key psychographic that forms decisions, putting a pin on it isn't as simple as it may seem. There are several factors that influence people's outlook, approach, and purchasing patterns; all of which marketers find hard to decipher.

In this blog, we'll tell you how to unlock its potential for better marketing.

Reward desired behavior with a loyalty program

Reward customers with loyalty points every time they do what you want. Positive reinforcement can do magic!

START NOW, IT'S FREE

What is Customer Behavior?

Consumer behavior means understanding the process and means of how audiences make end decisions. It includes focusing on their actions, engagement patterns, as well as usage or disposal of specific products or services – and the mental, emotional and behavioral responses that trigger those actions.

Why do marketers need to study it? Simply because the influence of psychology is a key factor in forming preferences, which eventually leads to purchases. The ability to connect with your audience right from the beginning of their journey, to instill trust and convey why and how you reign supreme over your competitors can win you life-long customers, a.k.a. recurring business opportunities.

Every business has an ICP and buyer personas they want to target. Consumer behavior marketing provides the fundamentals for understanding who you're targeting and what potentially would interest them. Consumer behavior research provides the underlying element that drives quality strategies and ensures business results.

Think about Apple. Behind its winning marketing strategies and an ever-increasing number of customers who swear by its products is its ability to change with the shifts in consumer behavior. With iTunes capturing first-party data, or the company leveraging reference groups, or hooking people on product experiences, Apple invests greatly in understanding its customers.

As per their customers' behavior, and preferences, Apple implements changes in its physical and digital elements to revamp its product offerings which in turn leads to customer delight.

Its monochromatic logo, feature updates, and brand messaging all are examples of how they use key-value propositions such as ease of use, sleek design, and exclusivity, to wow its customers.

After understanding how important is behavior analysis, the next obvious question would be: How to make your customers say yes to your products or services?

There are two categories of influences on consumer behavior:

External Influences

Demographics: an individual's interests can be greatly influenced by demographics (age, gender, culture, etc.).

Social factors: surrounding people, such as family, friends education, social media, and purchasing power, all influence consumers' behavior.

Internal Influences

Psychographics: Existing preferences, attitudes and beliefs can influence a customer's association with your brand.

How Does Customer Behavior Prediction Work?

Customer Behavior prediction is all about unveiling your customer motivations. It helps marketers with advanced segmentation and targeting based on behavior.

Gathering information about how your customers engage with your brand can give you insights into what's working and what's not; helping you create campaigns that are contextually relevant and revenue-generating.

Customer behavior prediction comes in various ways. It can be through collecting information through primary or secondary research such as analyzing online actions, feedback analysis, focus groups, conversational marketing, and more.

Data analytics has enabled eCommerce shops to dive deeper into consumer behavior by analyzing actions such as purchase history, cart abandonment, sessions, search histories, or social media profiles.

For marketers to receive these bits of information in real-time, there are data analytics platforms that help reveal predictive insights, but most struggle with the quantity of data and effective reporting to strategically act on them.

How to forecast behavior through modeling

Predictive analytics is based on behavior modelling, which uses statistical significance to evaluate customers' historical data and retrieve possible future actions. Modelling helps create a mathematical construct to highlight common behaviors among certain groups or segments of customers, which further hints at how similarly they'll behave to different external stimuli.

Customer behavior models are created using aggregated customer data, and they can help answer many questions based on which brands can pivot their marketing. This helps bring out the best marketing outcomes for each group of customers, and also helps understand how and what customers base their decisions on (e.g., purchase, bounce, upsell, churn).

What are the Customer Behaviors You Should Analyze?

When it comes to online shopping, the Covid-19 surge, virtual convenience, as well as global access, have increased customers' proclivity to buy online. In a recent survey, more than 65% of the customers vouched for online shopping than going into a store or event, etc. As it is established that ecommerce is positioned to grow, brands need to know how to capture customers' attention and drive sales.

For this, data gathering and analysis have been fueling more and more marketing efforts. From social listening to feedback calls, AI-driven message boards and more – companies are constantly on the lookout for cues and changes in behavior. If interpreted well, they can bring out key insights to scale campaigns and create content that "clicks" with your audience. Some behaviors that ecommerce brands are increasingly keeping tabs on are:

1. **Nature of the purchase:** When you analyze customer behavior, knowing if the nature of the purchase – i.e. if it is habitual (buying often), impulse purchase, variety-driven, promotional, etc. can give you insights into purchasing patterns and how your customers would respond to different marketing campaigns.
2. **Average spending:** Knowing about the spending habits (high-spend, low-spend, etc) of your audience can help you understand their purchasing power and push relevant services/products to them.
3. **Demographics:** This is the most pressed upon factor. e Demographics include gender, age, income, location, and others.
4. **Churn rate** – Also known as customer attrition, it's when a customer discontinues purchasing from a brand. It's one of the easier metrics to calculate, predicting the probability that someone cancels or fails to renew.
5. **Frequency and Recency:** Learn more about how frequently customers engage with your brand and what is the average time interval between their purchases. This can help you target customers with one-time purchases or those with repeated purchases.
6. **Preferred channel of engagement:** Understanding what sources/ mediums/ channels your customers refer to for consuming information is essential to target them with the right message at the right time.

Its Relation to Machine Learning

The biggest marketing challenge is mastering the art of influence. And it becomes even more challenging when we know no two customers are alike. To track each customer's intrinsic motivations, their actions, and preferences, aren't humanly possible. Here is where humans leverage machine intelligence.

Human beings, although rational beings, have decision-making abilities that can be erratic, and not always defined by logic. Therefore, to the best influence, it is

to observe and learn from it. Here is where machine learning comes into play. It can help you detect patterns not visible to the naked eye and estimate how existing and new customers are likely to engage.

Machine learning helps not just accumulate huge amounts of information about your customers but also establishes connections, builds segments and joins the dot between historical data and defining future customer behavior— i.e. what are their buying preferences, channel preferences, satisfaction levels and more.

However, predicting customer behavior through machine learning is only as good as the data available. Therefore, marketers need to make sure to constantly filter out archaic data and refer to specific data points that can derive meaningful conclusions.

How to Predict Customer Behavior in 4 Steps

1. Identifying prospects

Marketing today has been able to create one-on-one interactions with customers, but the catch is that these interactions are made precise and relevant by categorizing customers into certain segments.

Most companies start by building profiles of customers and clustering them with other customers that share the same traits. Data points such as demographics, geographic, product channels, and previous purchases can be used to segment customers.

But to create an even more advanced segmentation, you can leverage tools like Verfacto that capture technical attributes such as AOV, RFM (Recency, Frequency, Monetary Value, etc.), attention rate, sessions, returners, etc. for precise targeting.

2. Feature extraction to create profiles

Ask and evaluate critical questions like: who are your best customers, and why are they your customers? What do they buy, and why do they choose you over others?

Based on that, you can identify the different engagement patterns of your customers and understand how to market them.

3. Create models to analyze customer behavior

Once you create your customer segments, machine learning can help predict their future actions, such as high/ low purchase intent, churn, preferred products, etc.

