IITK CONSULTING GROUP

Summer Project

Science and Technology Council, IIT Kanpur

NAVIGATING CONSUMER DECISION-MAKING IN E-COMMERCE

Mentors-

SHARAH PS
ISHAN PRAKHAR
PARV GOYAL
KANISHK GOYAL

Consult & Finance

1.2 What are consulting frameworks?

Frameworks for consulting are tools that managers can use to deliver key tasks efficiently. These tools guide them to make sure their teams reach their goals and follow their strategic plans. Often, consulting frameworks involve analyzing a business's performance and outside factors that may prevent it from succeeding.

There are multiple consulting frameworks but we will see a few of them:

1. SWOT Matrix
2. Pestel analysis
3. Porter's Five Forces
4. Ansoff's Growth Matrix

1.2.1 SWOT Analysis

SWOT Analysis is more of a mini-framework, specifically for quickly evaluating a single company in an industry. It evaluates both intrinsic and extrinsic factors based on facts and data to create realistic predictions.

SWOT is effectively a quick, high-level market landscape/competitive dynamics analysis arranged using the following terminology:

1. **Strengths:** Company strengths within an industry
2. **Weaknesses:** Company weaknesses within an industry
3. **Opportunities:** Company opportunities available within the industry (or potentially by branching into a new industry)
4. **Threats:** Company threats within the industry (or potentially from companies whose primary business is in another, related industry, or from disruptive technologies that potentially threaten all companies in an industry)

For Example, let us take the case study of Zomato:

1. **Strengths**
    a. Effective Marketing Strategy
    b. Strong brand recognition
2. **Weaknesses**
    a. Low customer retention
    b. High platform fee
3. **Opportunities**
    a. Tapping into remote locations
    b. Market growth(by CAGR)
4. **Threats**
    a. Data security issues
    b. Legal threats

1.2.2 PESTEL Analysis

PESTEL analysis is a method that managers can use to evaluate any major external factors that may influence their operations to help them stay competitive in the market. Those using this

framework can maximize their current conditions and prepare for future changes that may bring opportunities or challenges.

The key areas of PESTEL analysis are:

- **Political**: The political analysis examines how government regulations and other factors, such as tax guidelines, could affect trading and business.
- **Economic:** The economic analysis reviews financial issues, such as inflation and interest rates, that may affect an organization's success.
- **Social:** The social analysis evaluates characteristics, such as lifestyle and education, of their customers to understand their purchase needs better.
- **Technological:** The technological analysis determines how technology factors, such as technological advances, can positively or negatively influence how customers respond to a new product or service.
- **Environmental:** This covers ecological and environmental aspects such as climate change, environmental regulations, and sustainability efforts.
- **Legal:** It includes laws regarding consumer protection, antitrust laws, employment laws, and health and safety regulations

Let's do a PESTEL analysis on whether Tesla should enter the Indian market or not:

1. **Political**
    a. The government is supporting local players which is a setback for Tesla.
    b. Increased custom duties(200%) in India cause resistance for international players to enter the Indian market.
2. **Economical**
    a. In India targeting to launch a car under 20 lakhs will be a good choice.
    b. But, Tesla has very high R&D costs per car(approx 3000$ /car)
3. **Social**
    a. A very high proportion of the market acquired by TATA EVs
    b. People usually have resistance to switching their vehicle preferences from petrol to EVs.
4. **Technological**
    a. Ground clearance issues are prominent in India.
    b. Safety concerns about self-driving features can pose a threat to Tesla.
    c. Tesla has to get better energy density and faster charging time in their batteries to lure Indian customers.
5. **Environmental**
    a. The adverse effects of high temperatures on batteries can potentially threaten the life of Tesla cars in India.
    b. High water consumption in manufacturing is also a big factor to look at.
6. **Legal**
    a. Navigating the GST framework and ensuring compliance with tax regulations could be challenging.
    b. Acquiring land for manufacturing plants can be a complex process due to India’s land use regulations and bureaucratic procedures

1.2.3 PORTER'S 5 FORCES

Porter's five forces are used to identify and analyze an industry's competitive forces. The model guides businesses in determining the intensity of competition and potential profitability within their market, helping them better understand where power lies in their sector.

1. **Competitive Rivalry**: This force examines the degree of competition between existing firms in the industry. High rivalry limits profitability as competitors engage in price wars, advertising battles, and new product launches

   Factors to Consider: Number of competitors, rate of industry growth, product or service differentiation, brand loyalty, barriers to exit and switching costs..

2. **Threat of New Entrants**: This force assesses how easily new competitors can enter the market. High threat of new entrants can decrease market share and profitability for existing companies

   Factors to Consider: Barriers to entry (such as economies of scale, capital requirements, access to distribution channels, and government regulations), brand recognition, and customer loyalty, capital requirements, government policies, access to distribution channels, switching costs.

3. **Bargaining Power of Suppliers**: This force analyzes the power that suppliers have over the prices and quality of materials or services they provide. High bargaining power of suppliers can increase costs and reduce profitability for firms in the industry.

   Factors to Consider: Number and size of suppliers, uniqueness of supplier products or services, switching costs to change suppliers, focal company's ability to substitute.

4. **Bargaining Power of Buyers**:This force examines the influence that customers have on prices and quality. High bargaining power of buyers can drive prices down and demand higher quality or more services, thus reducing profitability.

   Factors to Consider: Number of buyers, volume of purchases, differentiation of products or services, price sensitivity, buyer's information and switching costs to change suppliers.

5. **Threat of Substitutes**: This force evaluates the likelihood that customers will switch to different products or services that perform the same function. High threat of substitutes can limit the potential for price increases and reduce industry profitability.

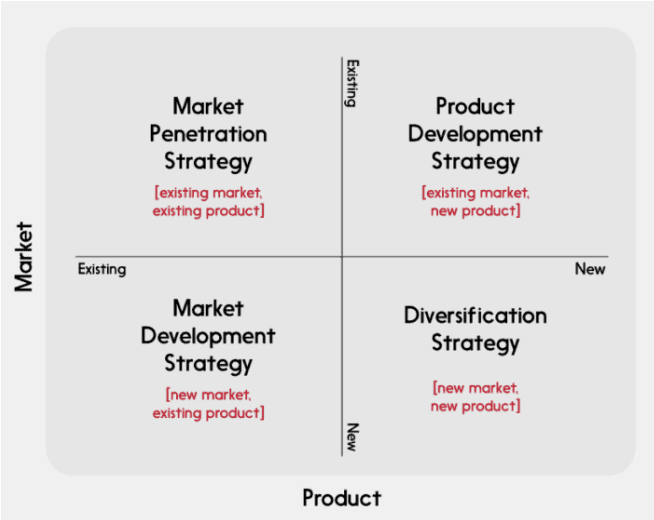
   Factors to Consider: Availability of substitute products or services, performance and cost of substitutes, switching costs and buyer willingness to switch to substitutes.

We did analysis of PhonePe vs Paytm , OTT vs cable operators, DABUR vs FORTUNE oil to understand the 5 forces in a better way

1.2.4 ANSOFF GROWTH MATRIX

The Ansoff Growth Matrix is a strategic planning tool designed to help businesses identify and plan their growth strategies. It focuses on two dimensions: products and markets. The matrix is divided into four quadrants, each representing a different growth strategy:

1. **Market Penetration**
2. **Market Development**
3. **Product Development**
4. **Diversification**



1. Market Penetration: Increasing market share for existing products in existing markets its objective is to grow market share within the current market by increasing the usage of current products among existing customers and attracting competitors' customers.

Strategies:

- Increase Usage: Encourage existing customers to use the product more frequently or in larger quantities.
- Attract Competitors' Customers: Offer better services, lower prices, or improved product features to lure customers away from competitors.
- Enhance Distribution: Expand the availability of the product by improving distribution channels.
- Promotions and Marketing: Use aggressive advertising, promotions, and sales tactics to boost market presence.

