A Study of Churn Prediction Analysis for Amazon e-commerce Platform in urban cities of India

NMIMS Global Access School for Continuing Education (NGA-SCE)

Name: Aditi
SAP ID: 77221412207
Course: Masters of Business Administration
(Marketing Management)

Acknowledgement

I extend my sincere appreciation to Prof. Deepak Gupta and Prof. Anuja Shukla for generously permitting and supporting me throughout the execution of this research. Their invaluable suggestions, encouragement, and thoughtful guidance played a pivotal role in the successful completion of this project.

I am profoundly grateful for the unwavering guidance provided by Prof. Anuja Shukla. Her consistent support and supervision were indispensable, and without her, the project would not have reached its current fruition. Prof. Shukla patiently elucidated concepts from the outset, often reiterating them to ensure my comprehension. Their composure and consistent guidance were instrumental, especially during moments of confusion. I express heartfelt thanks for their dedication, even amid their busy schedule, to ensure I stayed on track.

I would like to express my gratitude to the respondents who participated in the online questionnaire, as their time and effort provided the crucial data that formed the basis for analysis and yielded meaningful results.

Finally, I extend my thanks to my family and friends for their understanding and patience during challenging times. Their tolerance of my occasional stress-induced behaviour is deeply appreciated, and their support contributed significantly to navigating through the demands of this project.

Declaration

I Aditi, currently doing my Masters in Business Administration (Marketing Management) from NMIMS NGASCE, hereby declare that the submitted project entitled **"A Study of Churn Prediction Analysis for Amazon e-commerce Platform in urban cities of India "** is my own primary work and is a part of requirements for the fulfilment of my master's degree. It is an original work and has been submitted for the first time. The referred sites and research papers are duly credited in the references section at the end of this project.

Date: 29/11/23

Name: Aditi

SAP ID- 77221412207

Course: MBA (Marketing Management)

NARSEE MONJEE INSTITUTE OF MANAGAMENT STUDIES (NMIMS)

MUMBAI, 400056

Table Of Contents

Introduction

Over the past two decades, the rapid expansion of the retail e-commerce industry has fostered a competitive landscape marked by significant technological advancements and a focus on user-centric approaches (MacKenzie et al., 2013; Morgan, 2018; Terdiman, 2018). Maintaining a valuable user/customer base has become imperative for organizations. The motivation behind churn mitigation primarily stems from the contrast between the costs associated with user acquisition and retention (Gronwald, 2017), with additional advantages identified (Ascarza et al., 2018). The retention strategy is categorized into short-term and long-term objectives (Mezghani et al., 2012; Ascarza et al., 2018). In the short term, companies strive to capture behavioral and transactional dynamics to identify users at imminent risk of churning and retain them if profitable. In the long term, firms focus on enhancing customer satisfaction and loyalty by comprehending the drivers of churn. However, the success of both endeavors relies heavily on accurate representation. Brusilovsky (1996) introduces the concept of a user model to depict user features, including behavior, goals, knowledge, stereotypes, and preferences associated with practical actions. The quality of this model can be indirectly assessed through machine learning methods. Unfortunately, there is no unanimous agreement on a user/customer churn model encompassing both explanatory and explained characteristics.

Statement of problem

Maintaining customer loyalty poses a significant challenge for many organizations, especially those in the e-commerce sector. As highlighted by Wu et al. (2017), retaining current customers proves to be more challenging than acquiring new ones, given the substantial value existing customers bring to e-commerce businesses. While existing customers contribute significantly, attracting new ones requires substantial financial investments to foster loyalty. This research aims to construct a predictive model for the e-commerce sector, examining the essential factors associated with customer churn.

Objectives of the project

The significance of this project lies in its ability to enable e-commerce companies to proactively identify clients at risk of migrating through Churn Prediction and analysis. Anticipating churn in e-commerce provides insights into the actual value of potential losses from these clients, allowing companies to implement necessary retention measures to minimize or prevent their migration. The project aims to achieve the following objectives:

1. Develop prediction models for customer churn in Amazon E-commerce Company by analyzing various attributes related to churn.
2. Segmentation of customers using RFM model for better retention policy
3. Provide actionable recommendations for the business based on the research findings.

Scope of the project

The scope of this project ,Churn analysis in the context of e-commerce can provide valuable insights into customer behaviour, satisfaction, and factors influencing customer retention. Here are several aspects where churn analysis can be beneficial for e-commerce businesses:

Customer Segmentation:

- Identify segments of customers with varying churn rates. Understanding which customer segments are more prone to churn allows for targeted strategies to retain those specific groups.

Customer Lifetime Value (CLV) Analysis:

- Analyze the CLV for different customer segments. This helps in prioritizing efforts and resources towards retaining high-value customers who contribute significantly to revenue over time.

Product and Category Analysis:

- Evaluate the churn rates for different products or categories. Identify whether certain products are associated with higher churn and take actions to enhance the appeal of those products or improve customer experience.

Usability and Experience:

- Explore the relationship between website usability, user experience, and churn. If users find the website difficult to navigate or have a poor experience, they may be more likely to churn.

Purchase Frequency and Recency:

- Analyze how often customers make purchases and when they last made a purchase. Encourage repeat purchases through targeted marketing or promotions to re-engage customers who haven't made a purchase recently.

Customer Feedback and Reviews:

- Analyze customer feedback, reviews, and ratings. Negative sentiments or consistent complaints about certain aspects can be indicators of potential churn factors that need to be addressed.

Promotional Effectiveness:

- Evaluate the effectiveness of promotions and discounts in retaining customers. Analyze whether promotions lead to a short-term boost in sales or contribute to long-term customer loyalty.

Communication Channels:

- Assess the impact of communication channels (email, social media, etc.) on customer retention. Some customers may prefer specific channels, and tailoring communication strategies accordingly can be crucial.

Competitor Analysis:

- Understand how competitive factors influence churn. Analyze whether customers are leaving for competitors and identify areas where your business can differentiate itself.

Predictive Modelling:

- Use machine learning models to predict which customers are at a higher risk of churning. This enables proactive retention strategies before customers decide to leave.

Subscription Management (if applicable):

- If your e-commerce platform offers subscription services, analyze subscription cancellation rates and reasons for cancellations. This can inform adjustments to subscription plans or offerings.

Return Customer Incentives:

- Implement loyalty programs, discounts, or exclusive offers for returning customers. These incentives can encourage repeat business and reduce churn.

Churn analysis is an ongoing process, and its scope can evolve as customer behavior, market dynamics, and business strategies change. By understanding why customers churn, e-commerce businesses can develop targeted and effective retention strategies, ultimately improving customer satisfaction and long-term profitability.

Limitations Of Study

- Demographic Limitation: The study has majority of respondents belonging to the urban settings. Hence the findings from this study can not be extended to International or rural setup.
- Limited Historical Data: Churn prediction models heavily rely on historical data to identify patterns and trends. Our historical data is limited or does not capture diverse scenarios, the model's accuracy may be compromised.
- Imbalanced Data: Churn events are often less frequent compared to non-churn events, leading to imbalanced datasets. Imbalance can affect the model's ability to accurately predict the minority class (churn), as it may be biased toward the majority class. Though the technique of oversampling using SMOTE has been deployed yet certain bias exist
- Changing Customer Behavior: Customer behavior is dynamic and can change for various reasons. External factors, market trends, or changes in customer preferences may not be captured by historical data, affecting the model's ability to adapt to new patterns.
- Lack of Causation: Churn prediction models focus on identifying correlations between features and churn events but may not establish causation. Knowing that a customer is likely to churn does not necessarily explain why they are churning.
- Overfitting: Overfitting occurs when a model learns the training data too well but fails to generalize to new, unseen data. Overfit models may perform well on historical data but struggle with predicting churn in real-world scenarios.The issue of over fitting has been tackled by cross validation of test and train data , yet possibility of overfitting can’t be completely removed
- Limited Predictive Horizon: Churn prediction models are typically designed for short- to medium-term predictions. Predicting churn far into the future may be challenging due to the changing nature of customer behavior and external factors
- Customer Engagement Strategies:Even if a model accurately predicts churn, implementing effective strategies to retain customers is a separate challenge. The success of retention efforts depends on the feasibility and impact of intervention strategies.
- Despite these limitations, churn prediction analysis remains a valuable tool for businesses. It's important to be aware of these limitations and continuously refine models based on evolving data and business conditions. Regular model evaluation and updates are essential for maintaining the predictive accuracy of churn prediction systems.

Executive Summary

Churn Analysis in E-commerce

In the dynamic landscape of retail e-commerce, characterized by technological leaps and user-centric paradigms, organizations face a pivotal challenge of retaining a valuable customer base. The imperative to mitigate churn arises from the contrasting costs of user acquisition and retention. This executive summary delves into the short and long-term strategies employed by companies to address churn, emphasizing the critical role of accurate user representation. While machine learning methods offer insights, a universally agreed-upon churn model that comprehensively captures explanatory and explained characteristics remains elusive.

Objectives Achieved:

- Developed predictive models for customer churn in Amazon E-commerce.
- Utilized RFM segmentation for better customer retention policies.
- Provided actionable recommendations based on research findings.

Methodology

This comprehensive study focuses on analyzing customer churn in the e-commerce sector, specifically within Amazon. The findings are derived from a sample of 161 respondents via random sampling. Outliers were treated and data cleaned before running predictive models. Data imbalance and overfitting was handled to enhance performance of the model. Random forest model with an accuracy of 88% was chosen for predictive modelling post comparing the F1 score, precision and recall .Thus, providing valuable insights into churn probability and factors influencing churn and strategies for retention.

Key discoveries include:

- **Demographic Influences on Churn:**
  - Unemployed and non-disclosing income customers exhibit higher churn rates, suggesting a correlation between employment status and income disclosure with platform engagement.
- **Usage Patterns and Transaction Value Impact:**
  - Churn rates vary based on usage patterns, with infrequent users showing higher churn. Customers spending less than 500 INR per purchase have a notably high churn rate, emphasizing the importance of transaction value in customer loyalty.
- **Subscription and Usability Impact:**
  - Non-subscribers and users providing poor usability ratings have elevated churn rates, indicating the need for targeted retention strategies in these segments.
- **Device Choice Influence:**

- Login device choice influences churn, with customers using tablets showing the highest churn. Addressing device preferences and potential issues could improve user satisfaction.

Recommendations for Retention

- Targeted retention efforts for suburban residents, singles, and Bachelor's degree holders.
- Usability enhancements for customers with neutral or poor usability ratings.
- Incentivizing subscriptions and diversifying payment methods.
- Personalized engagement strategies for rare and occasional users.
- Encouraging income disclosure and optimizing loyalty programs.
- Communication improvement with customers unlikely to recommend.
- This strategic plan outlines customer segmentation based on RFM scores and corresponding retention policies for an e-commerce platform. The segmentation includes:
  - **Top Customers (RFM score > 4.5):**
    - **Policy:** Offer exclusive loyalty programs and VIP treatment.
    - **Action:** Provide personalized offers, early access, and dedicated support. Acknowledge loyalty with special rewards.
  - **High-Value Customers (4.5 > RFM score > 4):**
    - **Policy:** Encourage continued spending and engagement.
    - **Action:** Offer tiered rewards, promotions, and personalized recommendations.
  - **Medium-Value Customers (4 > RFM score > 3):**
    - **Policy:** Incentivize increased frequency and spending.
    - **Action:** Send targeted promotions, implement achievable loyalty milestones, and offer bundle deals.
  - **Low-Value Customers (3 > RFM score > 1.6):**
    - **Policy:** Nurture and convert into higher segments.
    - **Action:** Send re-engagement campaigns, offer incentives, and collect feedback to address issues.
  - **Lost Customers (RFM score < 1.6):**
    - **Policy:** Attempt to re-engage and understand reasons for disengagement.
    - **Action:** Send win-back campaigns with special offers, request feedback, and consider personalized communication.

Conclusion:

The study concludes with actionable recommendations based on nuanced insights into customer behavior and preferences. These recommendations aim to create a more personalized and engaging experience, contributing to customer retention and satisfaction enhancement. Despite certain limitations, this churn prediction model provides a valuable tool for businesses to refine strategies, ensuring continuous adaptation to evolving data and business conditions.

Future Scope

Continued model evaluation and updates are recommended for maintaining predictive accuracy. Further exploration of customer engagement strategies and assessing the feasibility and impact of intervention strategies are crucial for successful retention efforts. Regular refinement of models based on evolving data and business conditions is essential for sustained effectiveness.

Limitations:

The study acknowledges limitations, including demographic bias, limited historical data, imbalanced datasets, changing customer behavior, lack of causation establishment, potential overfitting, limited predictive horizon, and the challenge of effective retention strategy implementation.

This analysis serves as a foundational step for Amazon and similar e-commerce platforms, providing a roadmap for customer retention and satisfaction en

Methodology

This Study is an external primary study that has a sample size of 161. The study was fielded for a period of 1 month. The tool used for collection of data is Google forms. The analysis utilized software such as Python, MS Excel, Google Sheet. The questionnaire comprised of 41 questions out of which 39 were mandatory and 2 were optional. Here, 8 questions were open ended (6 mandatory numeric input based and 2 suggestive optional) and 3 were multi-select questions along with 30 single select multiple choice questions. The approximate time taken by a respondent to 5-6 minutes each.

General steps followed while conducting this research are:

- Data Collection: Gather historical customer data, including demographic information, usage patterns, transaction history, and customer service interactions.
- Data Preprocessing: Clean, transform, and prepare the data for analysis, addressing missing values and outliers.
- Handling data imbalance: The data imbalance is handled using various over sampling techniques like SMOTE etc.
- Model Development: Employ machine learning algorithms (e.g., logistic regression, decision trees, random forests) to build a predictive churn model.
- Model Evaluation: Assess the model's performance through metrics like accuracy, precision, recall, and F1-score.
- Applying the recency frequency monetary model to the data. The RFM analysis helps businesses identify and target specific customer segments with personalized marketing strategies, thereby improving customer engagement, retention, and overall marketing efficiency.
- Retention Strategy Analysis: Analyze the effectiveness of different customer retention strategies, such as targeted promotions, personalized recommendations, or improved customer support.