You can assign a certain bandwidth to each of these metrics (between 1-5 or 1-10). The idea is similar to lead scoring. This can help you better assess where each customer lies in terms of being a qualified/ unqualified lead.

Certain ecommerce analytics tools including Verfacto allow you to build clear identifications through data visualizations that help you compare and contrast each metric against the other – e.g. customer recency, frequency and monetary value of purchase.

4. Customizing communication to create a better pipeline.

By creating real-time customer profiles, you can clearly identify what communication you would want to trigger to a specific segment. Teams can utilize this information to create campaigns for their high-intent/ medium or low-intent prospects.

For example, the leads that are on the top of the funnel, content that is less sales-oriented and more informative, and value-driven (blogs, videos, infographics) can help you build the trust among visitors that ultimately leads to purchases.

While middle or bottom-funnel leads, who are aware of their problem and are looking for solutions, can be benefited from brand-driven content such as case studies, testimonials, live events, etc., to put more weight on your end when comparing other competitive offers.

Personalization in marketing and sales is a proven approach that creates a better pipeline, and opportunities, and helps improve the stage-to-stage conversion rate.

The 5 Best Tools to Predict it

Each customer behavior analytics tool offers different insights into your users' preferences and actions. While some tools can help you see the page-on-page engagement on your website or how are they moving forward through the funnel; others can get granular and capture different customer attributes and also yield out how they are interacting with specific elements.

If you wonder which tool would be the best fit for your eCommerce store, here are some of our top picks:

1. HubSpot

HubSpot is a CRM (Customer Relationship Manager) that offers brands to personalize their marketing outreach through behavioral targeting. It tracks behaviors based on customer personas and engagement activities such as web visits, email interactions, form submissions, etc.

You can leverage the tool to identify high-intent behaviors and develop journeys that bring your customers to a purchasing point. HubSpot is being currently used by more than 113,000 companies around the world.

2. Optimizely

Optimizely helps ecommerce brands move beyond just tracking data regarding churn, and conversions, to offering experience analytics that helps understand and optimize the entire customer journey. Often instating themselves as a digital experience platform (DXP), Optimizely helps brands create strategies that conform to customer behavior, preferences, and history.

With Optimizely, you can collect data at different touchpoints, monitor behavior across your channels and use the information to make intelligent decisions for each digital interaction.

3. Google Analytics

Being long in the game, Google Analytics has managed to stay relevant even with new-age tools emerging. Along with tracking standard metrics such as sessions, clicks, and conversions, the tool allows ecommerce brands to capture even more detailed interactions using event tracking.

For ecommerce business, in particular, Google Analytics offer "Enhanced Ecommerce" tracking to understand user behavior across their online experiences. With the tool being more or less sufficient, it is typically used by brands in tandem with other analytics tools for conducting behavior prediction.

Google Analytics is user-friendly and is the top choice for brands to analyze their website performance, study channel engagements, and create goals. Along with that, the reporting options are quick and comprehensive.

4. Verfacto

Verfacto is a new-age ecommerce analytics tool that answers and solves challenges in today's ecommerce marketing. This ranges from helping brands capture customer behavioral attributes to hyper-targeting, building and enriching customer experiences to driving campaign performances through contextual communication.

The tool helps convert raw and aggregate data into actionable insights, all in real-time, so that eCommerce stores can pivot by the minute. Apart from behavioral analytics, advanced segmentation and customer profiling, you can use the tool to create precise reports that inform about essential metrics.

5. Woopra

Woopra is an end-to-end customer journey and product analytics tool. It offers real-time insights into every action that your customers are taking including checkouts, web visits, open rates, etc. all in one platform. You can put the user

behavior data to use by optimizing your marketing campaigns, identifying gaps in existing processes, and improving the customer experience on your site.

It also offers customer journey reports and visualizes the browsing patterns so that marketers can study them more critically. You can also retrieve cohort reports to analyze and compare growth trends over time.

How To Forecast and Influence Customer Behavior In Real-Time with Verfacto

Understanding your customers' motivations is key to serving them better, which ultimately brings tangible value to your business. With Verfacto, analyzing customer behavior begins with realizing answers to the questions such as:

- What have been the customers' previous proclivities, and actions? (historical data)
- What are the factors influencing a customer to make purchase decisions?
- Why does a customer prefer one product/service over the other?
- How to segment them based on their interests, behavioral attributes or stage in the journey?

Through these questions, brands can perform customer behavior predictions that help better their marketing efforts. With Verfacto's Real-time customer Profiler, businesses can clearly segment audiences based on **behavior in the current session, customer historical data and Behavioral historical data.**

These attributes, which are further classified into certain events, can offer insights into buying intents, help optimize funnels and more. Brands can leverage the insights to further **communicate their relevancy to customers, making them their preferred choice for online purchasing.**

Real-time customer profiler is a giant leap forward from manual to analytics-fueled marketing. Removing the time and effort constraints of mapping multiple

customer events to produce actionable information, this feature can just be activated with an on-site tracking script. Learn more here.

Conclusion

In the ecommerce landscape, the customer is the king. So, there's a constant battle between brands to not just catch the eye, but retain themselves in customers' minds too. And the only way to do it is to work in their best interest- understanding the "why" of every purchase and bringing the "when" and from "who" in your favor.

Thanks to behavioral analytics tools, marketers can now track, record, and influence customer behaviors, all to improve user experiences. But how successful they get depends on the data they capture, as it should be contextual and structured.

Furthermore, choosing a tool custom-made for ecommerce business would be a smart choice. Tools such as Verfacto offer real-time insights that can help you gain an edge when it comes to optimizing strategies for ever-changing brand-customer dynamics!

Social Media Listening and Sentiment Analysis

How sentiment analysis and social listening can improve customer experience

Sentiment analysis may be popular with social media, but it probably has far more uses outside of social media. Sentiment analysis and social listening are invaluable tools that enable organisations to gather valuable insights into customer experiences. It involves the automated extraction of emotions, feedback and attitudes expressed in textual data. Machine learning (ML) techniques such as natural language processing (NLP) and associated calculations help analyse if the data collected is positive, negative or neutral.

When NLP is incorporated into chatbots and voicebots, both tools become almost human-like in their conversational and language skills, and can adjust tones during conversations. By leveraging these techniques and technologies, businesses can proactively address customer needs, resolve issues, and ultimately deliver an exceptional customer experience.

Points to ponder

Before embarking on this journey, organisations must be aware of certain facts:

- What is the purpose of the exercise? Is it to understand the competition in the industry or to identify the most discussed features of a product or service?
- What are the metrics that would answer the questions asked? While mentions in online posts can give a clear idea about the volumes of conversation about the product or service, the sentiments marked such as likes, comments, favourites - indicate whether the opinions will harm or help the brand discussed, and the engagement with the brand gives an idea about what features of the brand appeal to the customer.
- Analysis of all the information collected is the most challenging step. There are tools available or the organisation could have an internal team of scientists and data analysts to collect and analyse the data. While the scientists could predict patterns by using past patterns, data analysts could gather meaningful insights.