2. Market Development: Entering new markets with existing product, its objective is to grow by tapping into new geographic regions, new customer segments, or new uses for existing products.

Strategies:

- Geographic Expansion: Entering new regions or countries where the product is not currently sold.
- Target New Segments: Identifying and targeting new customer demographics that were previously untapped.
- New Distribution Channels: Using different distribution methods to reach new customers (e.g., online sales, direct sales).

3. Product Development: Developing new products for existing markets. Its objective is to grow by innovating and offering new products or improving existing products to the current customer base.

Strategies:

- Product Innovation: Creating new features or entirely new products that appeal to the existing market.
- Product Improvement: Enhancing the features, quality, or performance of existing products.
- Extending Product Lines: Adding new products that complement or are related to the existing product line.

4. Diversification: Entering new markets with new products its objective is to grow by developing new products for new markets, which can be related or unrelated to the current business.

Strategies:

- Acquisitions and Mergers: Acquiring or merging with companies in new industries or markets.
- New Ventures: Creating new businesses or subsidiaries to enter new markets.
- Strategic Alliances: Forming partnerships or alliances with companies in different industries.

We analyzed companies from various sectors of the market and had a thorough analysis of ITC from the FMCG sector.

1.3 What is E-commerce?

Ecommerce specifically refers to the online buying and selling of goods and services. It focuses on the commercial aspect of business conducted through digital platforms. E-commerce is powered by the internet. Customers use their own devices to access online stores. They can browse products and services those stores offer and place orders.

E-business, on the other hand, encompasses a broader scope including all digital operations of a company such as e-commerce, online marketing, customer relationship management, and supply chain management.

Some E-commerce models that we are going to analyze are :

1. B2C
2. B2B
3. D2C
4. Social
5. Recommerce
6. Roll Ups

1.3.1 B2C

The term business-to-consumer (B2C) refers to the process of selling products and services directly between a business and consumers who are the end-users of its products or services. Most companies that sell directly to consumers can be referred to as B2C companies.
Eg : Amazon , Flipkart , Myntra , Snapdeal , Voonik

1.3.2 B2B

It focusses on online transactions between businesses. This type of e-commerce involves companies that receive bulk orders of raw materials or office supplies online.
Eg : Industry Buying , Indiamart , Arzoo , ProcMart , Udaan

1.3.3 D2C

The D2C (Direct-to-Consumer) business model refers to a sales strategy where brands sell their products directly to customers, bypassing intermediaries such as retailers, wholesalers, and distributors.
Eg : Nykaa, Lenskart, Mama , Earth, BOAT ,Sugar

1.3.4 Social

It allows users to discover and purchase products directly through social media networks. It relies on influencers, user generated content, and personalized recommendations to drive sales.
Eg : Meesho , BulBul

1.3.5 Recommerce

Recommerce means buying and selling used or pre-owned products online. Online marketplaces for re-commerce allow individuals and businesses to sell second-hand goods.
Eg : Cashify, CarDekho, CredR, Olx

1.3.6 Roll Ups

A company acquires and invests in a number of smaller, independent brands. The goal is to create a portfolio of brands that can benefit from each other’s strengths and resources.

Eg : JOIN Ventures, Powerhouse91, Evenflow

1.4 Key Trends Shaping Indian E-commerce

- **M-commerce**: Consumers are progressively utilizing mobile applications for online shopping, resulting in the development of mobile-first ecommerce platforms and improved user experiences.
- **Hyperlocal Ecommerce**: Models for hyperlocal ecommerce have grown in popularity, especially for foodstuffs and essentials. These platforms connect local vendors with consumers to expedite and improve the dispatch of goods within a specific geographical area.
- **Individualisation & AI-powered Recommendations**: Advanced recommendation engines analyze customer behavior to provide product recommendations, resulting in increased customer engagement and conversions.
- **Omnichannel Retail:** The integration of offline and online channels is a developing trend in Indian ecommerce, referred to as omnichannel retail. Numerous traditional retailers are employing omnichannel strategies, enabling customers to shop online and pick up their purchases from nearby brick-and-mortar stores, thereby delivering a seamless purchasing experience.

1.5 Customer Behaviour Analysis

Customer behaviour analysis, as this process of understanding is called, can go to a granular level of detail on your customers. Customers don't always do what they say they will, so customer behaviour analysis can identify what's really happening when they encounter your brand.

1.5.1 What is the need for Customer Behaviour Analysis ?

- Identifying patterns helps you make accurate predictions for the future. Getting to grips with customer behaviour trends helps you to see the pattern in the way customers shop with you or use your services.
- Drilling down on existing customer behaviour helps you win over new customers.Completing customer behaviour analysis doesn't just give you insights into your existing customers – it can help you win over new ones.
- Personalizing customer experiences drive sales.You are constantly receiving relevant customer data. Customers are keen to tell you what they want, and when their experiences have not met or have exceeded their expectations. Tailoring your customer experience based on feedback and other customer data can go a long way to shaping customer behaviour, rather than just waiting for it to happen naturally.
- Understanding behaviour helps you to retain customers for longer. Analyzing customer behaviour will help you to identify pain points and inform the best solutions. Then you can optimize your journey accordingly.

1.5.2 How do you perform a customer behavior analysis?

- **Set out your segments :** You must have knowledge of your customer personas. For that you need to collect the personas and segment the customers accordingly.You should segment customers on the basis of:
  Demographics: Age, gender, income, location, family status, annual Income, education level
  Personal background: Hobbies, interests
  Professional information: Industry, job title, company size.
  Values and goals: Beliefs, aspirations both personal and professional
  Challenges: Personal pain points, worries, needs, problems to solve
  Use of your product/service: how your brand is relied upon in their life
  Identifying information: social media use, potential for being an influencer in their online or offline communities, communication preferences
  Objections or barriers to purchase: factors that might affect their choice to buy

- **Gather qualitative and quantitative data on customer behaviour :** Once you've identified who your customers are – particularly your revenue-generating customers – in segments, you're able to start evaluating their data for customer behaviour patterns. This information will fall into two types: **quantitative data and qualitative data.**

  Quantitative data will include information such as:

  -Purchase history (and product/service popularity)
  -Website visits and views
  -Social media engagement
  -Conversion reports for marketing/sales activity
  -How many customer service tickets they've raised and whether their issues were resolved quickly
  Quantitative data will describe what is happening when customers take action

  Qualitative data will cover information such as:

  -Direct customer feedback(collected through surveys)
  -Conversation Analytics data (such as emotion, intent and effort)
  Qualitative data can give you the "why" behind customer actions (or lack of, in some cases)in their own words.

- **Evaluate your data for behaviour insights :** Looking at the data you've gathered, you can start evaluating your information for behavioural insights.

  Customer behaviour: Consumer behaviour can fall into several types, and there are a few theories why a customer might behave the way they do. Your data might be reflected in some of the types of customer buying behaviour and theories outlined below.

➔ What are the 4 types of customer buying behaviour?
  - Extended Decision-Making: What research does a customer do and how much time do they invest before deciding to buy a product? This could include asking family and friends for references, reading reviews, looking at comparison sites and browsing a brand site for further information.

- Limited Decision-Making: Customers might be limited in what they buy due to availability. Are they buying a product because it's the only option on the market?
- Habitual Buying Behaviour: What do customers regularly seek out and buy? How does this differ from segment to segment?
- Variety-Seeking Buying Behaviour: Sometimes, there's several very similar options on the market. Customers might be driven to buy and try several of the same product over time to see what the differences are. Are your customers comparing you to others in your market offering the same thing?