Defining the variables

S.No	VARIABLE	DESCRIPTION OF THE VARIABLE	Nature Of Variable
1	**age**	This gives the age of the respondents	Independent variable
2	**gender**	This gives the gender of the respondents	Independent variable
3	**Area**	The area where a respondents resides	Independent variable
4	**Marital status**	Marital status of the respondent	Independent variable
5	**Qualification**	The highest qualification of the respondent	Independent variable
6	**Employment status**	Employement status of the respondent	Independent variable
7	**Annual Income**	Annual income of the respondent	Independent variable
8	**Freq_usage**	The frequency of usage of the portal	Independent variable
9	**Recency**	Number of days since the last usage	Independent variable
10	**spending_per_purchase**	The amount spent per purchase (given in INR)	Independent variable
11	**yearly_spend**	The amount spent per year (given in INR)	Independent variable
12	**Subscription**	If the respondent is a subscriber of premium version	Independent variable
13	**years_subs**	The number of years for which there has been an active subscription	Independent variable
14	**other_pref_portal**	Other portals preferred by respondents	Independent variable
15	**login_device_pref**	The preferred mode of login (device) for portal usage	Independent variable
16	**most_purchase_cat**	The most purchased category	Independent variable
17	**mode_of_payment**	preferred mode of payment during a purchase	Independent variable
18	**no_reg_device**	Number of devices registered against an account	Independent variable
19	**time_spent_weekly**	The number of hours spent on the portal in a week's duration	Independent variable
20	**no_of_address**	Number of addresses saved against an account	Independent variable
21	**no_coupon_used**	Number of coupons redeemed in the past year	Independent variable
22	**cashback_1yr**	The amount of cashback received in past one year	Independent variable
23	**last_ordered**	The number of days past since the last order placed via portal	Independent variable
24	**satisf_score**	The measure of satisfaction on a scale of 5(where 1 is least satisfied and 5 is most satisfied)	Independent variable

25	**usage_rating**	The measure of usability of the portal on a scale of 5(where 1 is least satisfied and 5 is most satisfied)	Independent variable
26	**supp_ser_satisf**	The measure of customer support satisfaction of the portal on a scale of 5(where 1 is least satisfied and 5 is most satisfied)	Independent variable
27	**delivery_time_satisf**	The measure of delivery time satisfaction of the portal on a scale of 5(where 1 is least satisfied and 5 is most satisfied)	Independent variable
28	**complain_in1yr**	Number of complaints raised against the portal in past one year	Independent variable
29	**pref_web_or_app**	preferred interface web or application	Independent variable
30	**Place_1stheard**	The place where the respondent first heard about the portal	Independent variable
31	**most_useful_feat**	According to the respondent, the most appealing feature of the portal	Independent variable
32	**reason_subsc**	The most prevalent reason to subscribe to a premium version	Independent variable
33	**likely_reco**	Likelihood to recommend the portal to others	Independent variable
34	**cont_using**	Likelihood to continue usage of the portal	Independent variable
35	**other_attempt**	If the respondent has attempted using other portals	Independent variable
36	**reason_other_attempt**	Reason for attempting usage of other portal	Independent variable
37	**improve_sugg**	Suggestion for improvement of the portal(if any)	Independent variable
38	**aware_discount**	Is the respondent aware of discounts and promotional offers run by the portal	Independent variable
39	**loyalty_prog**	Is the respondent interested in loyalty programs offered by the portal	Independent variable
40	**pur_promo_eff**	The effect promotions have on purchase behaviour of respondents	Independent variable
41	**churn pred**	Will the customer remain in the e-commerce ecosystem or would switch to another	Dependent variable

This study contains 41 variables (excluding key identifiers like name and email ID of the respondent) out of which the Churn pred (variable for predicting churn) is the dependent variable that we are interested in studying and rest 40 other variables are independent variable.

Literature Review

Marketing overview

The 7Ps is a marketing framework that outlines the key elements of a marketing strategy. Each "P" represents a different aspect of the marketing mix. Here's a breakdown of the 7Ps and their relevance to marketing:

1. Product: This refers to the actual goods or services a business offers to its customers. It involves product design, features, quality, branding, and the overall value proposition.
2. Price: Pricing strategy involves determining the right price for the product or service. Factors such as cost, perceived value, competitor pricing, and market demand influence pricing decisions.
3. Place: Place, or distribution, focuses on how products are made available to customers. It involves decisions about channels of distribution, logistics, inventory management, and ensuring that the product is accessible to the target market.
4. Promotion: Promotion involves the activities a business undertakes to communicate and promote its products to the target audience. This includes advertising, public relations, sales promotions, personal selling, and other promotional strategies.
5. People: People refer to both the employees involved in delivering the product or service and the customer themselves. Employee training, customer service, and the overall customer experience are crucial elements under the "People" category.
6. Process: Process refers to the procedures, systems, and methods used to deliver the product or service. It encompasses the entire customer journey, from initial awareness to purchase and post-purchase support.
7. Physical Evidence: Physical evidence relates to the tangible elements that customers can see and touch, reinforcing the overall brand image. It includes physical spaces, packaging, branding materials, and any other tangible cues that communicate the brand identity.

Integration of 7Ps with Marketing:

- Comprehensive Marketing Strategy: The 7Ps framework provides a comprehensive approach to developing a marketing strategy. It ensures that all aspects of the marketing mix are considered and integrated into a cohesive plan.
- Customer-Centric Approach:By including elements like people, process, and physical evidence, the 7Ps framework emphasizes the importance of a customer-centric approach. It recognizes that the customer experience goes beyond just the product itself.
- Holistic View of Marketing Mix: The 7Ps framework encourages marketers to consider all elements simultaneously. Changes in one P may impact others, and a holistic view helps ensure consistency and effectiveness in the overall marketing mix.
- Adaptability: The framework is adaptable to various industries and business types. Whether a business is focused on tangible products, services, or a combination, the 7Ps can be tailored to suit the specific needs of the organization.

- Strategic Decision-Making: Marketers can use the 7Ps to make strategic decisions about product development, pricing, promotion channels, and other critical aspects. It provides a structured approach for analyzing and optimizing marketing efforts.

Churn

Customer churn, denoting the departure of original customers from an enterprise to opt for competitors' services, is a prevalent concern in various sectors, including telecommunications (Wu et al., 2017). E-commerce customer churn specifically involves customers leaving an enterprise due to reasons like product quality concerns or delayed deliveries. This type of churn occurs in a non-contractual relationship setting, making it challenging for businesses to detect in advance (Shao, 2016).

In the context of e-commerce, accurately predicting high-value customer groups on the brink of churning is crucial. Simultaneously, understanding the purchasing habits of customers who remain loyal is essential for effective customer retention. The objective of e-commerce customer churn prediction is to amalgamate customer data over time, analyze purchase behaviors, and establish predictive models (Zhang, 2015). This enables the provision of targeted retention measures to minimize churn, identify high-value non-churn customers, and excel in customer retention efforts.

The significance of e-commerce customer churn prediction lies in mitigating the associated costs, as retaining existing customers proves more cost-effective than acquiring new ones (Shao, 2016). Distinctly, customer purchase behavior varies between existing and new customers, emphasizing the need to identify the factors contributing to customer loss. Analyzing customer loss, predicting potential churn, and implementing corresponding retention strategies are crucial steps in the e-commerce sector (Lu et al., 2018). Many e-commerce companies leverage data mining technologies for customer relationship management, employing methods like customer segmentation, churn prediction, and fraud analysis (Huang, 2018).

Customer Segmentation

Customer segmentation serves as a crucial foundation for enhancing targeted marketing efforts, aiming to recognize and leverage the value of customer relationships (Feng et al., 2018). Guided by the Pareto principle, which highlights that a significant portion of a company's profits is derived from a small percentage of its customers, understanding and tapping into customer value is imperative. The principle suggests that 80% of profits are generated by the top 20% of customers, while the bottom 30% of non-profit customers account for 50% of profit loss (Sun et al., 2020). Effective customer segmentation enables companies to focus on actual customer value, allocate resources strategically, and enhance core competitiveness.

In the dynamic landscape of e-commerce, the shift from a product-centric to a customer-centric model has become prominent, thanks to real-time and interactive online business activities (Dhote et al., 2020). According to Agrawal et al. (2018), customers are not only a source of profits for enterprises but also a vital resource for gaining a competitive edge. Given the relatively high churn rate in e-commerce, establishing long-term alliances and stable relationships with customers is essential. The extensive and complex nature of the e-commerce customer base requires a nuanced approach to identifying high-value customers and predicting and preventing churn (Saghir et al., 2019). Wu et al. (2021) emphasize that evaluating customer value enables companies to identify valuable customers and tailor

customer management strategies based on their value, maximizing the impact of limited resources.
In this context, the study by Saghir et al. (2019) raises the question of accurately identifying high-value customers and implementing proactive retention measures. Wu et al. (2021) advocate for using historical transaction data, particularly the amount of purchase, as a primary measure of customer value. The purchase amount is directly linked to the sales volume of enterprise products or services, serving as a tangible indicator of customer value and forming the basis for effective customer segmentation.

Univariate , Bivariate and Multivariate analysis

Univariate Analysis- Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable. There are many different ways people use univariate analysis. The most common univariate analysis is checking the central tendency (mean, median and mode), the range, the maximum and minimum values, and standard deviation of a variable.
Bivariate Analysis- Bivariate analysis is where you are comparing two variables to study their relationships. These variables could be dependent or independent to each other. In Bivariate analysis is that there is always a Y-value for each X-value.
Multivariate Analysis - Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables. For three variables, you can create a 3-D model to study the relationship (also known as Trivariate Analysis). However, since we cannot visualize anything above the third dimension, we often rely on other softwares and techniques for us to be able to grasp the relationship in the data.

Univariate	Bivariate	Multivariate
It only summarize single variable at a time.	It only summarize two variables	It only summarize more than 2 variables.
It does not deal with causes and relationships.	It does deal with causes and relationships and analysis is done.	It does not deal with causes and relationships and analysis is done.
It does not contain any dependent variable.	It does contain only one dependent variable.	It is similar to bivariate but it contains more than 2 variables.
The main purpose is to describe.	The main purpose is to explain.	The main purpose is to study the relationship among them.

Correlation

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

We describe correlations with a unit-free measure called the correlation coefficient which ranges from -1 to +1 and is denoted by r. Statistical significance is indicated with a p-value. Therefore, correlations are typically written with two key numbers: r = and p = .

- The closer r is to zero, the weaker the linear relationship.
- Positive r values indicate a positive correlation, where the values of both variables tend to increase together.
- Negative r values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease.

- The p-value gives us evidence that we can meaningfully conclude that the population correlation coefficient is likely different from zero, based on what we observe from the sample.
- "Unit-free measure" means that correlations exist on their own scale: in our example, the number given for r is not on the same scale as either elevation or temperature. This is different from other summary statistics. For instance, the mean of the elevation measurements is on the same scale as its variable.

Limitation of Correlation

Correlation can't look at the presence or effect of other variables outside of the two being explored. Importantly, correlation doesn't tell us about cause and effect. Correlation also cannot accurately describe curvilinear relationships.

Hypothesis testing

Hypothesis testing is a statistical method that is used in making a statistical decision using experimental data. Hypothesis testing is basically an assumption that we make about a population parameter. It evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

Parameters of hypothesis testing

- Null hypothesis(H0): In statistics, the null hypothesis is a general given statement or default position that there is no relationship between two measured cases or no relationship among groups. In other words, it is a basic assumption or made based on the problem knowledge.

- Alternative hypothesis(H1): The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis.

- Level of significance It refers to the degree of significance in which we accept or reject the null hypothesis. 100% accuracy is not possible for accepting a hypothesis, so we, therefore, select a level of significance that is usually 5%. This is normally denoted with \alpha and generally, it is 0.05 or 5%, which means your output should be 95% confident to give a similar kind of result in each sample.

- P-value The P value, or calculated probability, is the probability of finding the observed/extreme results when the null hypothesis(H0) of a study-given problem is true. If your P-value is less than the chosen significance level then you reject the null hypothesis i.e. accept that your sample claims to support the alternative hypothesis.

Steps in Hypothesis Testing

Step 1– We first identify the problem about which we want to make an assumption keeping in mind that our assumption should be contradictory to one another
Step 2 – We consider statical assumption such that the data is normal or not, statical independence between the data.
Step 3 – We decide our test data on which we will check our hypothesis
Step 4 – The data for the tests are evaluated in this step we look for various scores in this step like z-score and mean values.
Step 5 – In this stage, we decide where we should accept the null hypothesis or reject the null hypothesis

Predictive modelling

One popular statistical method for forecasting behavior is predictive modeling. As a type of data-mining technology, predictive modeling solutions analyze both historical and present data to create a model that can be used to forecast future events. Predictive modeling involves gathering data, creating a statistical model, making predictions, and validating (or updating) the model when new data becomes available. To increase underwriting accuracy, risk models can be developed that intricately link member data with lifestyle and demographic data from other sources. Predictive models evaluate historical data to determine the likelihood that a client would engage in a particular behavior going forward. This group also includes algorithms that look for minute patterns in data to address inquiries concerning customer performance, such as churn prediction models. Predictive models often perform calculations during live transactions—for example, to evaluate the risk or opportunity of a given customer or transaction to guide a decision.

Types of Machine Learning

To better understand Random Forest algorithm and how it works, it's helpful to review the three main types of machine learning

Reinforced Learning

The process of teaching a machine to make specific decisions using trial and error.

Unsupervised Learning

Users have to look at the data and then divide it based on its own algorithms without having any training. There is no target or outcome variable to predict nor estimate.

Supervised Learning

Users have a lot of data and can train your models. Supervised learning further falls into two groups: classification and regression.

With supervised training, the training data contains the input and target values.