Key benefits of sentiment analysis and social listening

Sentiment analysis, or opinion mining, and social listening enables organisational departments to attach measurable metrics to pieces of data and use them in daily functions. Data-driven decisions and targeted actions can be taken to boost customer satisfaction. Benefits of sentiment analysis include:

- **Real-time feedback and action**: Organisations can monitor customer sentiment in real-time and whenever possible, take prompt action to address any negative issues and experiences. Customer queries and complaints can be managed immediately too. A real-time alert system that gets triggered when a particular subject is mentioned or an influencer is active can help getting updates quickly. Early detection of such posts can help organisations manage the situations at the right time.
- **Identify trends**: By considering customer sentiments at various touchpoints, organisations can easily identify areas for improvement and also recurring trends. Strategies to address them can be developed proactively. Interest about the product or service can be gauged and brand owners get a better idea about how to connect with the customer.
- **Personalise customer experiences**: By understanding customer sentiment, organisations can customise their offerings and communication to meet individual needs, resulting in a more personalised and engaging customer experience.
- **Analyse the competition**: Organisations can gain a competitive edge by monitoring competitor activities and analysing customer sentiments. Gaps

can be identified and strategies can be developed to meet these requirements.

- **Identify individual or group influencers**: Social listening enables organisations to identify individuals or groups that influence customer action. Influencers can be engaged in conversation and persuaded to be brand advocates who would then drive positive sentiments.

Listening posts

There are many points or sources from which data for sentiment analysis can be procured. Some of them are:

Discussion points: Sentiment analysis sources are not limited to social media posts only. Products and services could be discussed in chatbot communication, phone logs, review websites, articles, emails, internal communication, support tickets and more. This is where artificial intelligence (AI) steps in. It can analyse text and human language from all these sources and provide real-time insights. Based on these data-driven insights, organisations can address specific situations.

Requests: Organisations are frequently overwhelmed by requests from customers on a wide variety of topics. It is with the help of NLP and natural language understanding (NLU) that customer emotions and tones are deciphered from both the written and spoken requests. The urgency of each request can be understood and sorted better by the AI techniques. Responders can then prioritise the jobs and also ensure that each request is diverted to the appropriate channel or department.

Project discussions: Sentiment analysis can be utilised within the organisation too. Project managers can get insights into employee sentiments regarding a project - how the project is perceived, how each team member feels about his/her role in it, and how effective the team communication is. Any negative sentiments,

if valid, can be addressed and nipped in the bud. The manager can steer the project better after being updated with all the information.

By keeping tabs on employment websites, internal messaging platforms and email communication, organisations can easily gather insights into employee sentiments and take necessary steps to reduce turnover.

Sentiment analysis is a way of life

When it comes to experience and customer service, consumer expectations are undoubtedly high. Automated sentiment analysis provides actionable insights to organisations enabling them to better understand customer pain points as well as what makes customers happy and satisfied. Metrics provided by sentiment analysis can help organisations make sound data-driven decisions into many unexpected areas of customer experience.

If the data sourcing techniques and the analysing algorithms are accurate and sound, most organisations can gain significantly from sentiment analysis in terms of customer support, employee retention, marketing efforts, product development and more.

** For organizations on the digital transformation journey, agility is key in responding to a rapidly changing technology and business landscape. Now more than ever, it is crucial to deliver and exceed on organizational expectations with a robust digital mindset backed by innovation. Enabling businesses to sense, learn, respond, and evolve like a living organism, will be imperative for business excellence going forward. A comprehensive, yet modular suite of services is doing exactly that. Equipping organizations with intuitive decision-making automatically at scale, actionable insights based on real-time solutions, anytime/anywhere experience, and in-depth data visibility across functions leading to hyper-productivity, Live Enterprise is building connected organizations that are innovating collaboratively for the future.*

Market Basket Analysis

The retail sector is especially benefiting from machine learning. It aids the retail industry in every way, from identifying customers to forecasting sales performance. One such prominent retail use of machine learning is market basket analysis (MBA). Knowing which goods customers frequently buy together enables merchants to organize their stores and websites consistently. It is mostly accomplished by looking at their prior purchase behavior. Businesses use it as a cross-sell tool for their itheon their web platform. But it's not just employed in the retail industry—false credit card transactions and insurance claims also use it.

Become The Highest-Paid Business Analysis Expert

What Is Market Basket Analysis?

Retailers utilize market basket analysis, a data mining approach, to boost sales by better understanding client buying habits. Identifying product groups and items that are most likely to be bought together, includes evaluating big data sets, such as purchase history.

Purpose of Market Basket Analysis

Finding items that buyers desire to buy is the major goal of market basket analysis. Market basket analysis may help sales and marketing teams develop more effective product placement, pricing, cross-sell, and up-sell tactics.

Types Of Market Basket Analysis

● Predictive Market Basket Analysis

This kind employs supervised learning methods like regression and classification. In essence, it seeks to imitate the market to examine what factors influence events. In essence, it determines cross-selling by taking into account things bought in a particular order.

● Differential Market Basket Analysis

For competition analysis, this kind of analysis is useful. To identify intriguing patterns in consumer behavior, it compares purchase histories across brands, periods, seasons, days of the week, etc.

Become an AI-powered Business Analyst

Algorithms Associated With Market Basket Analysis

The market study definition is based on Association Mining rules, as was already explained. Association mining is a technique used by the AIS, SETM, and Apriori algorithms. The Apriori Algorithm is the MBA algorithm that is used the most frequently.

How Does Market Basket Analysis Work?

The IF, THEN construct is used in association rule mining to replicate market basket analysis. When a customer buys bread, he is likely to also buy butter. Examples of association rules include the following: "Bread" -> "Butter"

Learn the following definitions to better understand market basket analysis:

● Antecedent

The entities or "itemsets" produced from the data are called antecedents. To put it another way, it's the IF element on the left. In the situation before, bread serves as the antecedent.

● Consequent

The term "consequent" refers to an item or group of items that are encountered along with the antecedent. The THEN part of the sentence is displayed on the right-hand side. The result in the aforementioned case is butter.

Metrics For Market Basket Analysis In Data Mining

You can put a lot of interesting controls on your association rules. These consist of

- Support
- Confidence
- Lift

Consider the following scenario: A well-known e-commerce site handled 4000 transactions. They are trying to determine how many transactions, how much lift, trust, and support there is for the two things, a phone, and a phone cover, out of 5000. The phone has 500 transactions, the phone case has 800 transactions, and the two together have 1000 transactions.

Benefits Of Market Basket Analysis

- Gaining market share: Once a business reaches its peak growth, finding new ways to do so might be difficult. Market basket analysis may be used to integrate gentrification and demographic data to locate the sites of new businesses or geo-targeted marketing.
- Campaigns and promotions: MBA is used to identify the goods that work well together as well as the products that serve as the cornerstones of their product range.
- Behavior analysis: A fundamental tenet of marketing is comprehending consumer behavior patterns. MBA may be used for anything, including UI/UX and basic catalog designs.
- Optimization of in-store activities: MBA is useful in deciding what goes on the shelves as well as at the back of the shop. Because geographic patterns are a major factor in determining the strength or popularity of particular products, MBA is increasingly used to manage inventory for each store or warehouse.