➔ **What are the five consumer behaviour approaches?**
- The economic man approach: The theory that customers always choose the lowest price product when offered a range of similar products at varying prices. Customers are believed to be driven by making the "rational" decision when reconciling their need and the limited money they have to meet that need.
- The cognitive approach: The theory that consumers act with a particular mental process in mind. This includes recognising they have a need, searching for information, evaluating their choices, making a purchase and then evaluating whether that purchase was a good one.
- The psychodynamic approach: Based largely on Sigmund Freud's theories, this theory suggests that consumers are motivated to reduce conflict between what they want and what they should do. Consumers are believed to search for the maximum amount of gratification they can find while also doing what they "should" be doing according to society.
- The behaviourist approaches: This theory suggests that consumers' behaviour is shaped by stimuli and past experience. Negative and positive experiences serve as lessons to either avoid or do the same action again.
- The humanistic approach: This theory posits that consumers are all individuals, with their own subjective reasons for taking action. They're always self-interested and their purchases will demonstrate their individuality in some way

Then you can determine accordingly across your segments, what patterns can you see? For example, you might want to answer the following questions:
How does a customer access your brand? Is it through online searches for products, social media posts, marketing emails? When are they most likely to purchase a product in terms of day, week, month, season? What stops them from completing a purchase? Is it a broken payment system or a price point? What functionality and design features of your website or purchase platform were a problem for your customers?

- **<u>Adjust your customer journey and experience for better customer life time value :</u>**
Once you've analysed your data for customer behaviour analysis insights, you can more accurately see what the optimised experience is for your customers and how to deliver on it. You can start to take steps to minimise behaviour that you don't want to see – cart abandonment, high bounce rates, failure to add payment details – and maximise the behaviour you do want to see.
However, this might mean tweaking your customer experience and buyer journey to better encourage this behaviour. For example, if customers often buy two products together, bundling them together and advertising this new bundle might sell more because you anticipate customers' needs and reduce the effort of having to search for both products. Or perhaps sending a reminder email to ensure your new customer who has just downloaded your taxi app add their card details so they can start using the app.

1.6 Key Performance Indicators (KPIs)

KPIs are quantifiable values that measure the performance of business objectives of a firm. They often directly reflect the success or shortcomings of business operations and help companies assess the best areas to focus their efforts to improve their performance.

There can be 3 classifications of KPIs

1.6.1. Quantitative Indicators :

These are measured using data and expressed in numbers. There can be further two types of Quantitative Indicators:

i)Continuous: These can take any values (including decimals). Eg. Average Order Value (AOV), Conversion Rate (CR).

ii)Discrete: Generally, count factors that can only include whole numbers. Eg. Number of complaints, accidents, new customers etc.

1.6.2. Qualitative Indicators:

These aren't measured using numbers instead to analyse these firms conduct surveys to understand the general situation associated with the functioning of the firm. Qualitative indicators tend to focus more on experiences or feelings and the intangible value we place on them. Some use cases for Qualitative indicators can be Product research and User research.

3.Process Indicators: Process indicators are used specifically to gauge the performance and efficiency of a process and facilitate any needed changes. A very common process indicator for support teams are KPIs focused on customer support tickets

There are a plethora of KPIs present but not all are equally relevant to every type of business. Different businesses use a combination of different KPIs according to their target market, their internal organization and their business model in order to effectively analyze the needs of their respective firms.

As this project is focused on E-commerce, we will briefly discuss some of the KPIs that are most relevant to an E- com business.

1. Conversion Rate (CR): Measure of users who complete a desired action. This desired action may click on an ad after seeing it or to buy a subscription after a trial period.

Importance: Tracking conversion rates allows you to measure the performance of your web pages and apps.

Factors influencing CR: UI/UX, AI based suggestions, Feedback mechanisms etc.

Formula: (Number of sales/Number of Visitors) * 100

2. Average Order Value (AOV): Self explanatory name.

Importance: Can help in developing price strategies to increase order size, Understanding consumer spending behaviour.

Strategies to increase AOV: Combo offers, Cross selling and Upselling, Offering free delivery over a threshold value and Features like flipkart coin etc.

Formula: (Total Order Revenue/ Number of Orders)

3. Customer lifetime value (CLV): Total revenue a company estimates to be generated from a single customer over the duration of relationship which may differ depending upon the services that company provides.

Importance: This metric lets companies analyse how much they should spend on customer acquisition and retention.

Strategies to increase CLV: Subscriptions based services, Supplementing offers, Warranty and customer service and Loyalty Programs etc.

Formula: Profitability* Customer revenue per year * duration of Relationship (in year) – Total cost of acquiring the customer.

4. Cart Abandonment Rate (CAR): Percentage of customers who add items to their cart but don't go ahead and complete the purchase.

Importance: This KPI can be instrumental in pointing out the issues with the checkout or the transaction part of the process.

Factors affecting CAR: More expensive product than competition, Complex ordering and payment interface, Additional charges like Shipping and Handling charges etc.

5. Customer Acquisition Cost (CAC): Cost associated with getting new customers aboard.

Importance**:** This KPI can help companies analyse the efficiency of their marketing features and other efforts in new customers aboard.

Factors affecting CAC: Quality of marketing campaigns, USPs of the platform, UI/UX, Discounts etc.

Formula: (total expense on marketing and sales)/(Number of customers acquired) (*for a particular time period)

6. Repeat Purchase Rate (RPR): Percentage of customers that are making more than one purchase from the platform.

Importance: Companies can't just depend upon the new customers for their revenue. And having a maintained customer base is very important for any firm to sustain. RPR is an indicator that helps firms track that aspect of the business.

Strategies for increasing RPR: Loyalty programs, Customer Satisfaction and Support.

Formula: (No. of Repeat customers)/ (total number of customers) * 100

7. Churn Rate: Percentage of customers who stop doing business with a company in a given period of time.

Importance: Its trend indicates a company's ability to hold a customer and it can indicate problems with the company's customer support system.

Strategies to reduce Churn Rate: Maintaining social media presence, Providing 24*7 solid technical support systems and Customer Relationship Management systems and Maintaining Communities and resolving issues that appear on them.

Formula: (Number of customers lost / total customers at the start) * 100

8. Purchase Frequency: Number of purchases made by a customer within a defined timeframe.

Importance: Can help a firm understand a customer's buying habits.

Strategies to increase Purchase frequencies: Periodic discount festivals like amazon's great Indian festival and Flipkart's big billion days sales.

Formula: Total number of Orders/ Number of Unique customers.

9.Inventory Turnover Ratio: It is the metric that shows how many times a company completely sold and replenished its inventory in a given period.

Importance: It can be instrumental in analyzing the company's functional and supply chain related efficiency.

A high Inventory Turnover Ratio means that the supply chain is efficient and in healthy state but a low Inventory Turnover means that the company may be struggling with overstocking issues or it can be slow in manufacturing finished products.

Formula: Cost of Goods Sold (COGS)/ Average Value of Inventory

1.7 Conversion Funnel Analysis

1.7.1 What is a conversion funnel?

A conversion funnel is a commonly used marketing and business-analytics tool. They help brands understand how effective their marketing is, how they can improve sales, spot and solve problems and improve the overall efficiency of their business.
In a nutshell, a conversion funnel is a distillation of the buyer's journey. The journey begins when a visitor discovers the brand or a specific product and ends when the visitor abandons the process or "converts" into a paying customer or buyer.

Broad Subdivisions of the funnel-

- The "top-of-funnel" (or "TOFU") is the broadest part of the funnel because it represents the greatest number of people. For example, every brand can expect to get more website visitors than it gets actual, paying customers from its website conversion funnel.

- The "middle-of-funnel" (or "MOFU") includes most of the stages of the consumer's journey.

- The “bottom-of-funnel” (or “BOFU”) describes the endpoint of the funnel when a visitor converts into a paying customer. It’s the narrowest part of the funnel because it’s the section that the fewest visitors reach (for a variety of reasons).

1.7.2 Main Stages of an E-Commerce Funnel-

1) **Awareness Stage**- Awareness includes attracting customers to the brand or making your target audience members aware of the company. During the awareness stage, the company pulls customers in via marketing efforts, reputational improvements, or other means of boosting brand awareness.
   Some examples of marketing to improve customer awareness include:

- Online ads
- Affiliate marketing
- Blogs and organic search
- Radio ads or TV ads
- Print ads
- Podcasts
- Social media marketing
- Video marketing on YouTube and other platforms

2) **Interest Stage**- Dubbed the interest or consideration phase, it’s here that the company tries to build trust and desire with your prospects or leads. It’s not enough for prospective customers to know that your brand makes products they can use. They have to *choose* your brand from all the competing companies or brands in the industry.