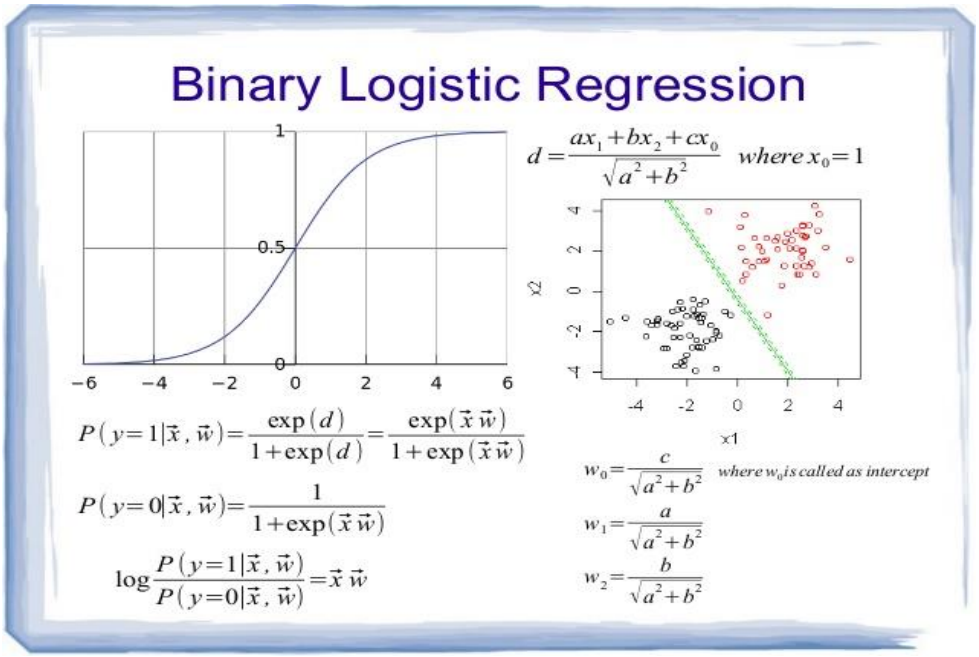
The algorithm picks up a pattern that maps the input values to the output and uses this pattern to predict values in the future. Unsupervised learning, on the other hand, uses training data that does not contain the output values. The algorithm figures out the desired output over multiple iterations of training. Finally, we have reinforcement learning. Here, the algorithm is rewarded for every right decision made, and using this as feedback, and the algorithm can build stronger strategies.

Logistic regression

Logistic regression models a relationship between predictor variables and a categorical response variable. For example, we could use logistic regression to model the relationship between various measurements of a manufactured specimen (such as dimensions and chemical composition) to predict if a crack greater than 10 mils will occur (a binary variable: either yes or no). Logistic regression helps us estimate a probability of falling into a certain level of the categorical response given a set of predictors. We can choose from three types of logistic regression, depending on the nature of the categorical response variable:

Binary Logistic Regression:

Used when the response is binary (i.e., it has two possible outcomes). The cracking example given above would utilize binary logistic regression. Other examples of binary responses could include passing or failing a test, responding yes or no on a survey, and having high or low blood pressure.



Nominal Logistic Regression:
Used when there are three or more categories with no natural ordering to the levels. Examples of nominal responses could include departments at a business (e.g., marketing, sales, HR), type of search engine used (e.g., Google, Yahoo!, MSN), and color (black, red, blue, orange).

Ordinal Logistic Regression:
Used when there are three or more categories with a natural ordering to the levels, but the ranking of the levels do not necessarily mean the intervals between them are equal. Examples of ordinal responses could be how students rate the effectiveness of a college course (e.g., good, medium, poor), levels of flavors for hot wings, and medical condition (e.g., good, stable, serious, critical).

Particular issues with modelling a categorical response variable include nonnormal error terms, nonconstant error variance, and constraints on the response function (i.e., the response is bounded between 0 and 1). We will investigate ways of dealing with these in the binary logistic regression setting here. Nominal and ordinal logistic regression are not considered in this course.

Logistic Regression in Churn

- **Churn prediction**: Specific behaviors may be indicative of churn in different functions of an organization. For example, human resources and management teams may want to know if there are high performers within the company who are at risk of leaving the organization; this type of insight can prompt conversations to understand problem areas within the company, such as culture or compensation. Alternatively, the sales organization may want to learn which of their clients are at risk of taking their business elsewhere. This can prompt teams to set up a retention strategy to avoid lost revenue.

Random Forest

A Random Forest Algorithm is a supervised machine learning algorithm that is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its

accuracy and problem-solving ability. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

Random Forest is a powerful predictive modeling algorithm that is commonly used in churn prediction in various industries, including telecommunications, finance, and e-commerce. The concept behind Random Forest lies in the ensemble learning technique, where multiple decision trees are trained independently, and their predictions are combined to achieve a more robust and accurate result.

1. Ensemble Learning:

- Random Forest employs an ensemble of decision trees, where each tree is constructed independently based on a subset of the data and a random selection of features. The final prediction is made by aggregating the predictions of all the trees.

2. Decision Trees:

- Each decision tree in the Random Forest makes predictions by recursively splitting the data based on features, creating a tree-like structure. Decision trees are prone to overfitting, but the ensemble nature of Random Forest mitigates this by combining multiple trees.

3. Bootstrap Aggregating (Bagging):

- Random Forest employs bagging by training each tree on a bootstrapped sample of the data. This involves randomly selecting samples with replacement, ensuring diversity among the trees.

4. Feature Randomization:

- At each node of a decision tree, only a random subset of features is considered for splitting. This randomization reduces the correlation between individual trees and contributes to the model's robustness.

5. Voting Mechanism:

- In classification problems like churn prediction, each tree "votes" for a class, and the class with the majority of votes becomes the final prediction. In regression problems, the average prediction of all trees is considered.

6. Handling Imbalanced Data:

- Churn prediction datasets often suffer from class imbalance, where the number of churn instances is significantly lower than non-churn. Random Forest handles this imbalance well by giving more weight to the minority class during training.

7. Hyperparameter Tuning:

- Random Forest has hyperparameters that can be tuned for optimal performance, such as the number of trees in the forest, the depth of each tree, and the minimum number of samples required to split a node.

8. Interpretability:

- While Random Forest models are often considered as "black boxes," efforts can be made to interpret feature importance. Features that consistently contribute to accurate predictions across trees are deemed more important.

9. Cross-Validation:

- Cross-validation techniques, such as k-fold cross-validation, can be employed to evaluate the model's performance, ensuring it generalizes well to unseen data.

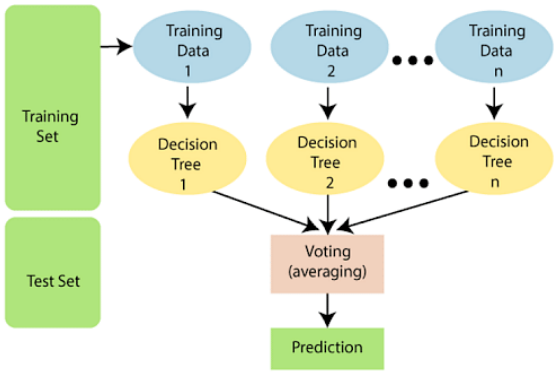
10. Scalability:

- Random Forest is computationally efficient and can be parallelized, making it suitable for large datasets.

In churn prediction, Random Forest's ability to handle complex relationships in data, manage imbalanced datasets, and provide robust predictions makes it a popular choice for building effective and reliable models.

Your AI/ML Career is Just Around The Corner!

Working of Random Forest Algorithm



The following steps explain the working Random Forest Algorithm:

Step 1: Select random samples from a given data or training set.

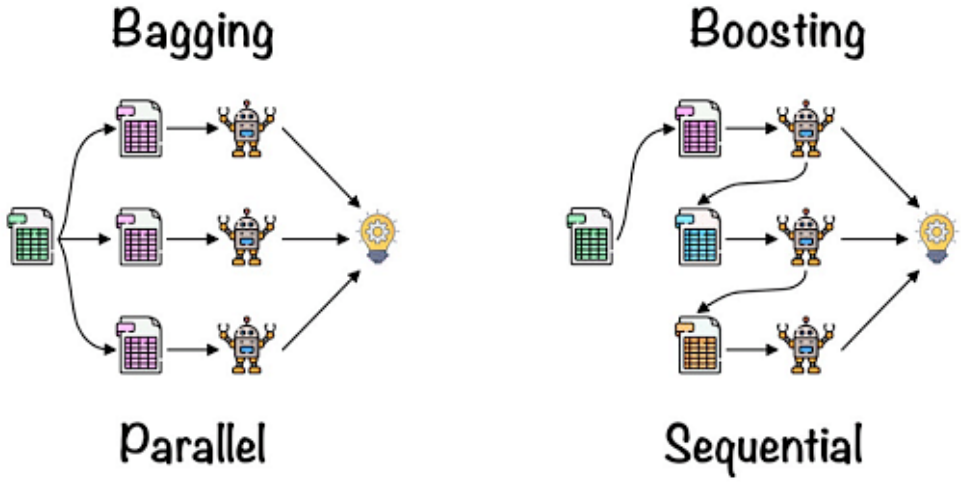
Step 2: This algorithm will construct a decision tree for every training data.

Step 3: Voting will take place by averaging the decision tree.

Step 4: Finally, select the most voted prediction result as the final prediction result.

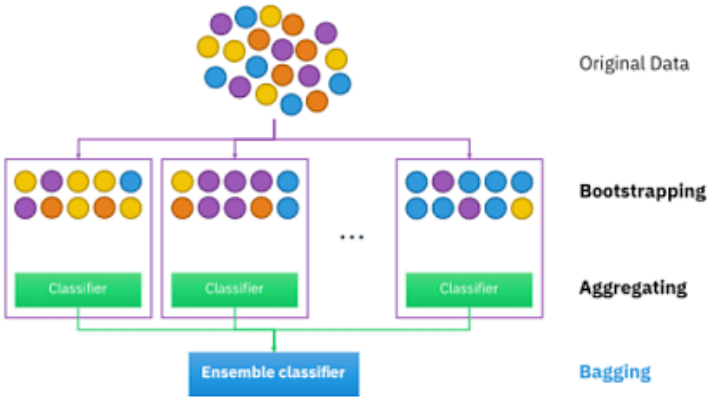
This combination of multiple models is called Ensemble. Ensemble uses two methods:

1. Bagging: Creating a different training subset from sample training data with replacement is called Bagging. The final output is based on majority voting.
2. Boosting: Combing weak learners into strong learners by creating sequential models such that the final model has the highest accuracy is called Boosting. Example: ADA BOOST, XG BOOST.



Bagging: From the principle mentioned above, we can understand Random forest uses the Bagging code. Now, let us understand this concept in detail. Bagging is also known as Bootstrap Aggregation used by random forest. The process begins with any original random data. After arranging, it is organised into samples known as Bootstrap Sample. This process is known as Bootstrapping.Further, the models are trained individually, yielding different results known as Aggregation. In the last step, all the results are combined, and the generated

output is based on majority voting. This step is known as Bagging and is done using an Ensemble Classifier.



Handling Imbalanced Data

The problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary.

One way to solve this problem is to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model.

An improvement in duplicating examples from the minority class is to synthesize new examples from the minority class. This is a type of data augmentation for tabular data and can be very effective.

Perhaps the most widely used approach to synthesizing new examples is called the **Synthetic Minority Oversampling Technique**, or SMOTE for short.

Over fitting and K -fold cross validation

Overfitting: Overfitting is a common issue in machine learning where a model learns the training data too well, capturing noise or random fluctuations in the data rather than the underlying patterns. As a result, an overfit model performs well on the training set but poorly on new, unseen data. Overfitting is a concern because the model fails to generalize to the broader population.

Causes of Overfitting:

- Complex Models: Models with high complexity, such as those with many parameters or deep neural networks, are more prone to overfitting.

- Insufficient Data: When the amount of training data is limited, the model may memorize the training set rather than learning the underlying patterns.
- Irrelevant Features: Including irrelevant or noisy features in the model can lead to overfitting. The model may capture random variations in these features.
- Lack of Regularization: Regularization techniques, such as L1 or L2 regularization, penalize overly complex models and help prevent overfitting.
- Data Leakage: If information from the test set accidentally leaks into the training set, the model may learn patterns specific to the test set, leading to overfitting.

K-Fold Cross Validation:

Cross-validation is a technique used to assess how well a model will generalize to an independent dataset. K-Fold Cross Validation is a specific method where the dataset is divided into k subsets (folds), and the model is trained and evaluated k times, each time using a different fold as the test set and the remaining folds as the training set. The performance metrics are averaged across the k iterations.

Steps of K-Fold Cross Validation:

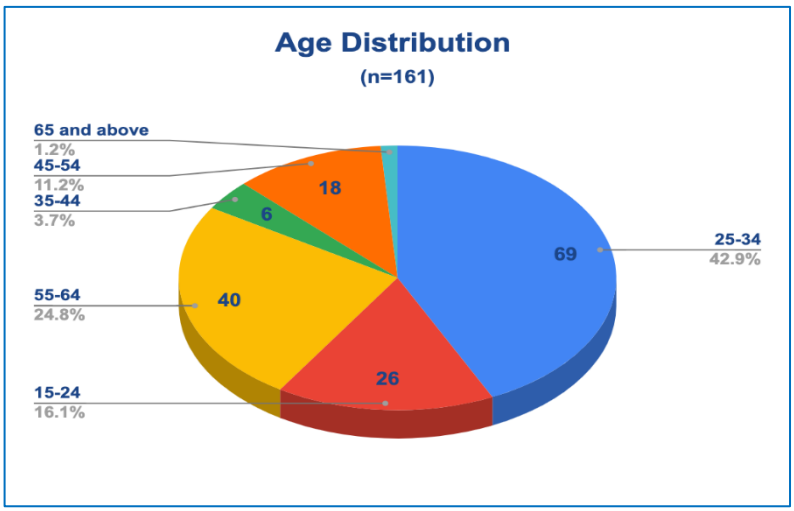
1. Data Splitting: The dataset is divided into k subsets or folds.
2. Model Training and Evaluation: The model is trained and evaluated k times. In each iteration, a different fold is used as the test set, and the remaining folds form the training set.
3. Performance Metrics: Performance metrics (e.g., accuracy, precision, recall) are calculated for each iteration.
4. Average Performance: The performance metrics from each iteration are averaged to obtain a more reliable estimate of the model's performance.