Become The Highest-Paid Business Analysis Expert

Examples Of Market Basket Analysis

Retail

The most well-known case study using market basket analysis is probably Amazon.com. As soon as you visit Amazon to look at a product, the product description will suggest "Items purchased together frequently." It is the clearest and most straightforward example of Market Basket Analysis cross-selling tactics. Along with e-commerce methods, consumer in-store retailers also greatly benefit from BA. For grocery stores, visual merchandising and shelf optimization is crucial. For instance, shower gel is almost usually kept close to one another at the grocery store.

IBFS

Examining credit or debit card history is a highly advantageous MBA opportunity for IBFS companies. For instance, Citibank frequently sends sales representatives to large malls to tempt potential customers with enticing on-the-go discounts. Additionally, they collaborate with services like Swiggy and Zomato to provide customers with a selection of offers that they may use their credit cards to redeem.

Telecom

Due to the intense competition in the telecom sector, businesses are paying close attention to the advantages that customers frequently utilize. For instance, telecom has started to combine TV and Internet bundles with other affordable internet platforms to reduce migration.

Choose the Right Program

Unlock the power of data with Simplilearn's Business Analytics courses. Gain the skills to analyze, interpret, and make informed business decisions. Conclusion

Market basket analysis may be used by more and more businesses to get relevant information about associations and unspoken linkages. A predictive form of market basket analysis is gaining traction across various industries in an effort to pinpoint sequential purchases as industry leaders continue to investigate the technique's use.

RFM Analysis

The “RFM” in RFM analysis stands for recency, frequency and monetary value. RFM analysis is a way to use data based on existing customer behavior to predict how a new customer is likely to act in the future. An RFM model is built using three key factors:

1. how recently a customer has transacted with a brand
2. how frequently they’ve engaged with a brand
3. how much money they’ve spent on a brand’s products and services

RFM analysis was born out of direct mail marketing, in particular a 1995 article by Tom Wansbeek and Jan Roelf Bult titled “Optimal Selection for Direct Mail,” which was published in the journal Marketing Science. Their work helped confirm the Pareto Principle — the idea widely held among marketers that 80% of sales come from 20% of a brand’s customers.

Benefits of RFM Analysis

RFM analysis enables marketers to increase revenue by targeting specific groups of existing customers (i.e., customer segmentation) with messages and offers that are more likely to be relevant based on data about a particular set of behaviors. This leads to increased response rates, customer retention, customer satisfaction, and customer lifetime value (CLTV).

Each of these RFM metrics has been shown to be effective in predicting future customer behavior and increasing revenue. Customers who have made a purchase in the recent past are more likely to do so in the near future. Those who

interact with your brand more frequently are more likely to do so again soon. And those who have spent the most are more likely to be big spenders going forward. RFM analysis enables you to target customers with messages that best match their relationship with your brand. For example, you are likely to have more success suggesting big-ticket items to customers who spend frequently and in large amounts. On the other hand, you are more likely to grow the customer value of your relationships with consumers who make purchases frequently, but only in small amounts, by rewarding them for their loyalty or offering referral promotions.

How Does RFM Analysis Work?

Market research has traditionally concentrated on demographic and psychographic data, which marketers use to conduct customer segmentation. Those data points are then used to predict customer behavior across much larger populations that share the same set of traits. However, these methods depend on data from a small sample of consumers.

With the advent of systems like customer data platforms (CDPs) that help gather, unify and synthesize customer behaviors, marketers have much more granular data about the habits of individual customers to inform segmentation. Rather than segmenting customers only using demographic and psychographic data, marketers can create segments based on the real-world behavior of individuals, including purchase history across any channel (online or offline), browsing history, prior campaign responses and more. Unsurprisingly, this type of segmentation is called behavioral segmentation.

And even a basic CRM system can perform rudimentary tracking of the three easily quantifiable characteristics that contribute to RFM analysis:

- **Recency value:** This refers to the amount of time since a customer's last interaction with a brand, which can include their last purchase, a visit to a

website, use of a mobile app, a “like” on social media and more. Recency is a key metric because customers who have interacted with your brand more recently are more likely to respond to new marketing efforts.

- **Frequency value:** This refers to the number of times a customer has made a purchase or otherwise interacted with your brand during a particular period of time. Frequency is a key metric because it shows how deeply a customer is engaged with your brand. Greater frequency indicates a higher degree of customer loyalty.
- **Monetary value:** This refers to the total amount a customer has spent purchasing products and services from your brand over a particular period of time. Monetary value is a key metric because the customers who have spent the most in the past are more likely to spend more in the future.

RFM Analysis for Customer Segmentation

Rather than analyzing the entire customer database, it’s better to segment customers by characteristics like age or geography and separate them into a customer group. By engaging them in a well-segmented marketing campaign, you are able to create a relevant, personalized offer for a high-value customer.

Computing RFM for real-world application typically requires special analytical expertise or advanced math skills. And, like any model, RFM models can vary in complexity from simple to sophisticated. RFM segmentation begins by ranking customers in each of the three categories: recency score, frequency score and monetary score. Typically, this is done on a scale of 1 to 10. A 10 indicates the top 10% in each category (i.e., the most recent to transact, the most frequent to transact and those who purchased the most), a 9 the next 10% and so forth. By using a RFM scoring system such as this you can construct an effective marketing strategy by creating customer RFM segments, including:

- **Your best customers:** These are the customers who earn top scores in every category. They're loyal, willing to spend generously and likely to make another purchase soon. Such customers are primed to respond well to loyalty programs. They're more likely to be interested in new products you launch. And because they're committed to your brand and its products, it probably makes less business sense to offer them discount pricing. Instead, increase CLTV by suggesting big-ticket items and recommending products based on past purchases.
- **Your big spenders:** This customer segment is based on only one of the three metrics: customers with top scores for monetary value. Typically, marketers target this segment with luxury offers, higher subscription tiers and value-add cross-/upsells that increase average order value. Again, it probably makes sense not to shrink margins by offering discounts.
- **Your loyal customers:** This is another customer segment that takes into consideration only one of the three metrics: customers with top scores for frequency. Despite making purchases often, they aren't necessarily your biggest spenders, so consider rewarding them with free shipping or similar offers. Advocacy programs and reviews can also be effective ways to engage these customers.
- **Your faithful customers:** Customers who score high for frequency but low in monetary value tend to respond best to product recommendations based on past purchases, as well as incentives tied to spending thresholds (e.g., a free gift for transactions above the brand's average order value).
- **Your at-risk customers:** Customers who have been in your top tier in the past (best, big spenders and/or loyal) but who now score low for recency and frequency present a special opportunity. Marketers should consider

targeting them with messages aimed at retention, such as discount pricing, exclusive offers and new product launches. With the help of your CDP, you can even create specific customer journeys aimed at re-engaging and retaining at-risk customers.