   The interest stage is significantly affected by things like:

- The type of content on social media channels and website
- The brand reputation, which may make it more likely that a potential customer will engage with the company
- Social media marketing

3) **Desire Phase**- During the “desire” phase, the brand must drive potential buyers’ interest and show them why they should purchase their product instead of another brand’s.
   In this phase, it's important to leverage attention-grabbing elements like:

- Phenomenal product descriptions for every item on your online store
- Excellent images, such as enticing photos
- Videos of products or services where applicable
- Downloadables like whitepapers and case studies
- Good product and website copy
- Excellent social media influencing

   If the brand already has some sales, then the company should leverage positive reviews to encourage new customers to buy their products. Can also increase desire in the customers through ancillary strategies, like generous promotional offers, free shipping and informative content marketings.

4) **Conversion Phase**- At the conversion stage of the funnel, things have begun to narrow significantly. It’s time to convert leads into customers.

Also sometimes called the action stage, the conversion stage is met when a customer makes a purchase or subscribes to the brand. It can also include other forms of conversion, like signing up for a newsletter.

One can improve the likelihood of prospects reaching this phase of the conversion funnel by:

- Using excellent email marketing strategies, if conversion rate refers to email sign-ups
- Streamlining the checkout process to remove distracting elements
- Using A/B testing where applicable
- Offering free or discounted shipping to sweeten the deal

5) **Re-Engagement Phase**- Many people think the funnel ends once the customer makes a purchase. But post-purchasers are part of the funnel too. These people are highly valuable - they can repurchase, leave a review, or refer a friend to purchase.

   Plus, it's almost always better to retain current customers than constantly attract new ones. This improves the customer lifetime value or CLV of each customer for the brand.

Ways to create an effective Conversion Funnel-

1) Determine an ideal buyer journey and map it out as a funnel
2) Set goals for each funnel stage
3) Implement strategies and create content to generate awareness
4) Generate interest and desire
5) Encourage users to take action

1.7.3 How to optimize your conversion funnel?

There are many ways you can optimize your company's conversion funnel. However, different strategies are best utilized for different stages of the funnel.

- Lower funnel stages are primarily about building trust and improving customer experiences and satisfaction with products. Here, optimization strategies should lean heavily into demonstrating the benefits of the brand or products and building trust with leads to convince them to buy.
- Middle funnel strategies should be about driving traffic to your site and specific products, creating personalized or targeted experiences, and convincing customers to buy from you rather than a competitor.
- Top-of-the-funnel stages are about bringing people to your brand in the first place. Optimization efforts heavily rely on marketing improvements.

1.7.4 Importance of Conversion Funnel-

- It helps to illustrate the consumer journey, which can be important for understanding how visitors navigate through the site/store
- It can break down which stage of the buyer's journey has the most failure

  These elements are important for boosting the brand's conversion rate. Conversion rate is the percentage of site visitors who "convert" into customers or subscribers. A higher conversion rate is always better since it means more people are engaging with your brand or making purchases.

Apart from this, there are a number of other KPIs used while measuring a conversion funnel like repurchase rate, retention rate, average order value, customer lifetime value, add to cart rate and cart abandonment rate.

2. DATA SCIENCE

2.1 Python Libraries

Python libraries pandas, NumPy, Matplotlib, and Seaborn were utilized to handle, analyze, and visualize data efficiently. These libraries were integral in transforming raw data into meaningful insights.

2.1.1 **Pandas**

Pandas is a powerful library for data manipulation and analysis.

- **DataFrames and Series**: Create and manipulate Data Frames and Series, which are essential data structures in pandas.
- **Data Cleaning**: Techniques such as handling missing values(**'dropna', 'fillna'**), filtering data using conditions, and converting data types are crucial.
- **Data Operations**: Grouping data (**'groupby'**), merging/joining datasets ('**merge', 'concat'**), and pivoting data ('**pivot_table**') prepare the data for analysis.
- Data Analysis: Statistical analysis (mean, median, mode, std, var, Data visualization (plot, hist, scatter, bar), Correlation and covariance analysis (corr, cov), Data filtering and subsetting (query, loc, iloc)

2.1.2 **NumPy**

NumPy is fundamental for numerical computations.

- **Arrays**: NumPy arrays are multidimensional arrays that are more efficient than Python lists for numerical operations. They support various data types, including integers, floating-point numbers, and complex numbers.
- **Mathematical Functions**: Functions like '**mean', 'median', 'std',** and mathematical operations ('**add', 'subtract', 'multiply', 'divide'**) are essential for statistical analysis.
- **Array Manipulation**: Reshaping arrays ('**reshape'**), stacking ('**vstack**', '**hstack'**), and splitting arrays ('**split'**) allow for flexible data manipulation.

2.1.3 **Matplotlib**

Matplotlib is a versatile plotting library.

- **Basic Plots**: Create line plots, bar charts, histograms, and scatter plots to visualize data trends.
- **Customization**: Customize plots by adding titles, labels, legends, and adjusting figure size to make the visualizations more informative.

- **Subplots**: Create multiple plots in a single figure (‘**subplot’, ‘subplots’**) to allow for comparative data analysis.

2.1.4 **Seaborn**

Seaborn builds on Matplotlib and provides high-level interfaces for drawing attractive statistical graphics.

- **Enhanced Visualizations**: Built-in themes and color palettes make the plots aesthetically pleasing.
- **Statistical Plots**: Advanced plots like pair plots, heatmaps, and violin plots provide deeper insights into data distributions and relationships.
- **Integration with Pandas**: Seamless integration with Pandas DataFrames makes it easy to create complex visualizations directly from the data.

2.2 Machine learning

Machine learning (ML) is a subset of artificial intelligence (AI) that involves developing algorithms and statistical models to enable computers to perform tasks without explicit instructions. Instead of being programmed to follow specific rules, machine learning systems learn from data, identify patterns, and use them to make decisions or predictions.
It has mainly three types:

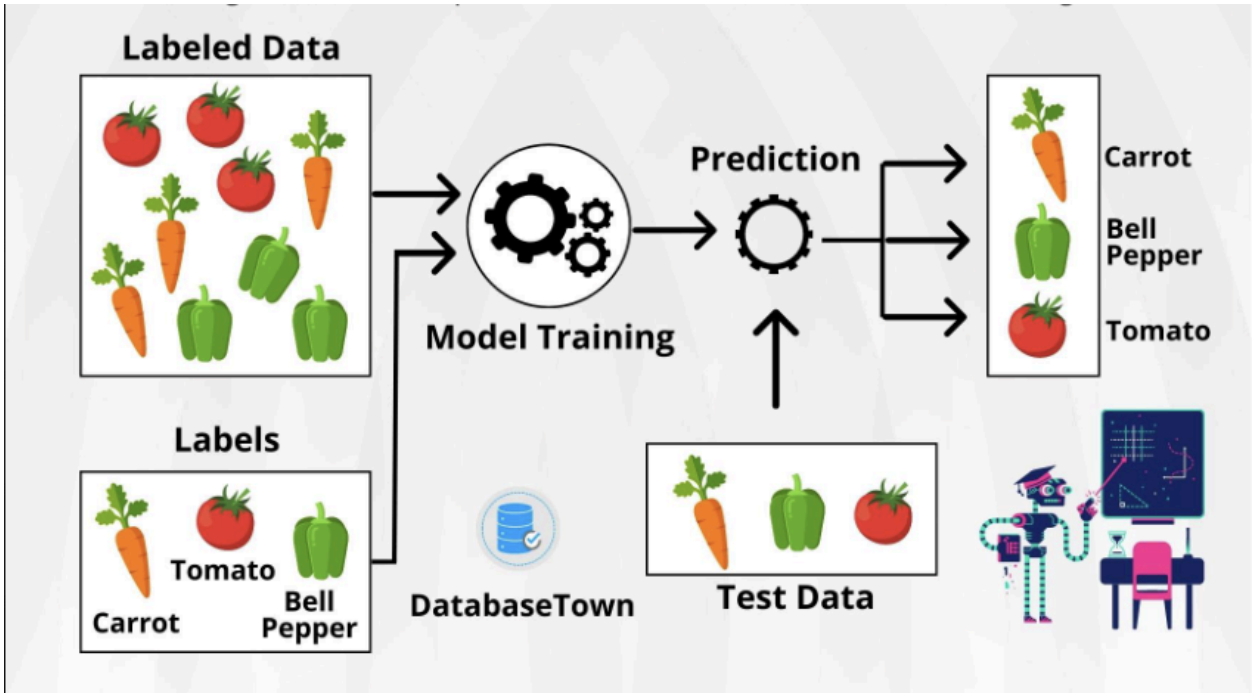
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