Benefits of K-Fold Cross Validation:

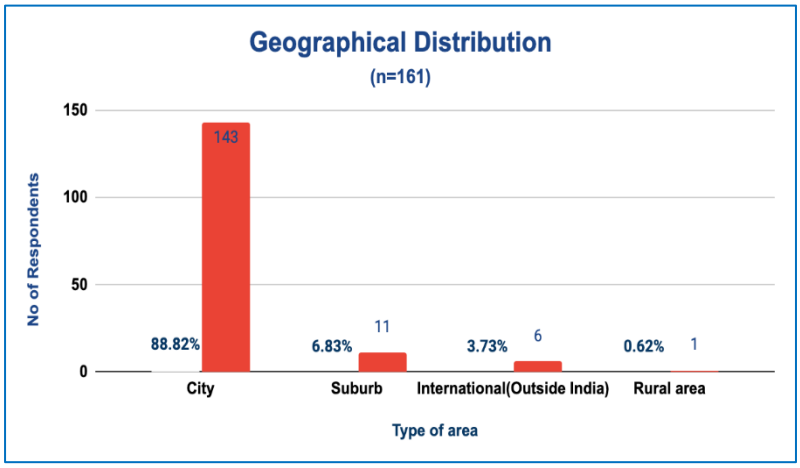
- Reduced Variance: By using multiple train-test splits, K-Fold Cross Validation provides a more stable estimate of model performance and reduces the impact of randomness in a single split.
- Better Generalization: The model's ability to generalize to new, unseen data is more accurately assessed with K-Fold Cross Validation.
- Data Utilization: All data points are used for both training and testing, ensuring that the model is evaluated on the entire dataset.
- Model Selection: K-Fold Cross Validation is often used for model selection and hyperparameter tuning by comparing the performance of different models.

Data Collection & Analysis

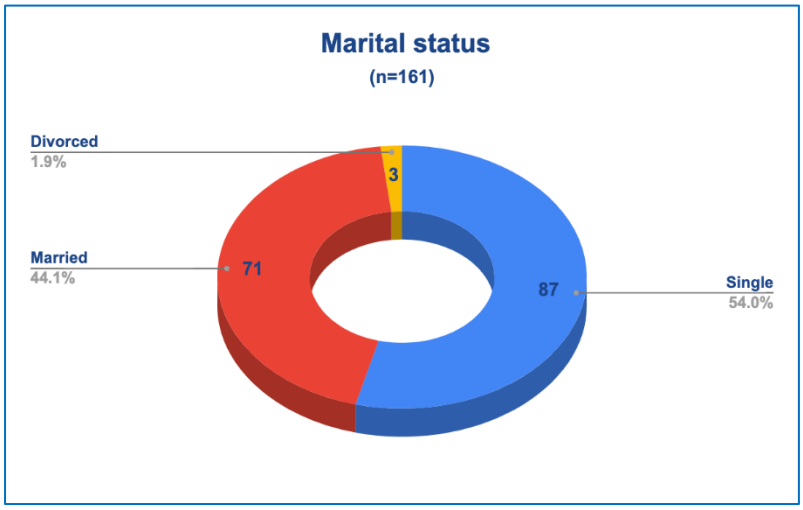
DEMOGRAPHICS ANALYSIS



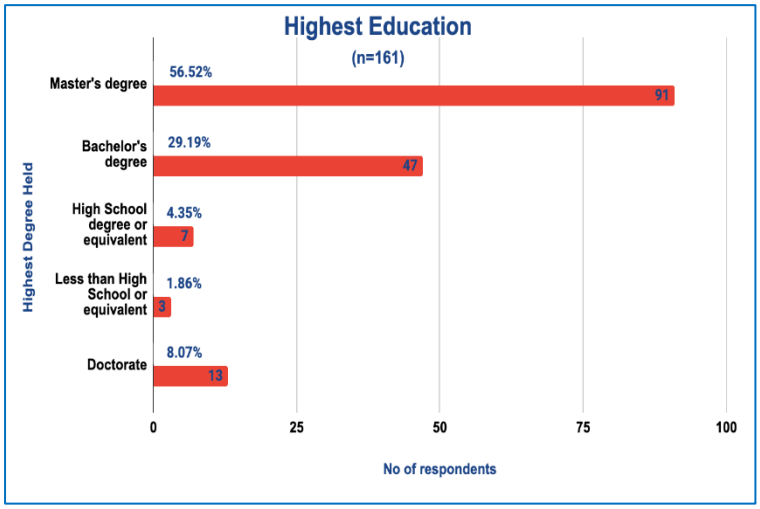
Majority of the respondents belong to the age bracket of 25 - 34 years (42.9%) and 55 - 64 years (24.8%), collectively comprising 68% of the sample size.



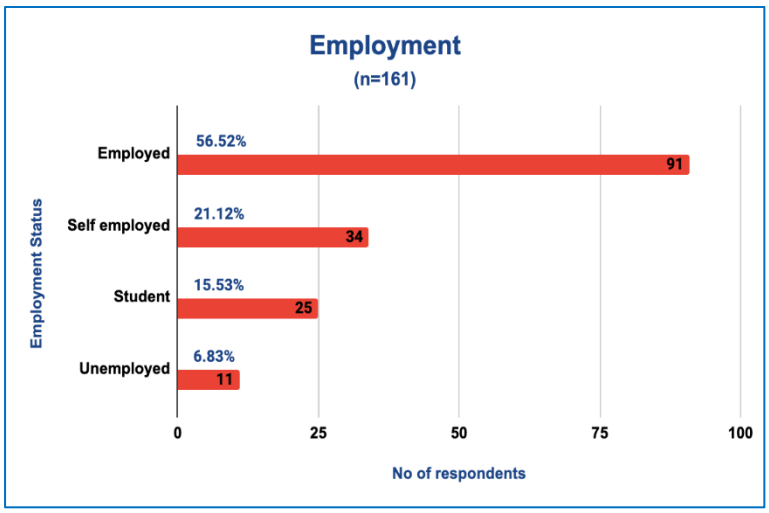
Maximum number of respondents (88.8%) are residing in the city.



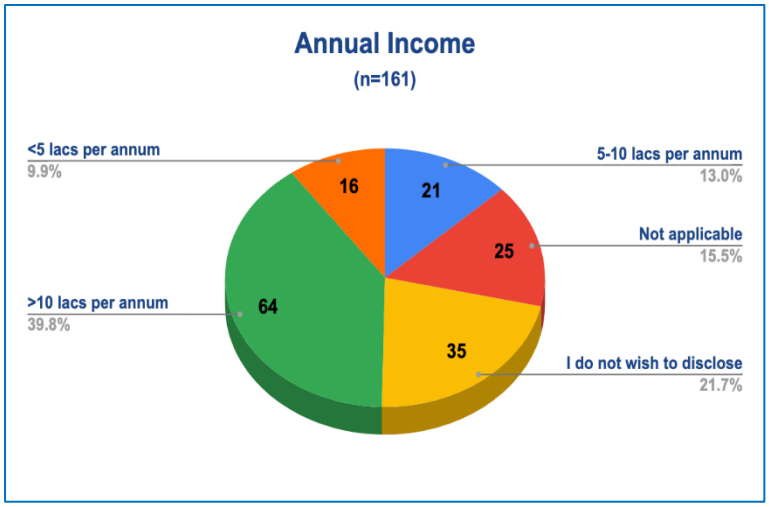
Single respondents take the lead (54%) followed by married respondents (44.1%) in the sample.



Majority of the respondents are either master's degree holder (56.5%) or bachelor's degree holder (29.1%).

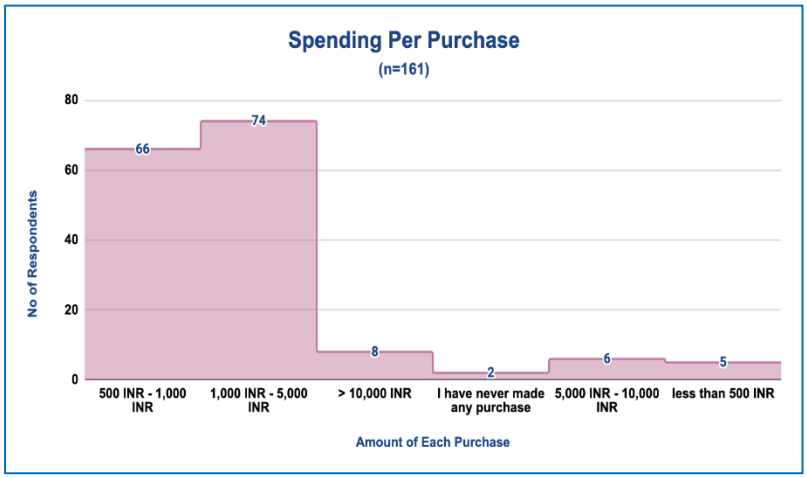


78.6% of the respondents in the sample reported earning an income

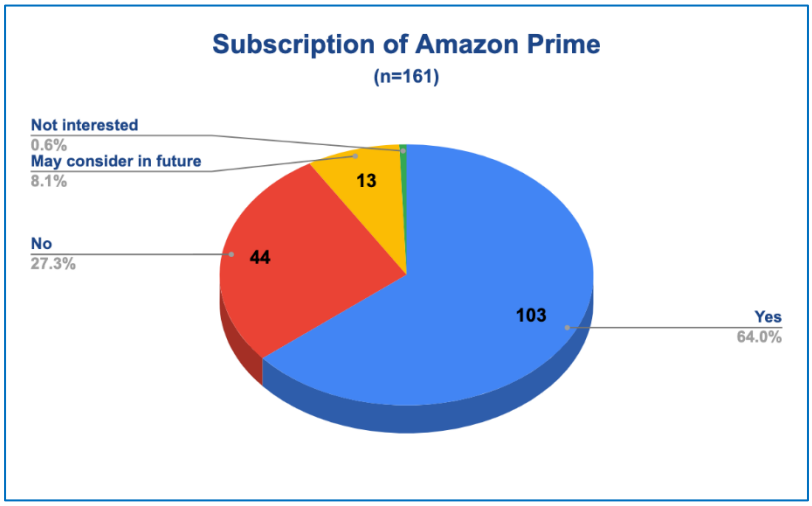


62.7% of the respondents have disclosed the income bracket to which they belong. Amongst them majority respondents fall in the >10 LPA (in INR) bracket (39.8%)

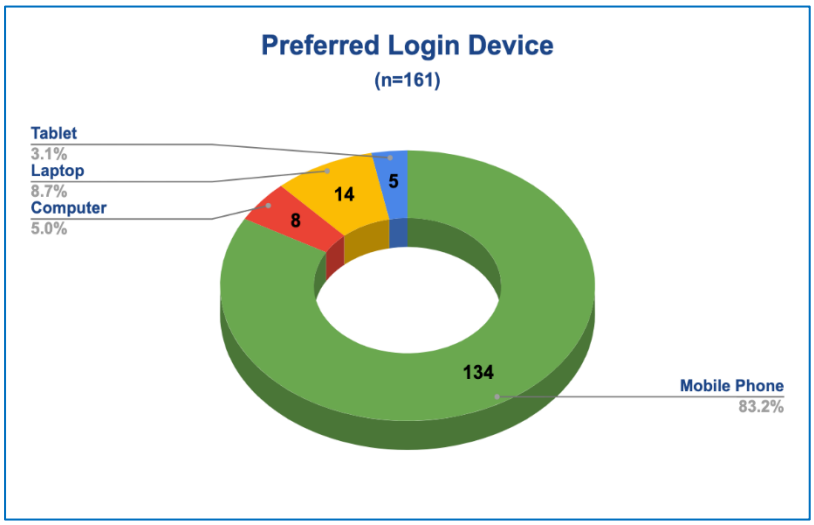
UNIVARIATE ANALYSIS



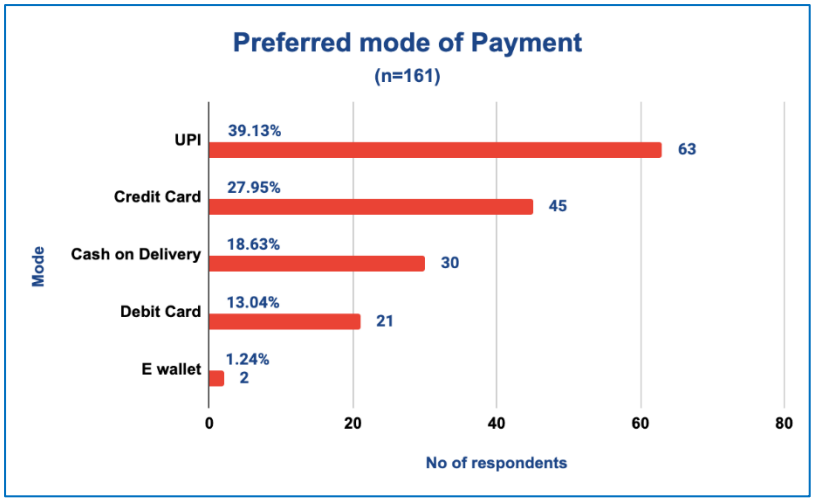
Most of the respondents (i.e., 86.9%) spend less than 5000 INR per purchase. Amongst them those who spend between 1000 INR- 5000 INR (46%) per purchase are 5% more in number than those who spend between 500 INR –1000 INR (41%)



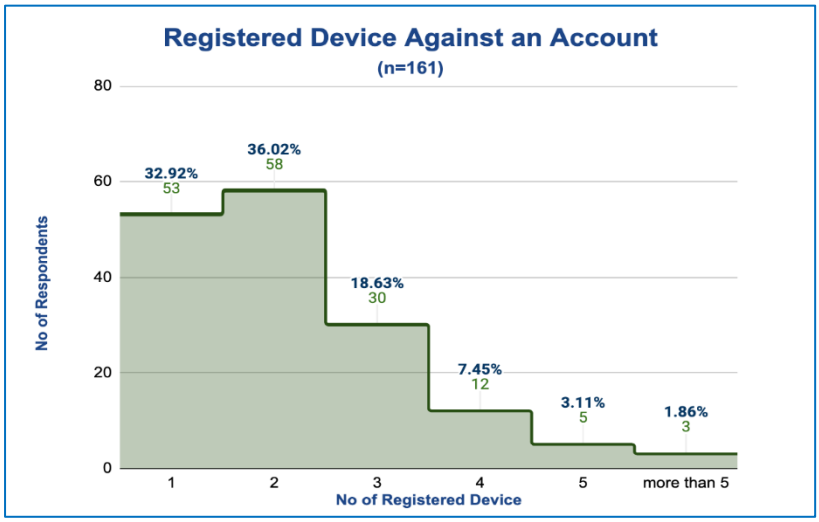
A major proportion of respondents (64%) have subscribed to the premium version of Amazon portal.