Steps of RFM Analysis

The steps below provide a high-level overview of how an RFM Analysis and segmentation is executed.

Build RFM Model

In order to build an RFM model, you need to assign a recency score, frequency score and monetary score to each unique customer. The raw data, which can be collected from a customer database from previous transactions, is then compiled in a spreadsheet or database.

Divide the Customer Segment

Next, divide the RFM database into tiered groups for each of the three values of the RFM score. Tier designation is based on the greatest to the least. For example, tier one for monetary value is assigned to the high spenders and tier five is assigned to the lowest spenders.

Select the Targeted Customer Group(s)

The third step involves the selection of the segmented customer group with high customer value. Organizing the RFM segment, you can begin to assign titles to segments of interest, such as your best customers, biggest spenders, faithful customers and at-risk customers.

Craft a Personalized Marketing Strategy

Finally, craft a unique marketing strategy designed for each RFM segment focused on their behavioral patterns. Utilizing the RFM Analysis, marketers are able to effectively communicate their messaging to customers in a way aligned with customer behavior.

Why RFM is Effective for Small and Medium-Sized Businesses

For startups and smaller retailers with limited marketing resources, RFM analysis can be a particularly effective tool because of its:

- **Simplicity:** RFM analysis does not, on its own, require complex tools or sophisticated analytical capabilities. The principles are easy to understand and the results are easy to interpret and act on.
- **Affordability:** In many cases, it's possible for marketing professionals without advanced statistical or analytical training to perform RFM customer segmentation with only a standard spreadsheet.
- **Effectiveness in direct marketing:** RFM analysis, which grew out of database marketing and direct mail marketing, has been shown to be effective with relatively inexpensive digital direct marketing strategies that smaller brands can afford, such as an email marketing campaign.

Scaling RFM to the Enterprise

That said, as a business scales, you will also need technology that scales with the complexity and volume of interactions across all your channels, regions and more. With advanced RFM, you can create more authentic experiences at scale, using a range of customer traits as inputs to your model and going beyond scores and segments to achieve one-to-one personalization.

The most advanced enterprise-class CDPs serve as an engine for creating these types of RFM-driven experiences. They empower business users to orchestrate campaigns and journeys quickly and seamlessly leverage the full breadth and depth of all your customer data across any and all channels.

The Limits of RFM Analysis: What to Avoid

While RFM segmentation is powerful, it does have limits. When performed manually, it's prone to human error. RFM analysis is also based on just a few

behavioral traits, lacking the power of the advanced predictive analytics now available.

Some businesses may use RFM analysis as an excuse to bombard high-ranking customers with messages and thus reduce response rates on campaigns that could otherwise be highly effective. On the other hand, it can cause marketers to neglect customers with low rankings even though many of them may be worth cultivating. For example, your RFM model may fail to account for the impact of past promotions or seasonality on RFM analysis. Likewise, a customer may have very little activity with your brand one month, yet be ready to engage in purchasing behavior the following month due to a birthday or anniversary.

How Relevant the RFM Model is Today

RFM analysis remains a perennial favorite of marketers. It's simple and intuitive, yet data-driven. It has the power to provide actionable insights down to the individual customer level — all without any input from data scientists or complex tools. That isn't to say you can't do sophisticated things with RFM analysis. For example, you can use RFM techniques to identify your best customers and turn them into a seed audience within an advertising platform that uses lookalike modeling to automatically identify prospects who share similar key traits.

Nevertheless, thanks to CDPs, marketers are now able to combine RFM data with other behavioral and demographic traits — everything from geolocation to recent products purchased — to create even more effective segmentation. Better yet, they can quickly and easily apply lookalike models and other sophisticated analytics to predict what messages are most likely to resonate and how and when those messages are most likely to prompt action.

With or without these more sophisticated approaches, marketers can use RFM analysis to:

- **Increase the effectiveness of email marketing campaigns:** Build an automated drip campaign with messages tailored to each segment.
- **Increase loyalty and user engagement:** Follow up with recent customers or new customers with timely promotions and educational content likely to increase their engagement with your brand.
- **Decrease churn:** Send personalized messages, offer repeat purchases at a discount or provide surveys that help you understand and address potential concerns.
- **Reduce marketing costs and increase ROI:** Reduce costs by focusing quickly and easily on smaller segments that are more likely to produce revenue and use insights from RFM analysis to optimize campaigns going forward.

Recommender Systems Development

Recommendation systems are built to predict what users might like, especially when there are lots of choices available. They can explicitly offer those recommendations to users (e.g., Amazon or Netflix, the classic examples), or they might work behind the scenes to choose which content to surface without giving the user a choice.

Either way, the "why" is clear: they're critical for certain types of businesses because they can expose a user to content they may not have otherwise found or keep a user engaged for longer than they otherwise would have been. While building a simple recommendation system can be quite straightforward, the real challenge is to actually build one that works and where the business sees real uplift and value from its output.

Recommendation systems can be built using a variety of techniques, from simple (e.g., based only on other rated items from the same user) to extremely complex. Complex recommendation systems leverage a variety of different data sources (one challenge is using unstructured data, especially images, as the input) and machine learning (including deep learning) techniques. Thus, they are well suited for the world of artificial intelligence and more specifically unsupervised learning; as users continue to consume content and provide more data, these systems can be built to provide better and better recommendations.

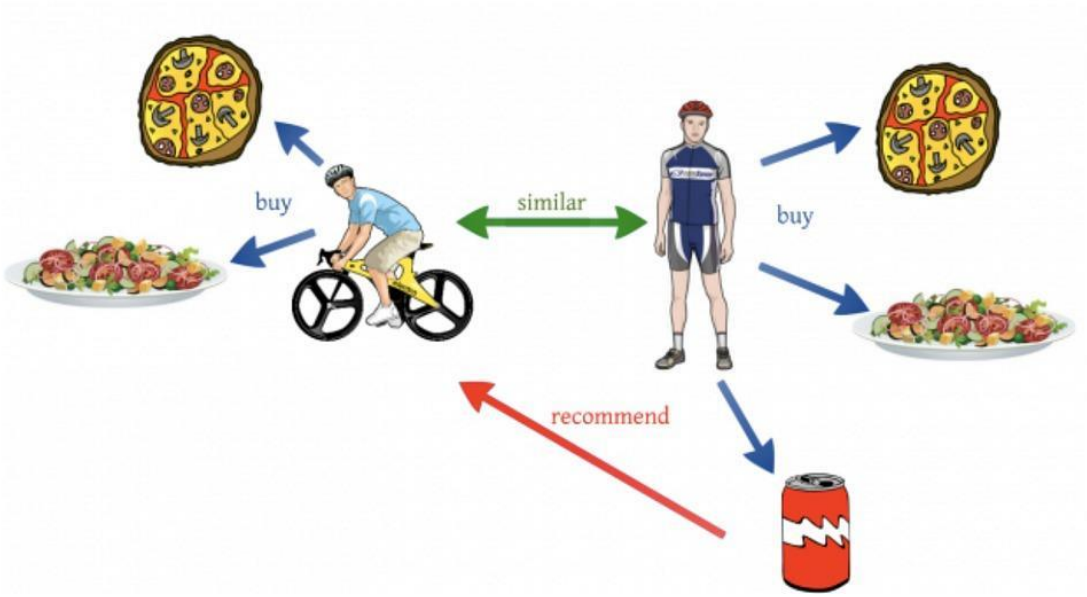
In this post and those to follow, I will be walking through the creation and training of recommendation systems, as I am currently working on this topic for Master Thesis. Part 1 provides a high-level overview of recommendation systems, how they are built, and how they can be used to improve businesses across industries.