2.3 Supervised learning

Supervised learning is a machine learning task that involves learning a function to map an input to an output based on example input-output pairs. The data used in supervised learning is labeled. Supervised learning models are trained on labeled datasets, and their performance is evaluated based on how accurately they predict the correct outputs.

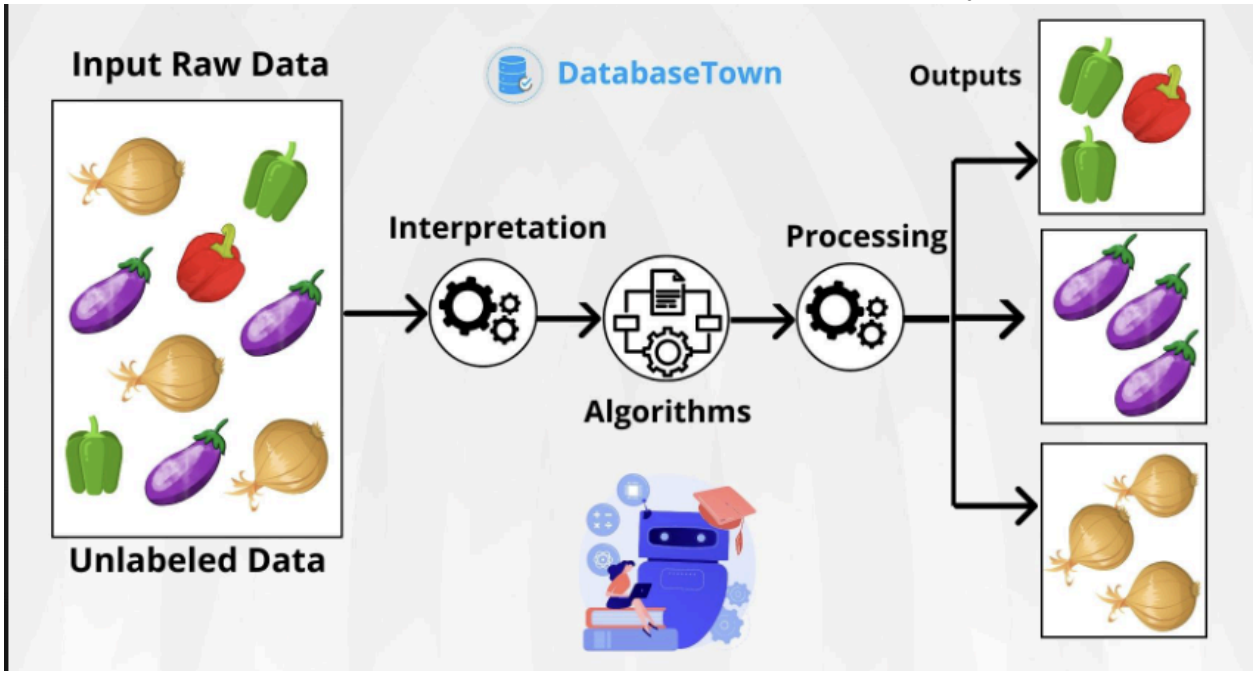
Classification and Regression are the two main types of supervised learning problems:

- **Classification**: The goal is to categorize examples into discrete classes. For instance, predicting whether an email is spam or not (binary classification), diagnosing diseases, and recognizing hate speech. Features like text in a review are analyzed to determine the category an example belongs to.
- **Regression**: The aim is to predict continuous numerical values. For example, predicting house prices based on features such as size and location. Regression answers questions like "How much?" or "How many?".



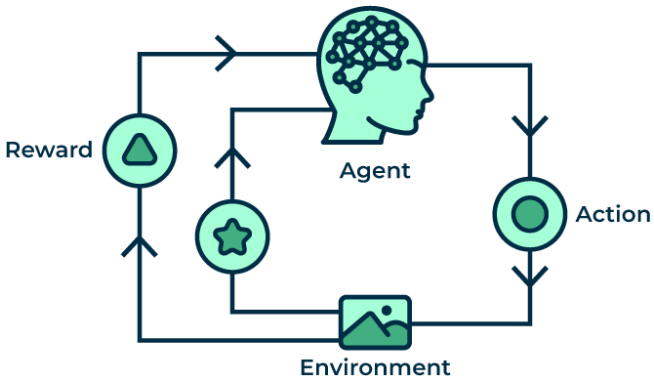
2.4 Unsupervised learning

Unsupervised learning is a machine learning technique where an algorithm discovers patterns and relationships using unlabeled data. It doesn't require labeled target outputs. As a kind of learning, it resembles the methods humans use to figure out that certain objects or events are from the same class, such as by observing the degree of similarity between objects. The goal is to uncover hidden patterns, similarities, or clusters within the data for purposes like data exploration, visualization, and dimensionality reduction.



2.5 Reinforcement learning

1. Reinforcement learning (RL) is a type of machine learning where an agent learns to make decisions by taking actions in an environment to maximize some notion of cumulative reward. In RL, the learner needs to be told what actions to take as in most forms of machine learning but instead must discover which actions yield the most reward by trying them out. For example, think of teaching a dog a new trick: we cannot tell the dog what to do or what not to do, but we can reward or punish it based on whether it does the right or wrong thing.



2.6 EDA

Exploratory Data Analysis (EDA) is a method of analyzing datasets to understand their main characteristics. It involves summarizing data features, detecting patterns, and uncovering relationships through visual and statistical techniques. EDA helps in gaining insights and formulating hypotheses for further analysis.

Exploratory Data Analysis (EDA) is crucial for understanding datasets, identifying patterns, and informing subsequent analysis

Step 1: Import Python Libraries

Import all libraries that are required for our analysis, such as Data Loading, Statistical analysis, Visualizations, Data Transformations, Merge and Joins, etc.

- Pandas and Numpy have been used for Data Manipulation and numerical Calculations
- Matplotlib and seaborn have been used for data visualizations

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt
```

```
import seaborn as sns
```

Step 2: Reading Dataset

The Pandas library offers a wide range of possibilities for loading data into the Pandas.

Most of the data are available in a tabular format of CSV files. It is trendy and easy to access. Using the read_csv() function, data can be converted to a pandas DataFrame.

```
df=pd.read_csv(‘used_cars.csv”)
```

Step 3: Analyzing the Data

Before we make any inferences, we listen to our data by examining all variables in the data. The main goal of data understanding is to gain general insights about the data, which covers the number of rows and columns, values in the data, their data types, and missing values in the dataset.

- df.shape() will display the number of observations(rows) and features(columns) in the dataset
- df.head() will display the top 5 observations of the dataset
- df.tail() will display the last 5 observations of the dataset
- df.info() helps to understand the data type and information about data, including the number of records in each column, data having null or not null, Data type, the memory usage of the dataset

Step 4: Data Cleaning

- Identify and handle missing values using methods like df.isnull().sum()
- Find and address duplicates with df.duplicated.sum()

Step 5: Data Reduction

Some columns or variables can be dropped if they do not add value to our analysis.