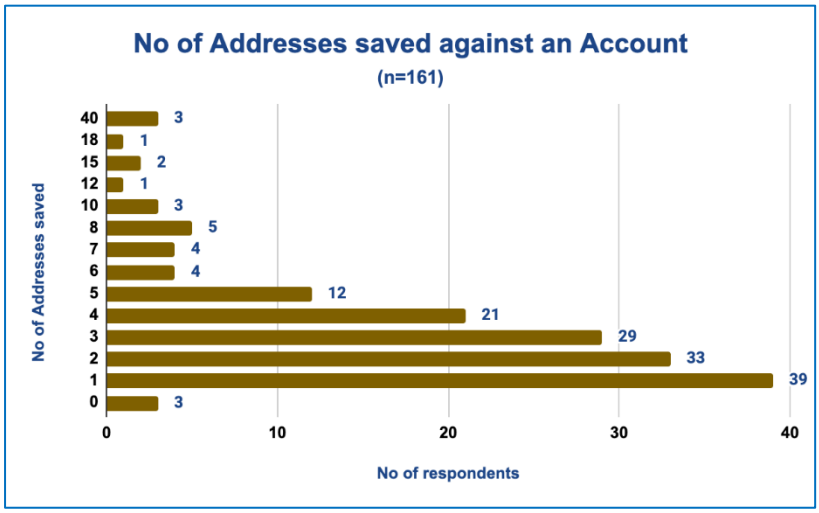
Majority of the respondents (i.e., 83%) prefer to login via their mobile phones.



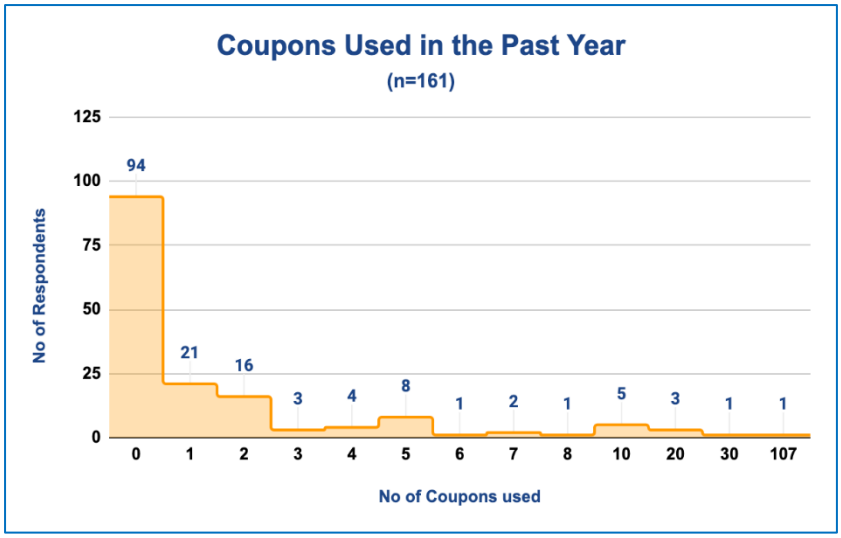
Most preferred mode of payment while purchasing today is UPI (39.1%) followed by credit card (27.9%)



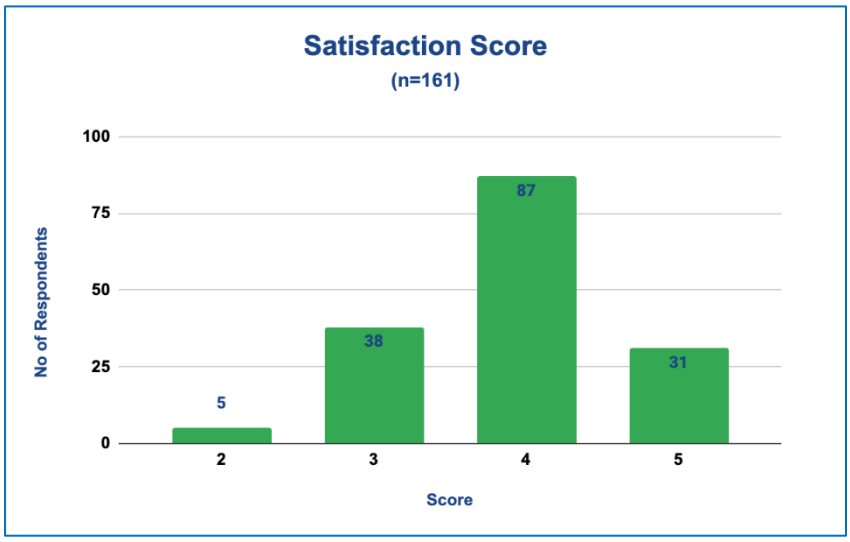
Up to 3 devices registered against an account constitute the maximum proportion (87.5%) of sample preference. Within this the majority (36%) have two devices registered against an account.



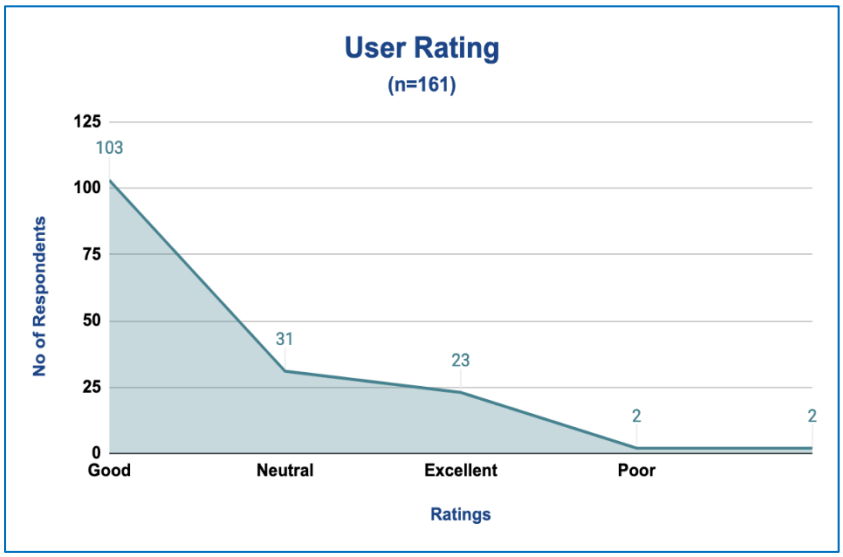
Most number of users (39) have one address saved against a particular account, Until 4 addresses post which the number of users having larger number of addresses decreases drastically.



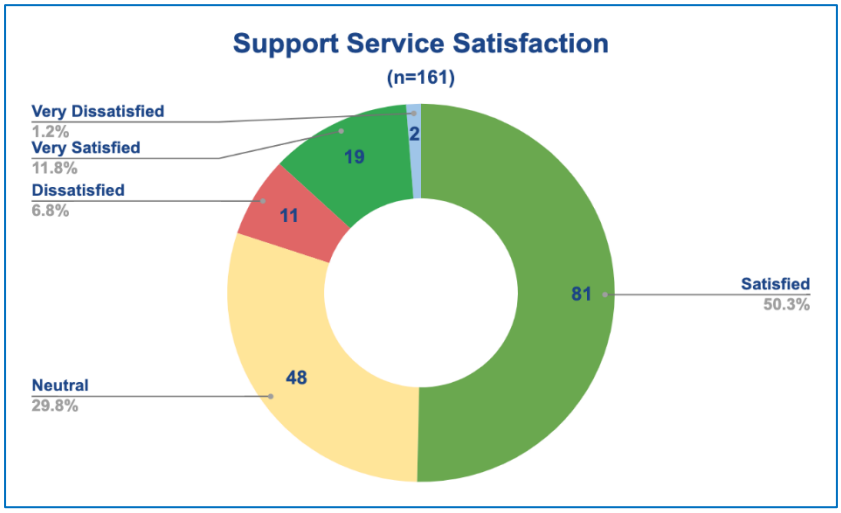
Most respondents (58.3%) have not used any coupons in the past year. Amongst the people who have redeemed coupons, most of them have used either 1, 2 or 5 coupons in the past year.



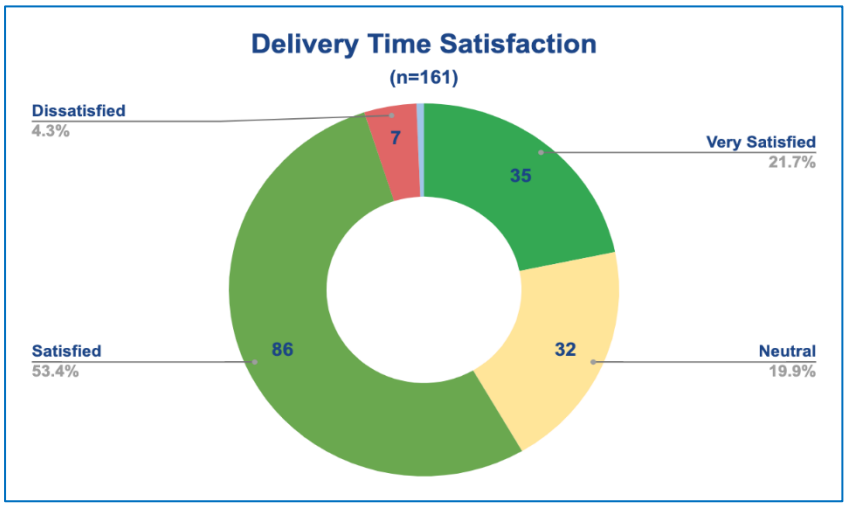
73% of the users are either satisfied or extremely satisfied with the portal (rating it a 4 or 5). Amongst these and on an overall basis, most users have a satisfaction score of 4, i.e., they are satisfied with the portal. It is also seen that none of the users are extremely dissatisfied (giving a score of 1).



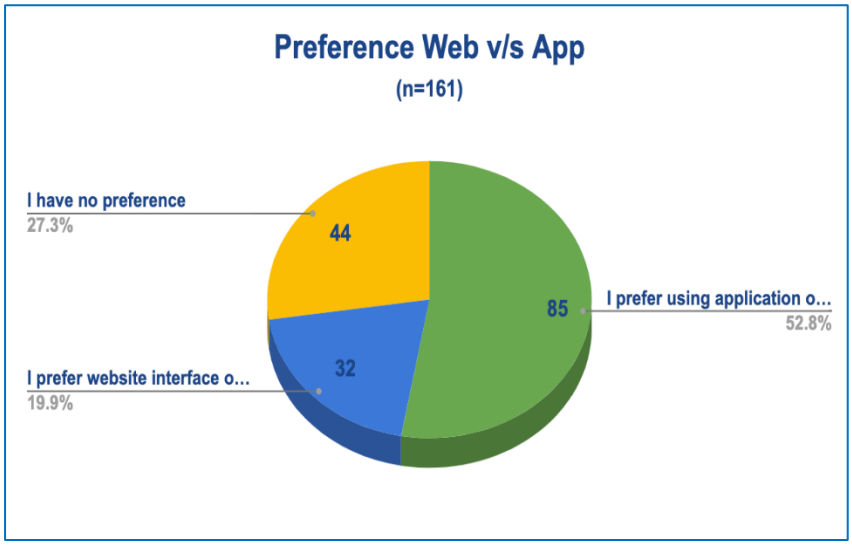
Most users (64%) have rated the usability as good followed by 19.3% users rating it as neutral



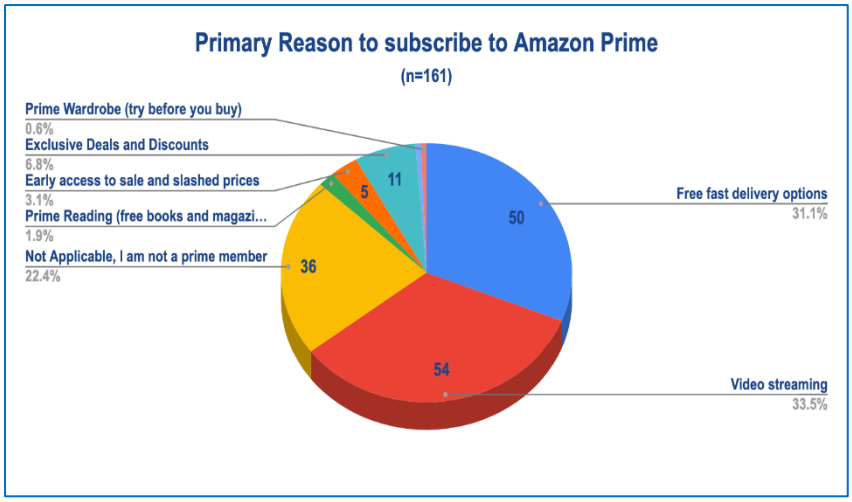
Over 50% of respondents are satisfied with the support service of the portal. Whereas only 8% users are either dissatisfied (6.8%) or extremely dissatisfied (1.2%) with the support service of the portal.