The 2 Types of Recommendation System

There are two primary types of recommendation systems, each with different sub-types. Depending on goals, audience, the platform, and what you're recommending, these different approaches can be employed individually, though generally, the best results come from using them in combination:

1 — Collaborative Filtering

It primarily makes recommendations based on inputs or actions from other people (rather than only the user for whom a recommendation is being made).



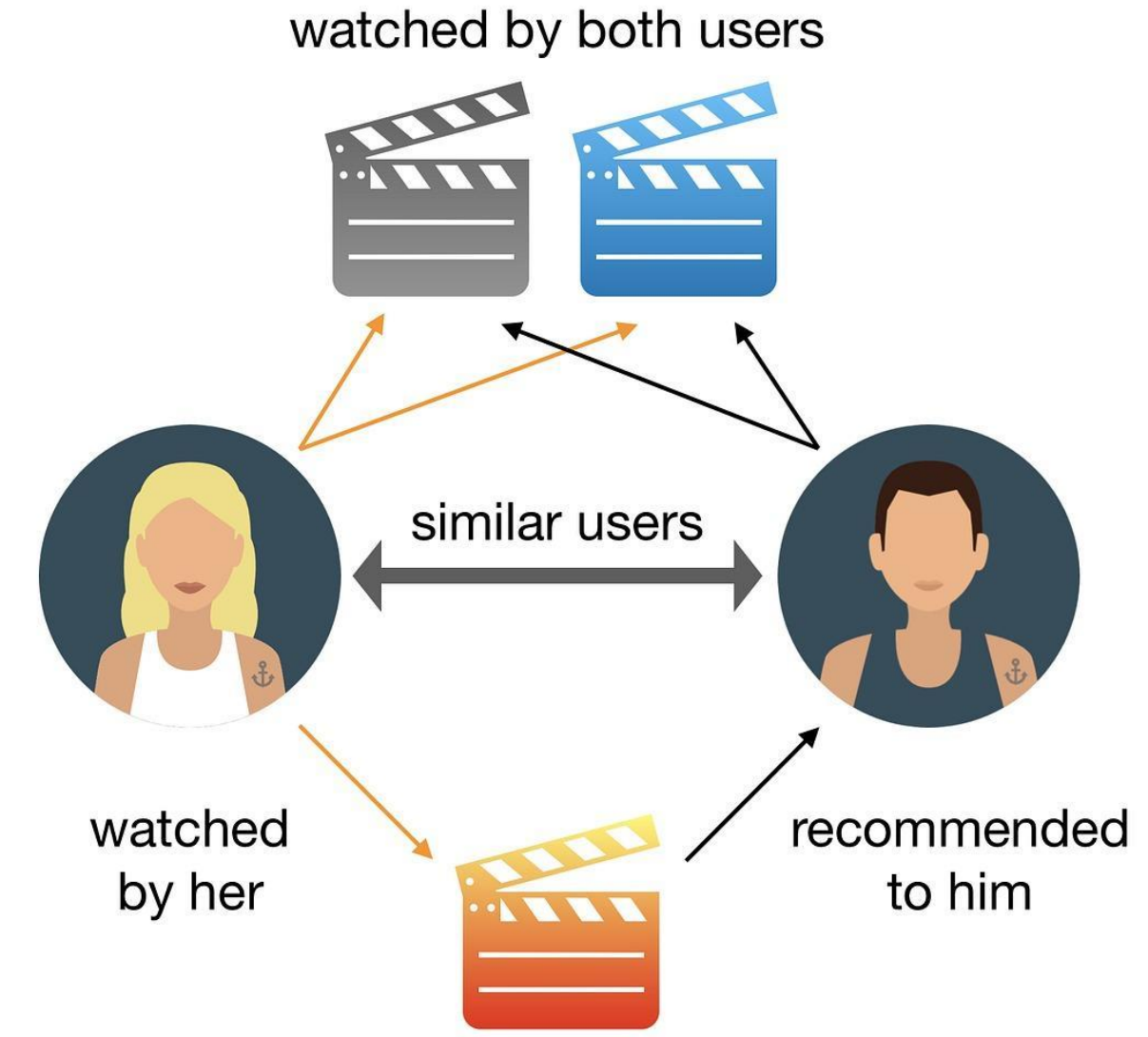
Variations on this type of recommendation system include:

- **By User Similarity:** This strategy involves creating user groups by comparing users' activities and providing recommendations that are popular among other members of the group. It is useful on sites with a strong but versatile audience to quickly provide recommendations for a user on which little information is available.
- **By Association:** This is a specific type of the one mentioned above, otherwise known as "Users who looked at X also looked at Y." Implementing this type of recommendation system is a matter of looking at purchasing sequences or purchasing groups, and showing similar content. This

strategy is useful for capturing recommendations related to naturally complementary content as well as at a certain point in the life of the user.

2 — Content-Based

Content-based systems make recommendations based on the user's purchase or consumption history and generally become more accurate the more actions (inputs) the user takes.



More specific types of content-based recommendation systems include:

- **By Content Similarity:** As the most basic type of content-based recommendation system, this strategy involves recommending content

that is close based on its metadata. This approach makes sense for catalogs with a lot of rich metadata and where traffic is low compared to the number of products in the catalog.

- **By Latent Factor Modeling:** Going one step further than the content similarity approach, the crux of this strategy is inferring individuals' inherent interests by assuming that previous choices are indicative of certain tastes or hobbies. Where the previous strategy is based on explicit, manually filled catalog metadata, this strategy hinges on discovering implicit relationships. This is done by using the history of users' larger interactions (e.g., movie watched, item purchased, etc.) to learn these tastes.
- **By Topic Modeling:** This is a variant of the Latent Factor Modeling strategy, whereby instead of considering users' larger actions, one would infer interests by analyzing unstructured text to detect particular topics of interest. It is particularly interesting for use cases with rich but unstructured textual information (such as news articles).
- **By Popular Content Promotion:** This involves highlighting product recommendations based on the product's intrinsic features that may make it interesting to a wide audience: price, feature, popularity, etc. This strategy can also take into account the freshness or age of the content and thus enable using the most trendy content for recommendations. This is often used in cases where new content is the majority.

The 6 Steps to Build a Recommendation System

Building a successful and robust recommendation system can be relatively straightforward if you're following the basic steps to grow from raw data to a prediction. That being said, there are some particularities to consider when it

comes to recommendation systems that often go overlooked and that, for the most efficient process and best predictions, are worth introducing (or reiterating). This section will walk through the six fundamental steps to completing a data project in the context of building a recommendation system.