```
# to remove S.No. column from data

df=df.drop([‘S.No.’], axis=1)
```

Step 6: Statistical Summary

This information gives a quick and simple description of the data.

Can include Count, Mean, Standard Deviation, median, mode, minimum value, maximum value, range, standard deviation, etc.

A statistics summary gives a high-level idea to identify whether the data has any outliers, data entry errors or distribution of data such as the data is normally distributed or left/right skewed.

This can be achieved using the describe() function that gives all statistics summary of data.

Step 7- EDA Univariate Analysis

Analyzing/visualizing the dataset by taking one variable at a time:

Data visualization is essential; we must decide what charts to plot to better understand the data. Here, we visualize our data using Matplotlib and Seaborn libraries.

Matplotlib is a Python 2D plotting library used to draw basic charts.

Seaborn is also a Python library built on top of Matplotlib that uses short lines of code to create and style statistical plots from Pandas and Numpy

Univariate analysis can be done for both Categorical and Numerical variables.

Categorical variables can be visualized using a Count plot, Bar Chart, Pie Plot, etc.

Numerical Variables can be visualized using Histogram, Box Plot, Density Plot, etc.

Step 8: EDA Bivariate Analysis

Bivariate Analysis helps to understand how variables are related to each other and the relationship between dependent and independent variables present in the dataset.

For Numerical variables, Pair plots and Scatter plots are widely used to do Bivariate Analysis.

A Stacked bar chart can be used for categorical variables if the output variable is a classifier.

A bar plot can be used to show the relationship between Categorical variables and continuous variables

2.7 Feature Engineering

2.7.1 What is a feature?

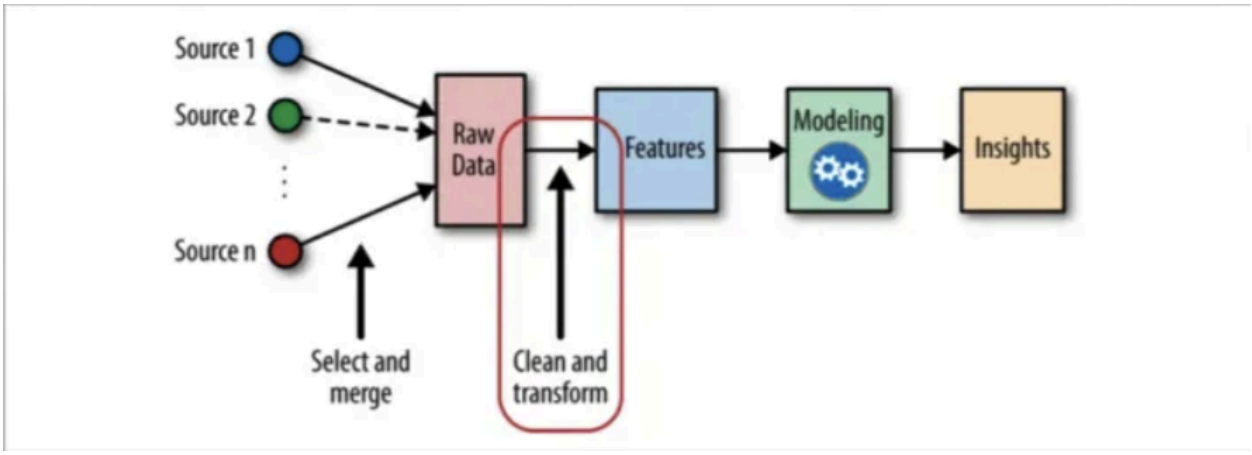
Feature is an attribute or property shared by all the independent units on which analysis or prediction is to be done. Common feature types:

- Numerical: Values with numeric types (int, float, etc.). Examples: age, salary, height.
- Categorical Features: Features that can take one of a limited number of values. Examples: gender (male, female, non-binary), color (red, blue, green).
- Ordinal Features: Categorical features that have a clear ordering. Examples: T-shirt size (S, M, L, XL).
- Binary Features: A special case of categorical features with only two categories. Examples: is_smoker (yes, no), has_subscription (true, false).

- Text Features: Features that contain textual data. Textual data typically requires special preprocessing steps (like tokenization) to transform it into a format suitable for machine learning models.

2.7.2 Steps of Feature engineering

Feature engineering, in data science, refers to manipulation — addition, deletion, combination, and mutation of your data set to improve machine learning model training, leading to better performance and greater accuracy. Feature engineering involves transforming raw data into a format that enhances the performance of machine learning models.



The key steps in feature engineering include:

- Data Exploration and Understanding: Explore and understand the dataset, including the types of features and their distributions. Understanding the shape of the data is key.
- Handling Missing Data: Address missing values through imputation or removal of instances or features with missing data. There are many algorithmic approaches to handling missing data.
- Variable Encoding: Convert categorical variables into a numerical format suitable for machine learning algorithms using methods.
- Feature Scaling: Standardize or normalize numerical features to ensure they are on a similar scale, improving model performance.
- Feature Creation: Generate new features by combining existing ones to capture relationships between variables.
- Handling Outliers: Identify and address outliers in the data through techniques like trimming or transforming the data.
- Binning or Discretization: Convert continuous features into discrete bins to capture specific patterns in certain ranges.

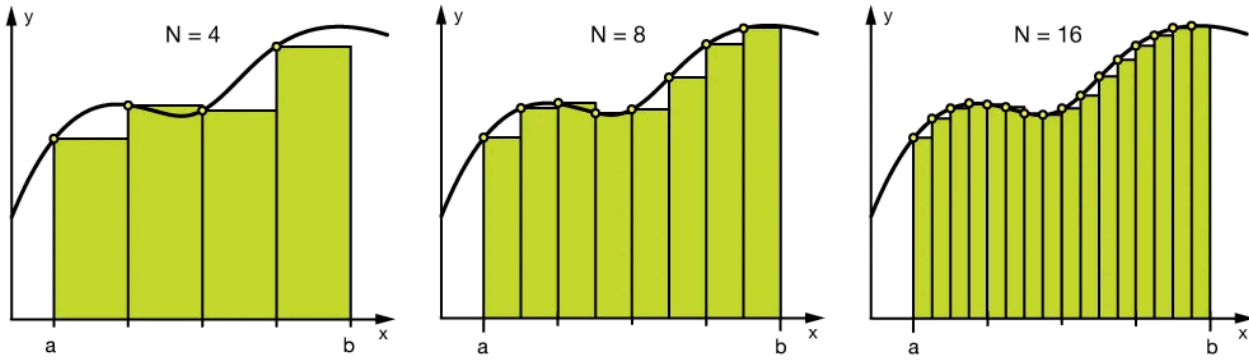
Here we will explore a few steps.

1. **Imputation**: Missing values are one of the most common problems you can encounter when you try to prepare your data for machine learning. **Imputation** is the act of replacing missing data with statistical estimates of the missing values. The goal of any imputation technique is to produce a complete dataset that can be used to train machine learning models.
- There are multiple techniques for missing data imputation. These are as follows:-
    1. Complete case analysis: Complete case analysis implies analyzing only those observations in the dataset that contain values in all the variables. In other words, in complete case analysis, we remove all observations

with missing values. This procedure is suitable when there are few observations with missing data in the dataset.

2. Mean / Median / Mode imputation: We can replace missing values with the mean, median, or mode of the variable. Mean/median / mode imputation is widely adopted in organizations and data competitions. Mean/median imputation consists of replacing all occurrences of missing values (NA) within a variable with the mean (if the variable has a Gaussian distribution) or median (if the variable has a skewed distribution). This is for numerical imputations. For categorical variables, replacement by the mode is also known as replacement by the most frequent category.

2. **Handling outlier:** Outliers are values that are unusually high or unusually low with respect to the rest of the observations of the variable. The outliers can be dropped or capped. There are many techniques for outlier handling. Here are a few:
    a. **Outlier Detection with Standard Deviation:** If a value has a distance to the average higher than ***x * standard deviation*****,** it can be assumed as an outlier. Then what **x** should be? There is no trivial solution for x, but usually, a value between 2 and 4 seems practical.
    b. **Outlier Detection with Percentiles:** Another mathematical method to detect outliers is to use percentiles. You can assume a certain percent of the value from the top or the bottom as an outlier. The key point is here to set the percentage value once again, and this depends on the distribution of data.
3. **Binning:** It can be done on both numerical and categorical variables.



The main motivation of binning is to make the model more **robust** and prevent **overfitting**, however, it has a cost to the performance.