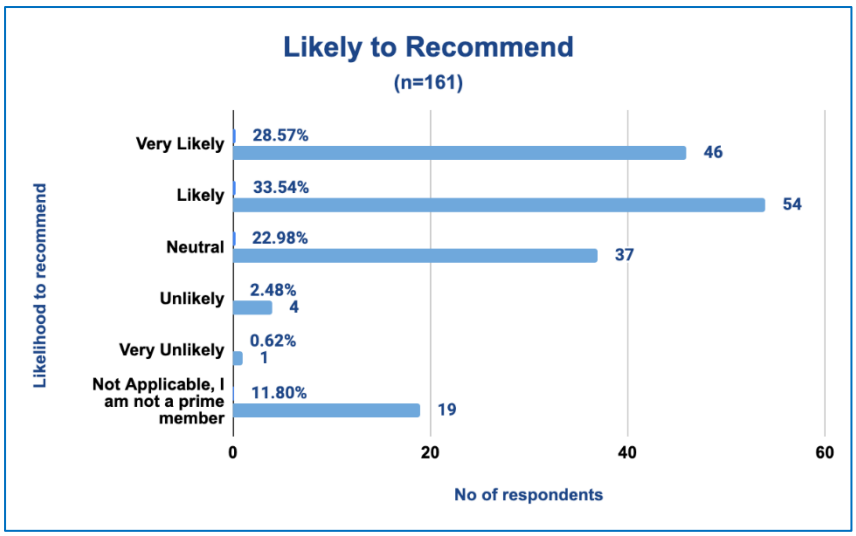
A little over 75% of respondents are either satisfied (53.4%) or very satisfied (21.7%) with the delivery time of the portal.



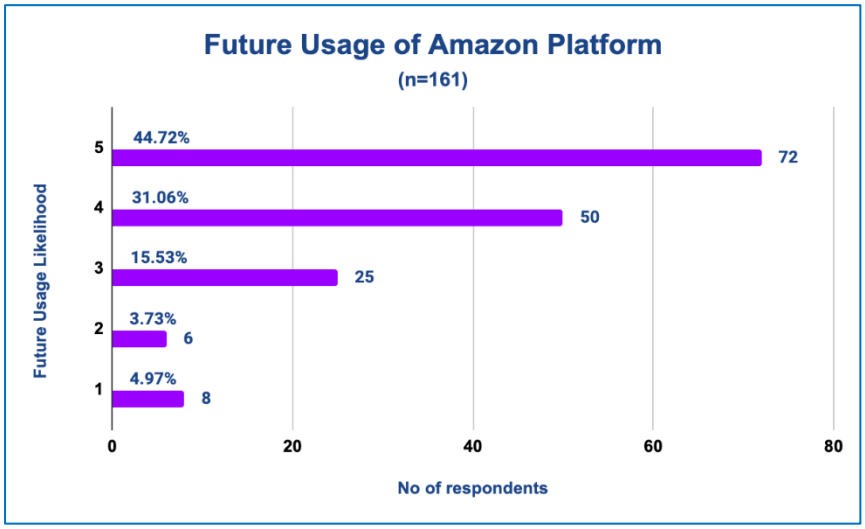
Nearly half (52.8%) of the respondents prefer to use the application over the website. Followed by no preference for either (27.3%)



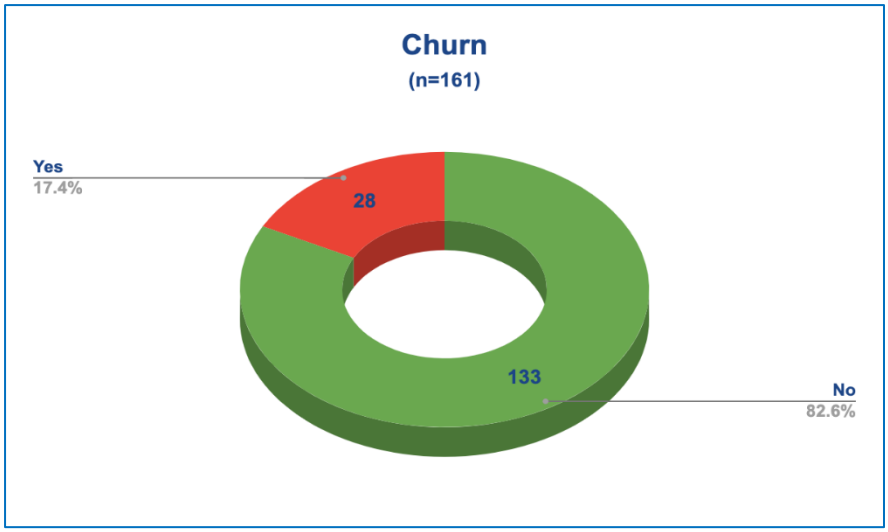
Most respondents who have subscribed to the premium version is because of video streaming (33.5%) in the first place followed by free and fast delivery options (31.1%)



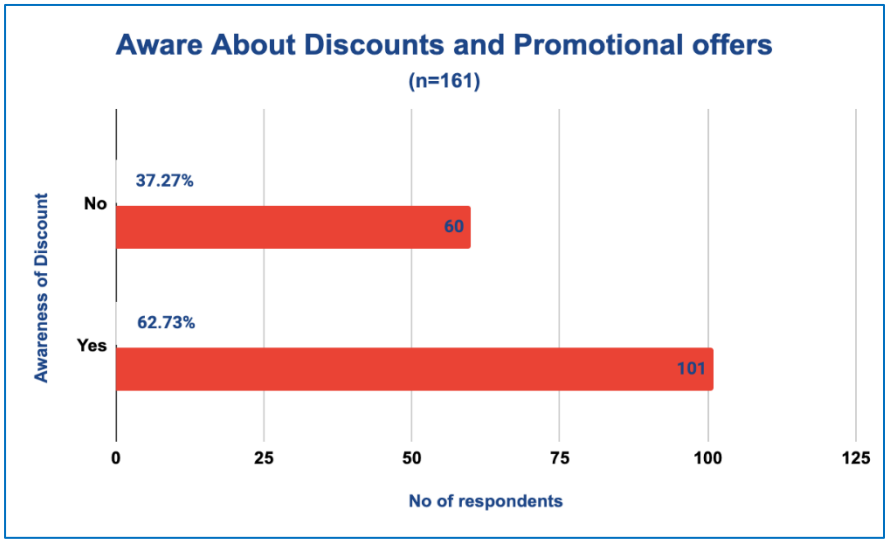
62.1% of users are Likely (33.5%) and very likely (28.5%) to recommend prime subscription to others. Whereas only about 3% would not recommend prime to others.



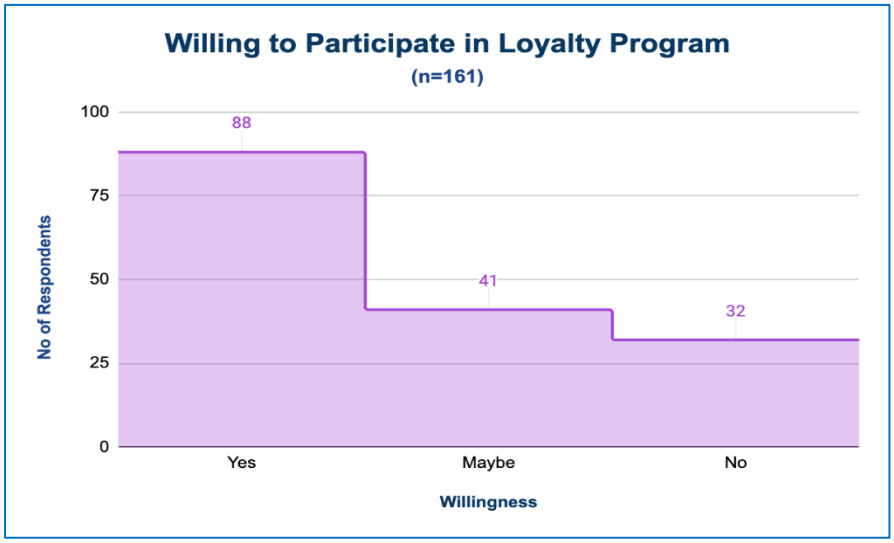
75.7% of users (who have rated 4 and 5) are likely to continue future usage of the amazon platform.



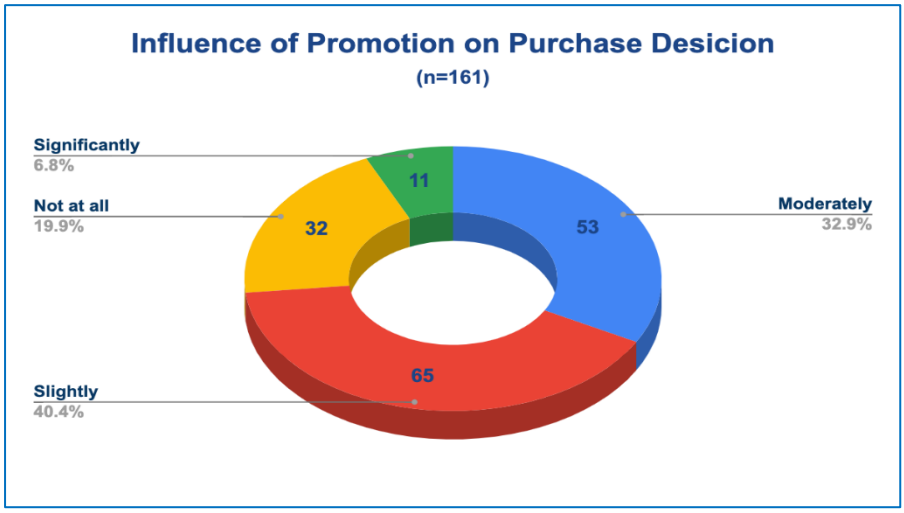
17.4% of the respondents are willing to churn out of the amazon ecosystem or willing to switch to other available alternatives



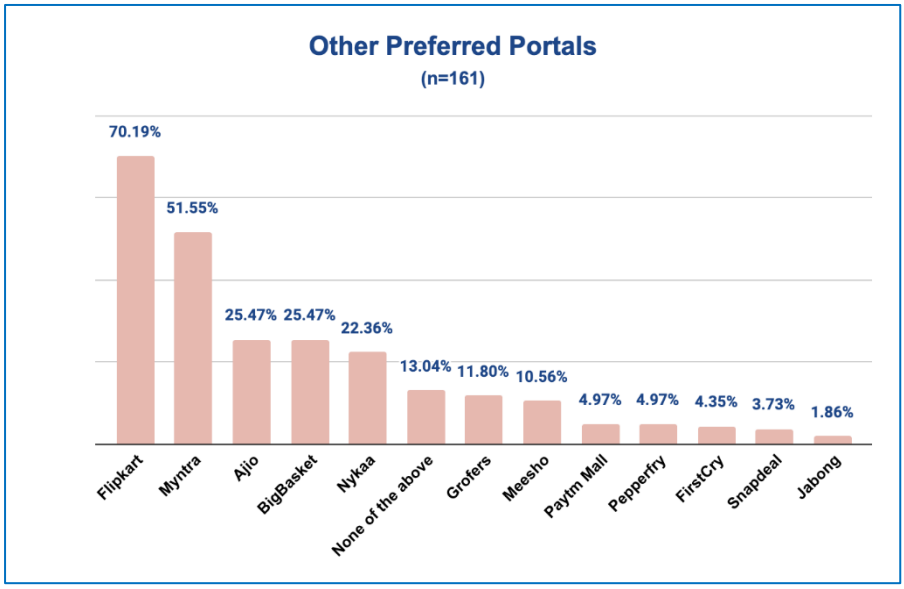
Majority of respondents (62.7%) are aware of discounts and promotional offers run by amazon



54.6% of respondents are willing to participate in Amazon Loyalty programs.

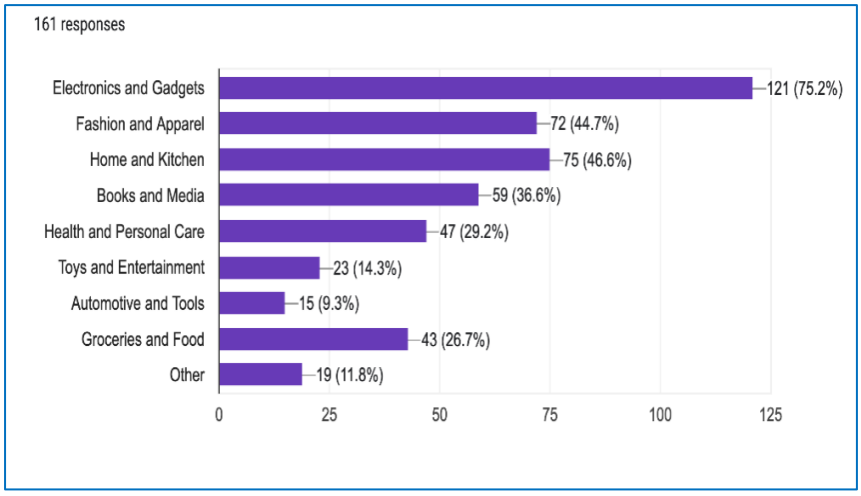


Most respondent's purchase decision is slightly (40.4%) or moderately influenced (32.9%) by Amazon's promotion



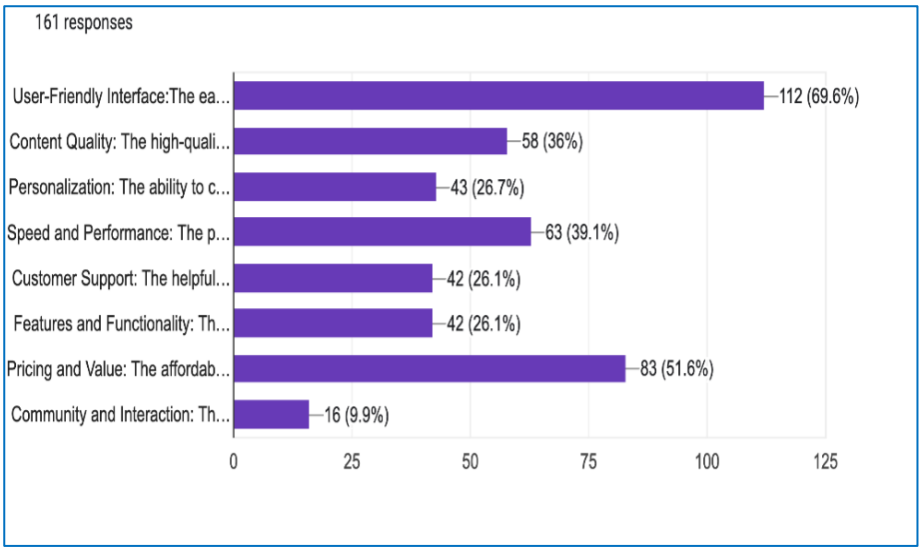
In comparison to Amazon, Flipkart (with 70.1% votes) is the most preferred portal followed by Myntra (51.55%). The most voted for combination is that of Flipkart and Myntra (8.7%)