1 — Understand the Business

Extremely simple and critical but often overlooked, the first step in building a recommendation system is defining the goals and parameters of the project. This will most definitely involve discussions between and input from both the data team as well as business teams (which might be product managers, operations teams, even partnership or advertising teams, depending on your product).

Here are some specific topics to consider to understand the business need more deeply and kickstart the discussion between these teams:

- *What is the end goal of the project?* Is the idea to build a recommendation system to directly increase sales / achieve a higher average basket size /

reduce browsing time and make a purchase happen faster / reduce the long tail of unconsumed content / improve user engagement time with your product?

- *Is a recommendation really necessary?* This is perhaps an obvious question, but since they can be expensive to build and maintain, it's worth asking. Can the business achieve its end goal by driving discovery via a static set of content instead (like staff/editor picks or most popular content)?
- *At what point will recommendations occur?* If recommendations make sense in multiple places (i.e., on a home screen upon first visiting the app or site as well as after purchasing or consuming content), will the same system be used in both places, or are the parameters and needs distinct for each?
- *What data is available on which to base recommendations?* At the time of recommendation, approximately what percentage of users are logged in (in which case there may be much more data available) vs. anonymous (which could complicate things for building the recommendation system)?
- *Are there product changes that must be made first?* If the team wants to build the recommendation system using more robust data, are there product changes that must be made first to identify users earlier (i.e., invite them to log in sooner), and if so, are they reasonable changes from a business perspective?
- *Should all content or products be treated equally?* That is, are there particular products or pieces of content that the business team wants to (or has to) promote aside from organic recommendations?
- *How can users with similar tastes be segmented?* In other words, if employing the model based on user similarity, how will you decide what makes users similar?

2 — Get the Data

The best recommendation systems use terabyte(s) of data. So when it comes to rounding up data to use for your recommendation systems, in general, the more the better. This can be difficult if users are unknown when you're trying to make a recommendation for them — i.e., they're not logged in or, even more challenging, they're brand new. If you have a business where most users are unknown, you may need to rely on external data sources or general data not explicitly tied to preferences, like demographics, browsing history, etc.

When it comes to user preferences, there are two kinds of feedback: explicit and implicit.

- **Explicit user feedback** is anything that requires user effort, like leaving a review/rating or initiating a complaint or product return (often from customer relationship management, CRM, data).
- By contrast, **implicit user feedback** is information that can be gathered about a user's preferences without them actually specifying those preferences. For example, past purchase history, time spent looking at certain offers, products, or content, data from social networks, etc.

Good recommendation systems usually employ a combination of these types of feedback since there are advantages and disadvantages to each.

- Explicit feedback can be very clear: a user has literally stated their preferences, likes, or dislikes. But by the same token, it's inherently biased; a user doesn't know what he doesn't know (in other words, he might like something but has never tried it and therefore wouldn't list it as a preference or interact with that type of item or content normally).

- By contrast, implicit feedback is the opposite — it can reveal preferences that a user didn't — or wouldn't — otherwise, admit to in a profile (or perhaps their profile information is stale). On the other hand, implicit feedback can be more complicated to interpret; just because a user spent time on a given item doesn't mean that (s)he likes it, so it's best to rely on a combination of implicit signals to determine preference.

3 — Explore, Clean, and Augment the Data

One thing to consider when exploring and cleaning your data for a recommendation system, in particular, is changing user tastes. Depending on what you're recommending, the older reviews, actions, etc., may not be the most relevant on which to base a recommendation. Consider only looking at features that are more likely to represent the user's current tastes and removing older data that might no longer be relevant or adding a weight factor to give more importance to recent actions compared to older ones.

Datasets for recommendation systems can be challenging to work with because they are commonly high dimensional, but at the same time, it's also common that many of the features don't have any values, which can make clustering and outlier detection difficult.

4 — Predict the Ranking

Given the work done in the previous steps, you could have already built a recommendation system, simply by ranking those scores by users and you'll have products to recommend. This strategy doesn't use machine learning or a predictive element, but that's totally fine. For some use cases, this is sufficient.

But if you do want to build something more complex, there are lots of subtasks that can be done after users consume recommended content that can be used to further refine the system. There are several ways to leverage the hybrid approach to try for the highest-quality recommendations:

- Presenting recommendations from different types of systems together side-by-side.
- Maintaining multiple algorithms in parallel where the decision of which algorithm is preferred over another is itself subject to machine learning (e.g., multi-armed bandit).
- Using a pure machine learning approach to combine multiple recommendation systems (logistic regression or other weighted regression methods). One specific example would be using a weighted average of two (or more) recommendations using different techniques.

It's also possible that different models will work better in different parts of the product or website. For example, the homepage where the user has yet to take action vs. after the user has clicked or consumed content in some way.

5 — Visualize the Data

In the context of recommendation systems, visualization serves 2 primary purposes:

1. When still in the exploration phases, visualizations can help reveal things about the data set or give feedback on model performance that would otherwise be difficult to see.
2. After putting the recommendation system in place, visualizations can help convey useful information to the business or product teams (e.g., which content does well but isn't being discovered, similarities between users'

tastes, content or products commonly consumed together, etc.) so they can make changes or decisions based on this information.

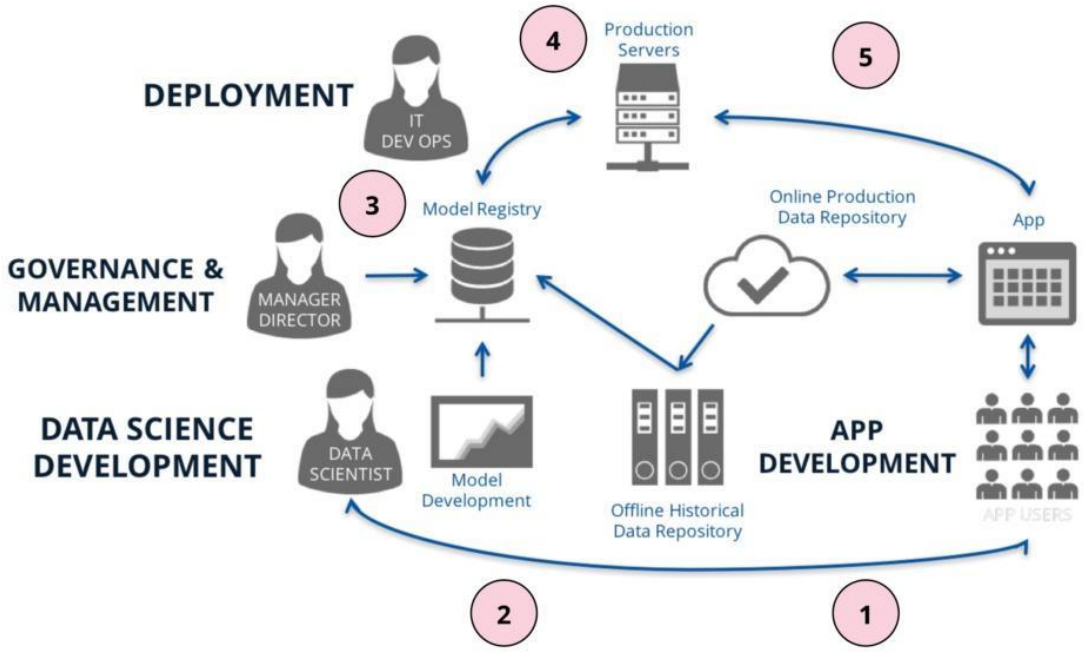
The primary issue with visualizing this type of data is the amount of data present, which can make it difficult to cut through the noise in a meaningful way. But by the same token, a good visualization will help make sense out of lots of data from which it would be otherwise difficult to derive meaningful insights.