```
#Numerical Binning Example

Value      Bin
0-30   ->  Low
31-70  ->  Mid
71-100 ->  High

#Categorical Binning Example

Value      Bin
Spain  ->  Europe
Italy  ->  Europe
Chile  ->  South America
Brazil ->  South America
```

4. **Variable encoding or categorical encoding**: Categorical data is data that takes only a limited number of values. For example, if you people responded to a survey about which brand of car they owned, the result would be categorical (because the answers would be things like Honda, Toyota, Ford, None, etc.). Responses fall into a fixed set of

categories. Categorical variable encoding is a broad term for collective techniques used to transform the strings or labels of categorical variables into numbers. There are multiple techniques under this method:

a. One-Hot encoding (OHE)
b. Ordinal encoding
c. Count and Frequency encoding
d. Target encoding / Mean encoding
e. Weight of Evidence
f. Rare label encoding

a. **One-Hot Encoding (OHE)**: OHE is the standard approach to encode categorical data. One hot encoding (OHE) creates a binary variable for each one of the different categories present in a variable. These binary variables take 1 if the observation shows a certain category or 0 otherwise. OHE is suitable for linear models. But, OHE expands the feature space quite dramatically if the categorical variables are highly cardinal, or if there are many categorical variables. In addition, many of the derived dummy variables could be highly correlated. For example, for the categorical variable "Gender", with labels 'female' and 'male', we can generate the boolean variable "female", which takes 1 if the person is female or 0 otherwise. We can also generate the variable male, which takes 1 if the person is "male" and 0 otherwise.