Most Shopped category

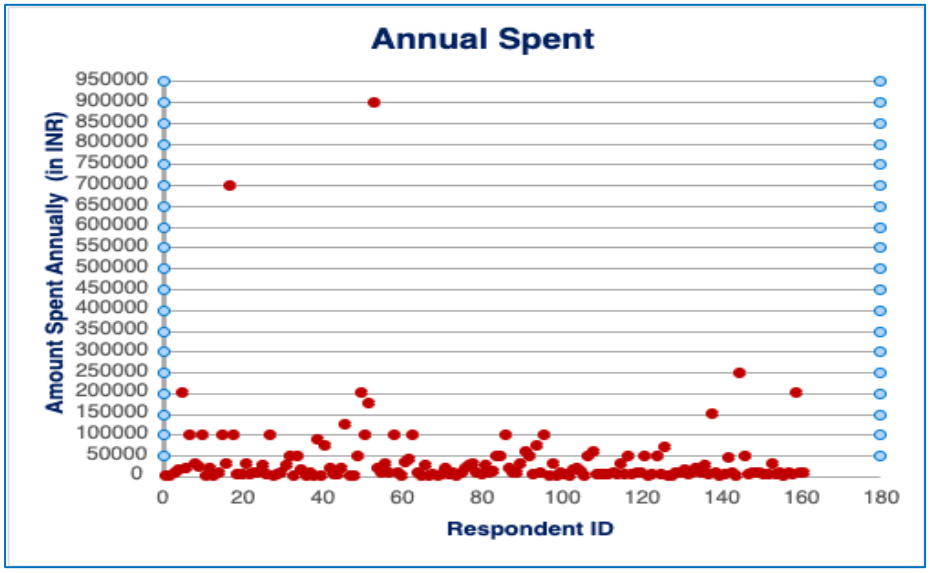


The most shopped category on Amazon as voted by respondents is electronics and gadgets (75.2%) followed by Home and kitchen at 46.6% with fashion and apparel at 44.7%. Most voted for combination are -(Home and Electronics + fashion and electronics) with 4.35% votes each

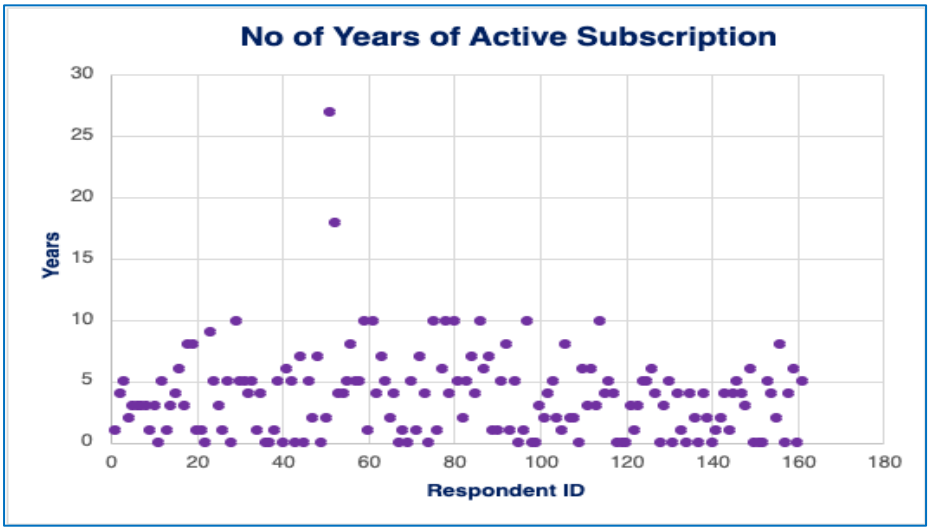
Most Appealing Feature of Amazon



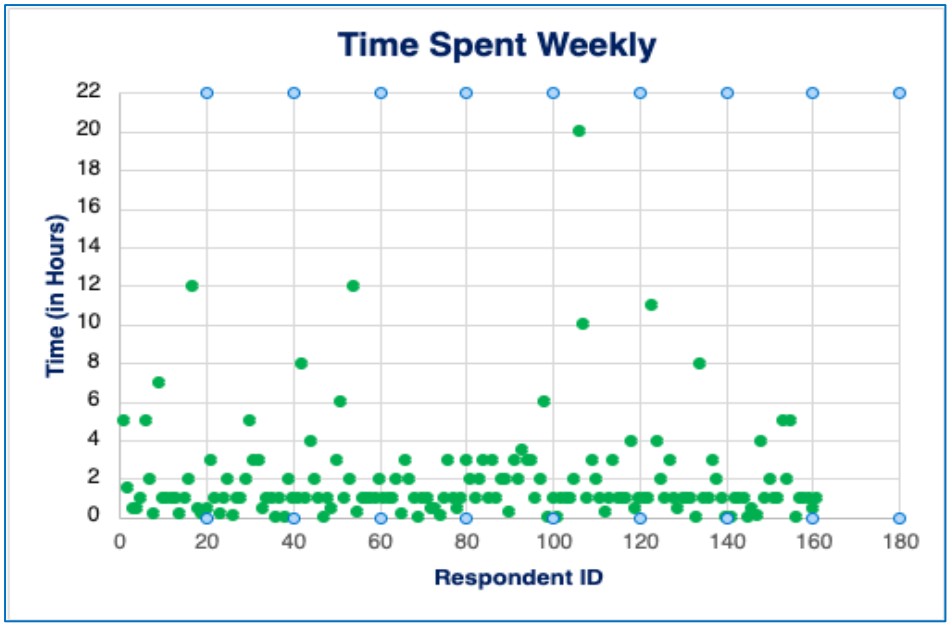
The feature that stands out the most is user friendly interface (69.6%) of Amazon platform. This is followed by the pricing and value that it provides (51.6%)



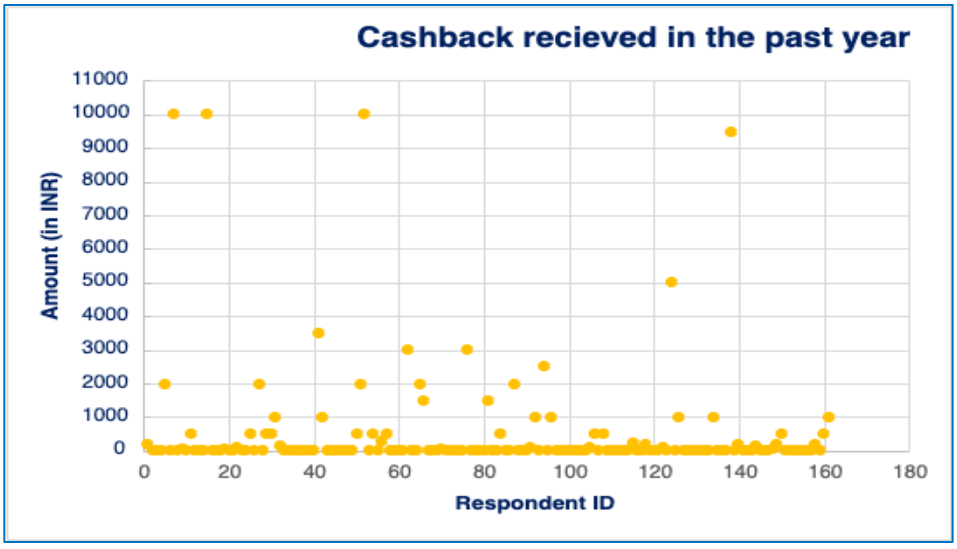
Majority of the users spend less than 50,000 INR per year on amazon.



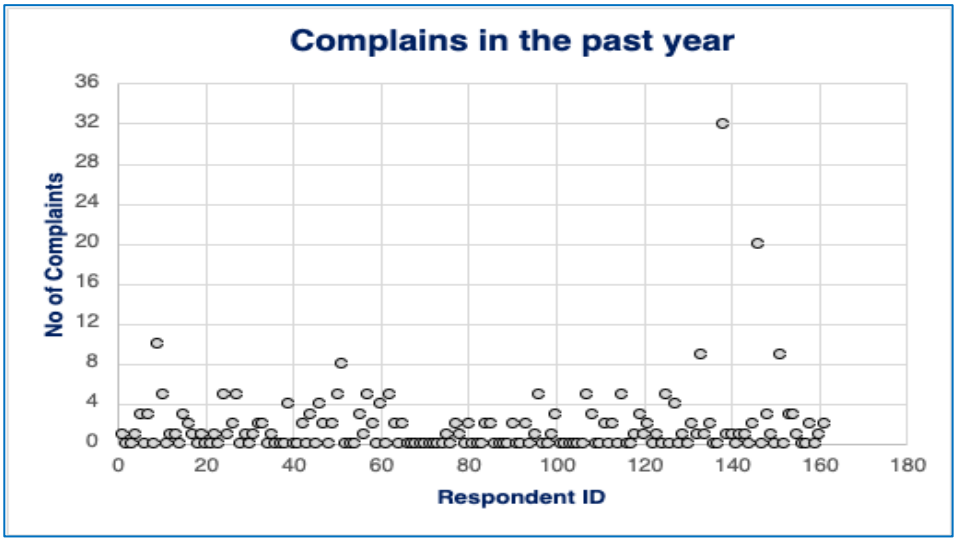
Most of the respondents have had an active subscription for a time span of 5 years or less.



On an average, majority of respondents spend less than 2 hours per week while using the platform



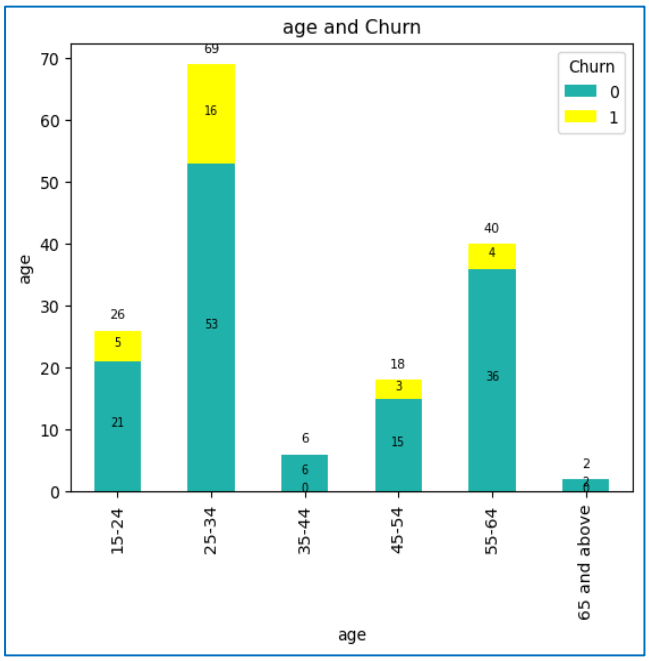
Majority of the respondents have not received any cashback in the past year. Amongst those who have received it most of them have received a cashback of 2000 INR or less.



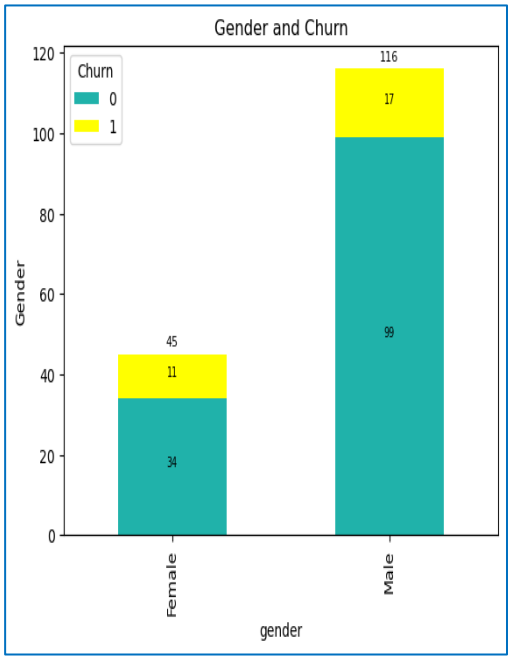
Majority of respondents have no complaints filed in the past year followed by less than 4 complaints per year.

BIVARIATE ANALYSIS

A



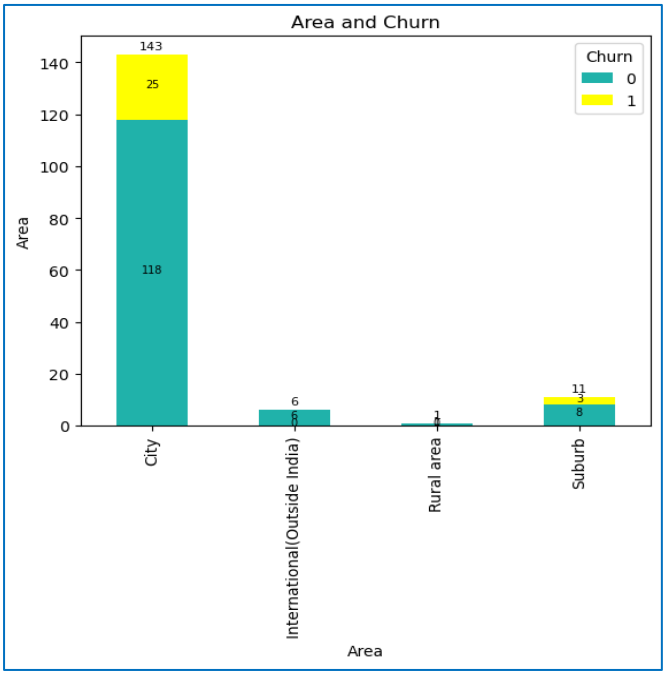
B



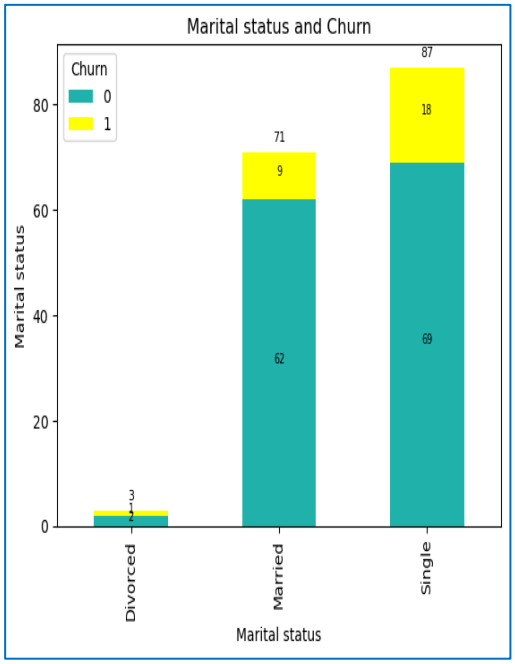
A. 23% of customers belonging to 25-34 yrs tend to churn out, followed by 19% of 15-24 yrs.