6 — Iterate and Deploy Models

Recommendation systems that are working in a development environment or sandbox don't do any good. It's all about putting the system into production so that you can begin to see the effect on the business goals you've laid out in the beginning.

Additionally, keep in mind that the more data you have with which to feed the recommendation system, the better it can become. So with this type of data project perhaps more so than others, it's critical to evaluate performance and

continue to fine-tune, like adding new data sources to see if they have a positive effect.



In fact, making sure your recommendation system is built to adapt and evolve by regularly monitoring its performance is one of the most important parts of the process — a recommendation system that isn't properly adjusting to tastes or new data over time likely will not help you ultimately achieve your initial project goal, even if the system performed well at first. Building a feedback loop to understand whether or not users care about recommendations will be helpful and provide a good metric for making refinements and decisions going forward.

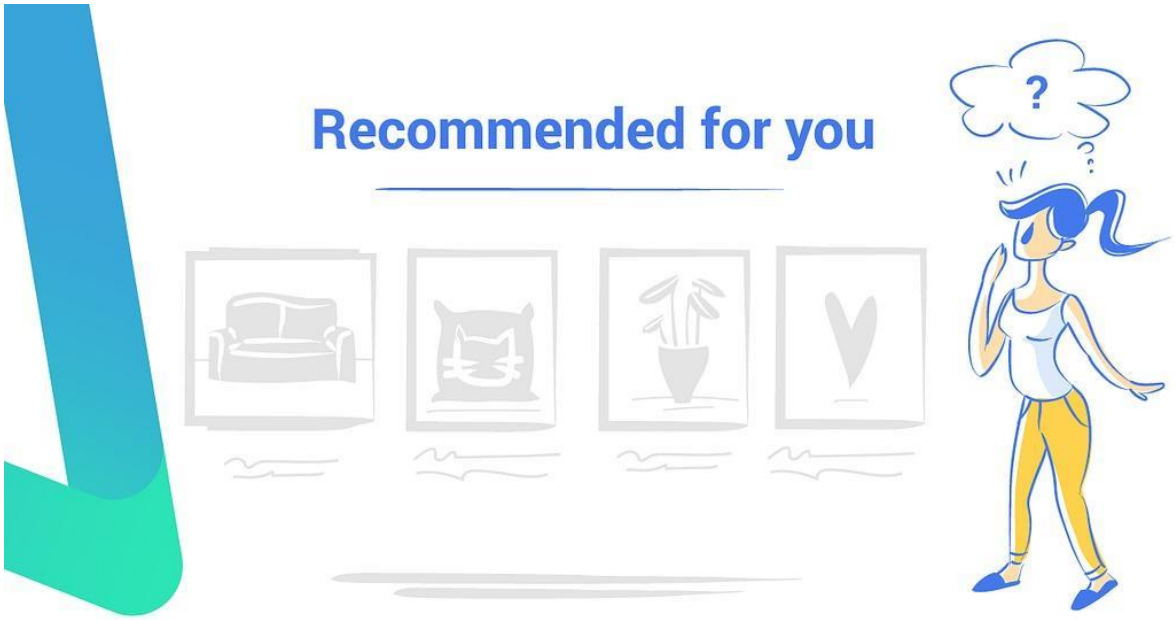
If recommendations are core to your business, constantly trying new things and evolving the initial model you've created will be an ongoing task; recommendation systems are not something you can create and cast aside.

Challenges

It's important to create a recommendation system that will **scale** with the amount of data you have. If it's built for a limited dataset and that dataset grows, computation costs grow exponentially, and the system will be unable to handle

the amount of data. To avoid having to rebuild your recommendation system later on, you must ensure from the beginning it is built to scale to expected data volumes.

It's also possible that after spending time, energy, and resources on building a recommendation system (and even after having enough data and good initial results) that the recommendation system only makes very **obvious recommendations**. The crux of avoiding this pitfall really harkens back to the first of the seven steps: understand the business need. If there isn't enough of a content long-tail or no need for the system, perhaps you need to reconsider the need to build a recommendation system in the first place.



Finally, people's tastes don't stay static over time, and if a recommendation system isn't built to consider this fact, it may never be as accurate as it could be. Similarly, there is a risk of building a recommendation system that doesn't get better over time. As users continue to consume content and more data is available, your recommendation system should learn more about users and adapt to their tastes. A recommendation system **not agile enough** to continue to adapt can quickly become obsolete and won't serve its purpose.

Future Work

Basic recommendation systems have been around for quite some time, though they continue to get more complex and have been perfected by retail and content giants. But what's next? What are the latest trends and developments that businesses should consider if they are looking to develop a truly cutting-edge system?

Context-aware recommendation systems represent an emerging area of experimentation and research, aiming to provide even more precise content given the context of the user in a particular moment in time. For example, is the user at home, or on the go? Using a larger or smaller screen? Is it morning or night? Given the data available on a certain user, context-aware systems may be able to provide recommendations a user is more likely to take in those scenarios.

Deep learning is already in use by some of the biggest and most powerful recommendation systems in the world (like YouTube and Spotify). But as the amount of data continues to skyrocket and more businesses find themselves up against a huge corpus of content and struggling to scale, deep learning will become the de facto methodology for not only recommendation systems but all learning problems.

Solving the cold-start problem is also something that cutting-edge researchers are starting to look at so that recommendations can be made for items on which there is little data. This is a critically important area for businesses with lots of turnover in content to examine so that they can successfully push items that will sell well (even before they know how that item will perform).

Conclusion

Recommendation systems can be an effective way to expose users to content they may not have otherwise found, which in turn can forward larger business goals

like increasing sales, advertising revenues, or user engagement. But there are a few key points to find success with recommendation systems. Namely, recommendation systems should be, above all, necessary.

Building a complex system that requires experienced staff and ongoing maintenance when a simpler solution will do is a waste of data team resources that could be spent elsewhere for more impact. The challenge lies in building a system that will actually have a business impact; building the system in and of itself shouldn't be the end goal.

Recommendation systems should also be agile. That is, adaptable and able to evolve as users do. Putting a recommendation system into production isn't the final step in the process; rather, it's an ongoing evolution, looking at what works, what doesn't, thinking about additional data sources that might help make better recommendations, etc.