```
0      male
1    female
2    female
3    female
4      male
Name: Sex, dtype: object
```

	Sex	female	male
0	male	0	1
1	female	1	0
2	female	1	0
3	female	1	0
4	male	0	1

b. **Ordinal encoding**: Categorical variables which can be meaningfully ordered are called ordinals. For example, a Student's grade in an exam (A, B, C or Fail). When the categorical variable is ordinal, the most straightforward approach is to replace the labels by some ordinal number. In ordinal encoding, we replace the categories by digits, either arbitrarily or in an informed manner. If we encode categories arbitrarily, we assign an integer per category from 1 to n, where n is the number of unique categories. If instead, we assign the integers in an informed manner, we observe the target distribution: we order the categories from 1 to n, assigning 1 to the category for which the observations show the

highest mean of the target value, and n to the category with the lowest target mean value.

c. **Count or frequency encoding**: In count encoding, we replace the categories with the count of the observations that show that category in the dataset. Similarly, we can replace the category by the frequency or percentage of observations in the dataset. That is, if 10 of our 100 observations show the colour blue, we would replace blue by 10 if doing count encoding, or by 0.1 if replaced by the frequency. These techniques capture the representation of each label in a dataset, but the encoding may not necessarily be predictive of the outcome.
d. **Target encoding**: In target encoding, also called mean encoding, we replace each category of a variable, with the mean value of the target for the observations that show a certain category. For example, we have the categorical variable “city”, and we want to predict if the customer will buy a TV provided we send a letter. If 30 percent of the people in the city “London” buy the TV, we would replace London with 0.3.

- This technique has 3 advantages:
    1. it does not expand the feature space,
    2. it captures some information regarding the target at the time of encoding the category, and
    3. it creates a monotonic relationship between the variable and the target.
- Monotonic relationships between variable and target tend to improve linear model performance

e. **Label encoding**: It refers to simply assigning integer values to the categorical variable without considering any order.

2.8 Linear and Logistic Regression

2.8.1 Linear Regression

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data. When there is only one independent feature, it is known as Simple Linear Regression, and when there is more than one feature, it is known as Multiple Linear Regression. Similarly, when there is only one dependent variable, it is considered Univariate Linear Regression, while when there is more than one dependent variable, it is known as Multivariate Regression.

Simple Linear Regression

Simple linear regression focuses on understanding the relationship between a single independent variable (X) and a dependent variable (Y). The model equation is:

- $Y = \beta_0 + \beta_1 X + \varepsilon$

where:

- Y is the dependent variable
- X is the independent variable
- β_0 is the y-intercept
- β_1 is the slope
- ε is the error term

Multiple Linear Regression

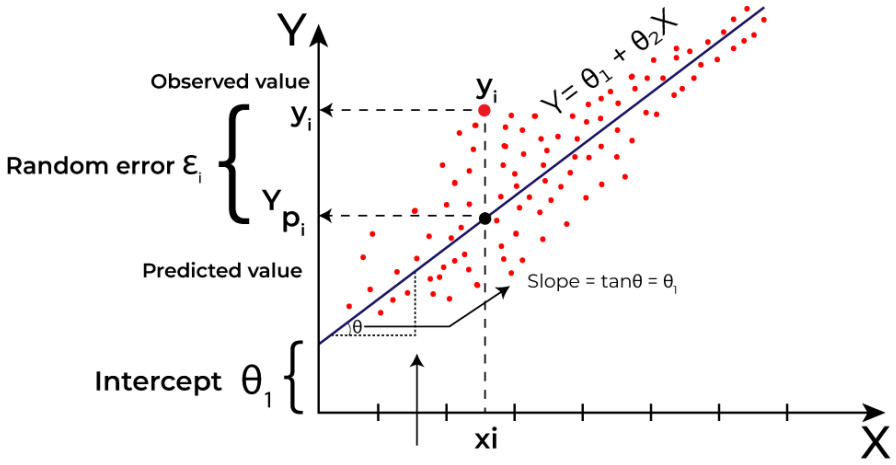
Multiple linear regression extends the concept to analyze the relationship between a dependent variable (Y) and two or more independent variables (X_1, X_2, ..., Xn). The model equation becomes:

- $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta\square X\square + \varepsilon$

where:

- Y is the dependent variable
- X_1, X_2, ..., Xn are the independent variables
- β_0 is the y-intercept
- β_1, β_2, ..., $\beta\square$ are the slope coefficients for each independent variable
- ε is the error term

Our main goal in using linear regression is to identify the best-fit line, meaning that the error between the predicted and actual values is minimized. The best-fit line has the smallest error. The best-fit line equation represents the relationship between the dependent and independent variables as a straight line. The slope of this line indicates how much the dependent variable changes for each unit change in the independent variable(s).



Cost Function of Linear Regression

Cost function or Loss function The cost function, or loss function, measures the error or difference between the predicted value and the true value. In linear regression, the Mean Squared Error (MSE) cost function is used, which calculates the average of the squared errors between the predicted values and the actual values. The objective is to find optimum slope coefficients and intercepts.

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

We try to find a best-fit line such that the above error is minimal, for this we use:

Gradient Descent: Gradient Descent is an optimization algorithm used to minimize the cost function in linear regression. It iteratively adjusts the parameters (coefficients) of the model to find the best-fit line that predicts the target variable most accurately.

Steps Required in Gradient Descent Algorithm

Step 1: we first initialize the parameters of the model randomly

Step 2: Compute the gradient of the cost function with respect to each parameter. It involves making partial differentiation of cost function with respect to the parameters.

Step 3: Update the parameters of the model by taking steps in the opposite direction of the model. Here we choose a hyperparameter learning rate which is denoted by alpha. It helps in deciding the step size of the gradient.

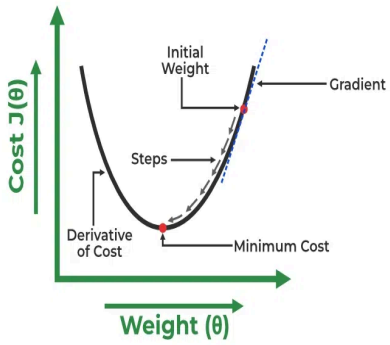
$$\theta j := \theta j - \alpha \frac{\partial}{\partial \theta j} J(\theta)$$

where theta is the parameter.

Step 4: Repeat steps 2 and 3 iteratively to get the best parameter for the defined model.

The learning rate determines how large the steps are in the update process. A small learning rate leads to smaller steps and potentially slower convergence, while a large learning rate can cause the algorithm to jump past the minimum and potentially diverge.

$$\theta_1 = \theta_1 - \alpha\,(J'_{\theta_1})$$
$$= \theta_1 - \alpha \left(\frac{2}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \right)$$
$$\theta_2 = \theta_2 - \alpha\,(J'_{\theta_2})$$
$$= \theta_2 - \alpha \left(\frac{2}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \cdot x_i \right)$$

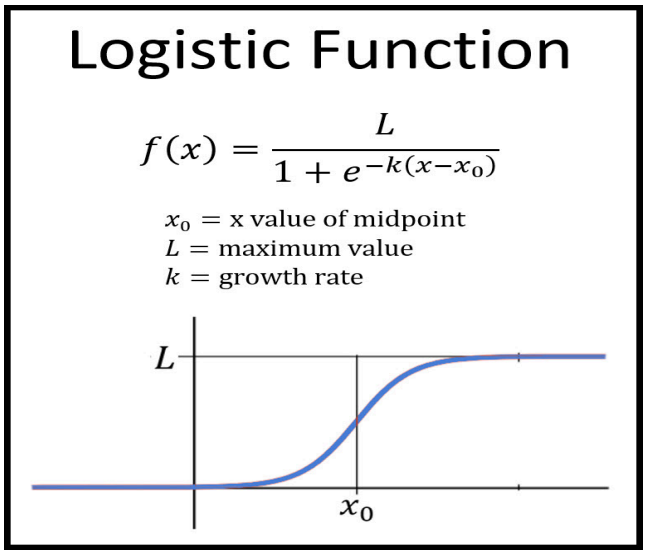


2.8.2 Logistic Regression

Logistic regression is used for binary classification where we use a the sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but

instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. In Logistic regression, instead of fitting a regression line, we fit an “S” shaped logistic function, which predicts two maximum values (0 or 1).

Logistic function: The sigmoid function is a mathematical function used to map the predicted values to probabilities.It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the “S” form.



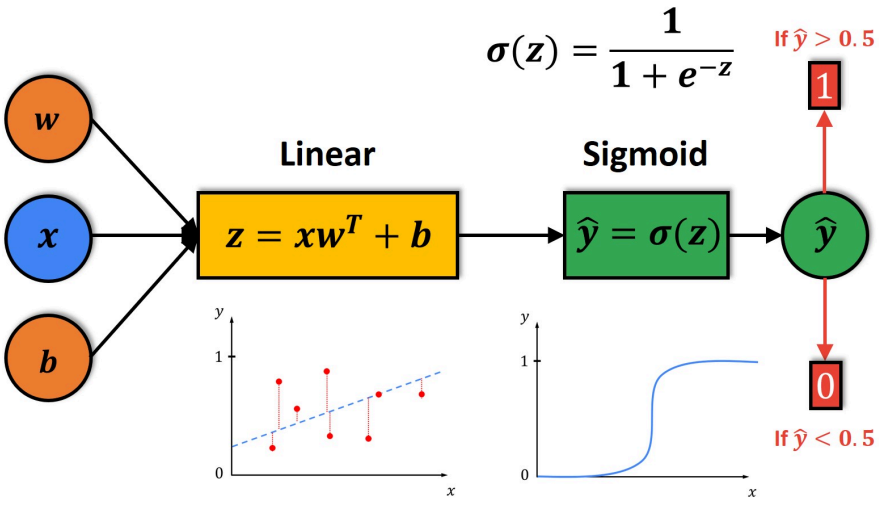
The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between 0 and 1.

$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$

where:

- Y is the dependent variable
- $X_1, X_2, ..., X_n$ are the independent variables
- β_0 is the y-intercept
- $\beta_1, \beta_2, ..., \beta_n$ are the slope coefficients for each independent variable

$y=\sigma(z)=1/1+e^{-z}$



To classify the outcome, a threshold value is used. The most common threshold is 0.5:

- If y>=0.5, predict class 1 (True).
- Else, predict class 0 (False).

Log Likelihood function: The log-likelihood value of a regression model is a way to measure the goodness of fit for a model. The higher the value of the log-likelihood, the better a model fits a dataset.

$$\text{Log likelihood} = \sum[y_i ln\hat{y}_i + (1 - y_i)\ln(1 - \hat{y}_i)]$$

Where y is the actual value and y-cap is the predicted valueTo find the best-fit curve which gives good predictions we use the gradient descent method where we try to maximize likelihood.

Final Assignment

Preprocessing Pipeline Overview

To ensure high-quality model performance and robust predictions, the dataset underwent a comprehensive preprocessing pipeline:

1. **Irrelevant Column Removal:**
   Columns that did not contribute meaningful predictive value — namely `Customer ID`, `Purchase Date`, and `Customer Name` — were dropped. This step helped eliminate noise and prevent overfitting on non-generalizable features.

2. **Categorical Variable Encoding:**
   Categorical features such as `Product Category`, `Payment Method`, and `Gender` were transformed using **one-hot encoding**. This conversion into a numerical format enabled compatibility with machine learning algorithms while preserving the categorical distinctions.

3. **Outlier Detection via Z-Score Filtering:**
   Numeric attributes were scanned for anomalies using **z-score-based outlier detection**. Any data point with a z-score greater than 3 was considered a statistical outlier and removed. This step significantly enhanced model reliability by reducing skewed learning from extreme values.

4. **Feature Scaling:**
   The numeric data was scaled using **StandardScaler**, standardizing the feature distributions to have a mean of 0 and a standard deviation of 1. This was particularly critical for distance-based models and to ensure faster and more stable convergence in optimization algorithms.

Modeling & Evaluation

To predict whether a customer would return a product, several machine learning models were trained and evaluated:

- **Logistic Regression:**
  Served as the baseline model for binary classification. It provided interpretable coefficients and reasonably strong initial performance.

- **Random Forest Classifier:**
  An ensemble of decision trees that improved robustness by reducing variance. It effectively captured complex patterns and feature interactions, offering superior performance over the linear model.

- **Decision Tree Classifier (Vanilla):**
  Offered a simple, interpretable structure to understand decision boundaries. However, it was prone to overfitting in its raw form.

- **Pruned Decision Tree:**
  To combat overfitting, hyperparameters such as `max_depth`, `min_samples_split`, and `ccp_alpha` were tuned. Pruning the tree helped reduce complexity and improved generalization, particularly in the presence of noisy data.

Z-Score Filtering Impact

The application of **z-score anomaly detection** had a significant effect on data quality and model performance. By filtering out statistical outliers, the models learned more stable patterns from the core data distribution, resulting in:

- Reduced training noise
- Improved prediction accuracy
- Better generalization on unseen data

Comparative Analysis & Insights

All models were benchmarked using accuracy, precision, recall, F1-score, and confusion matrices. The Random Forest classifier consistently outperformed others in both balanced accuracy and class-wise F1-scores. The pruned Decision Tree achieved better interpretability while mitigating the overfitting issue seen in the unpruned version. Logistic Regression, though simpler, provided a valuable baseline for comparison.

This rigorous experimentation and comparison allowed for a holistic understanding of which model best suited the problem of return prediction, while highlighting the importance of preprocessing in influencing model efficacy.