B. With a 24.4% of churn rate, female have a higher tendency to churn out than that of male with a churn rate of 14%

A



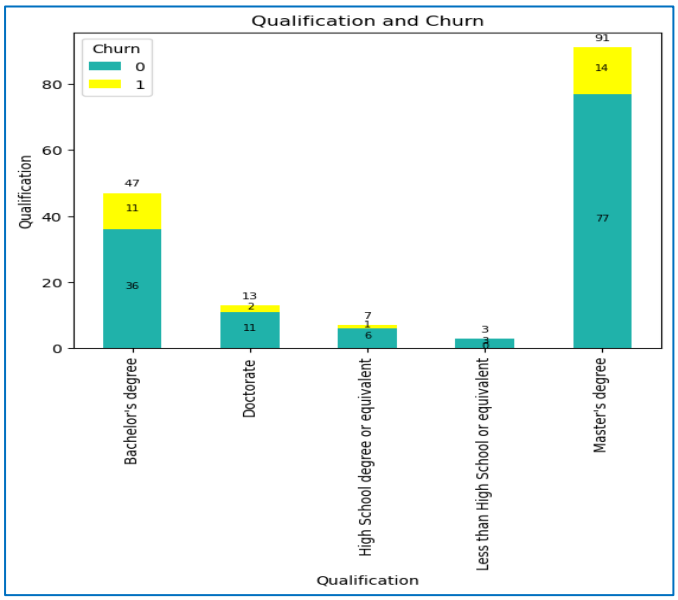
B



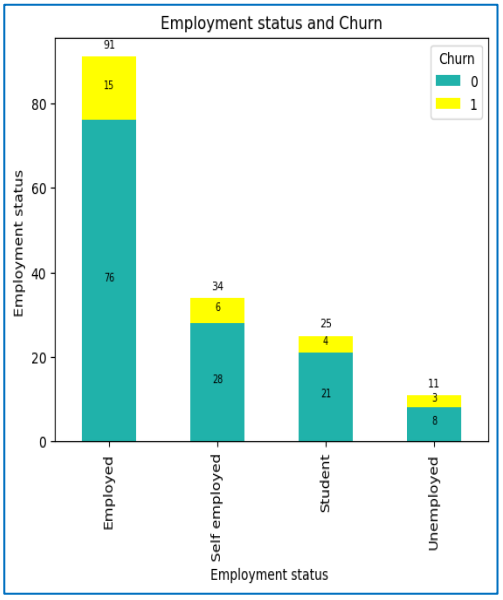
A. With a 27.2% of churn rate, customers in the suburb have a highest tendency to churn out followed by that of city with a churn rate of 17%.

B. Married customers have churned out the least (12.6%),20% single churn out along with 33% divorcee.

A



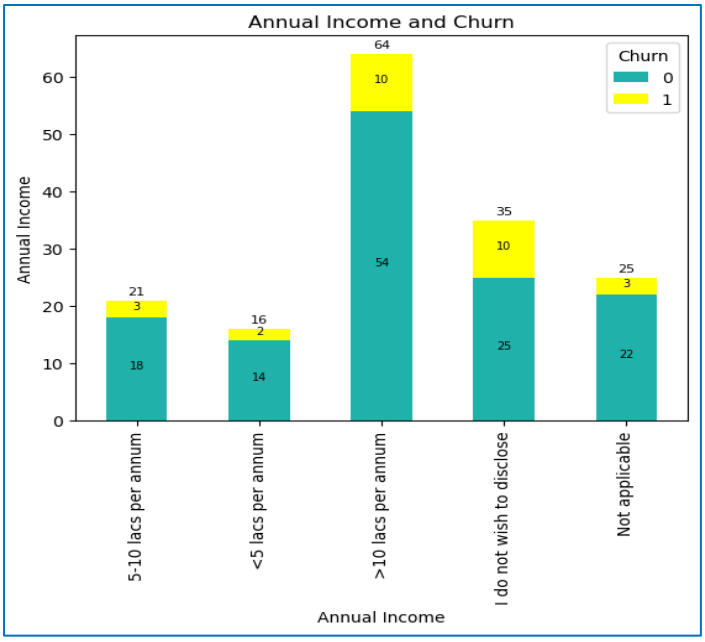
B



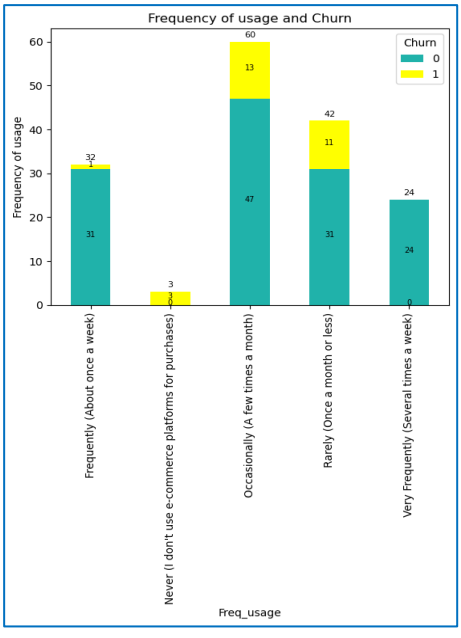
A. Customers with Bachelor's degree churn out the most (23%) while those with degree lesser than high school don't churn out.

B. Unemployed customers churn out the most (27.2%). The rest of the categories stand at an even churn rate of 16–17%

A



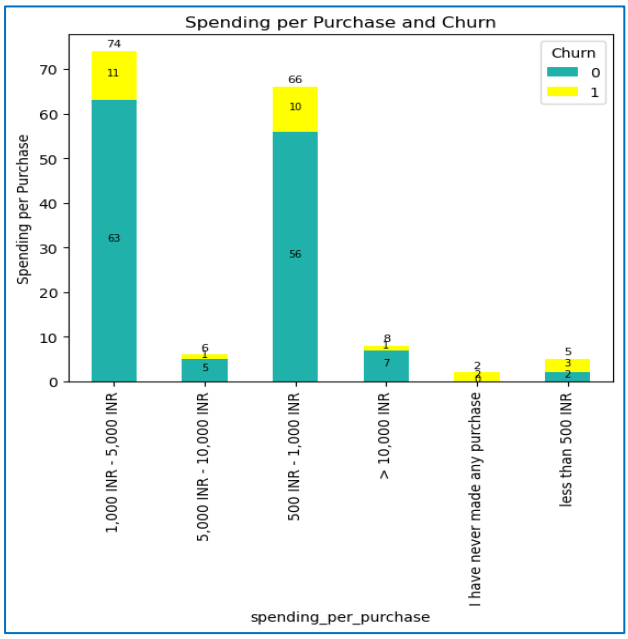
B

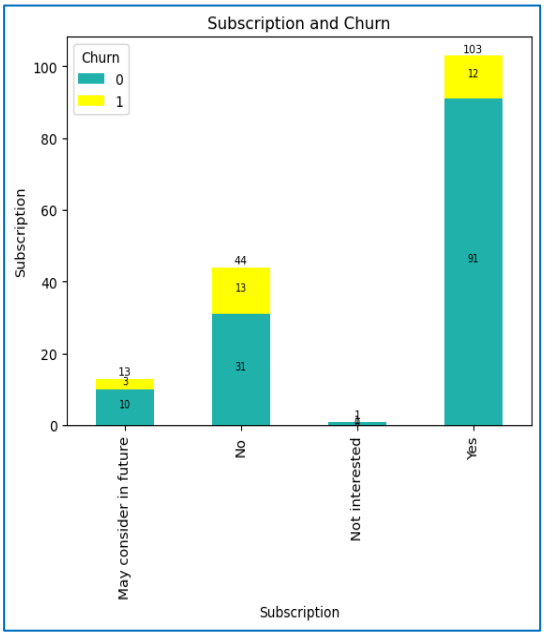


A. Customers who do not to disclose their income churn out the most (28.5%). Remaining categories show a churn rate increasing with the increase in income category ranging from 12% to 15.6%

B. Rare(26.1%) and occasional users(21%) churn out most whereas the frequent users churn out the least(3.1%)

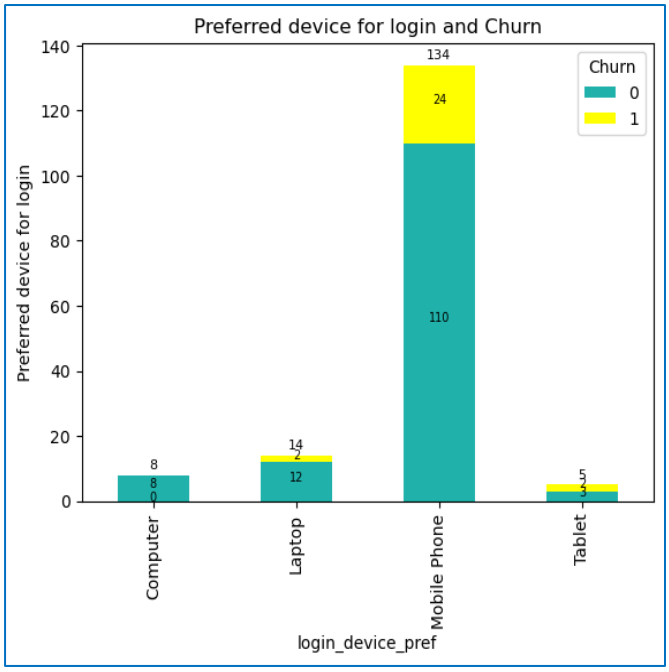
A B

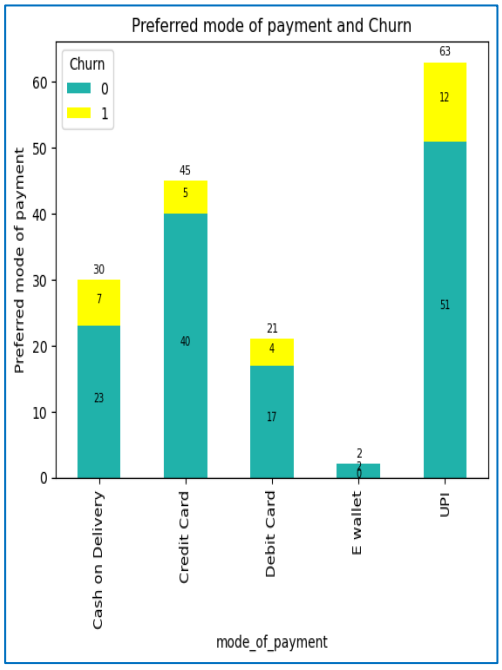




A. Customers who spend less than 500 INR per purchase have a very high churn rate of 40%. For remaining categories, the rate ranges from 12.5% to 16.6%.
B. Customers who are not a subscriber churn out the most with a rate of 29.5%

A B

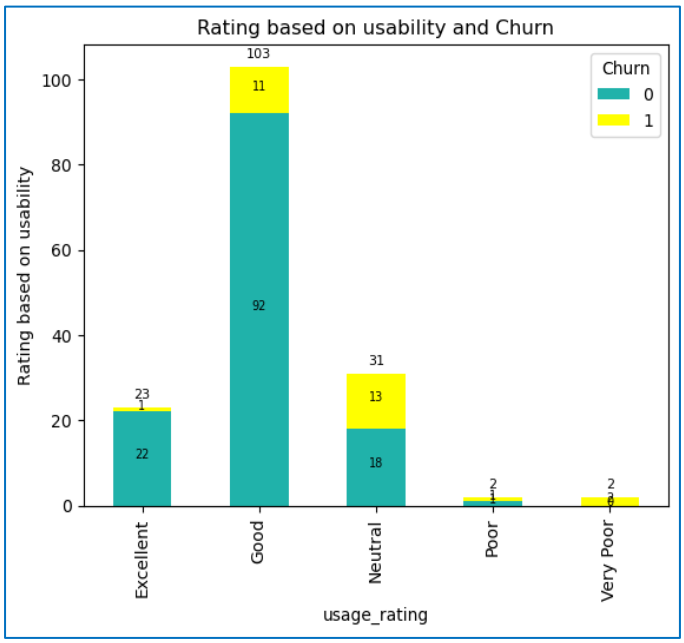




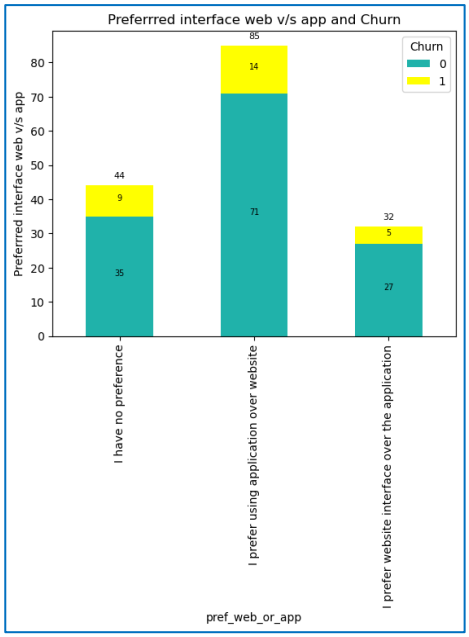
A. Customers using tablet to login, churn out the most (40%) followed by mobile (17.9%) and laptop at 14.2%

B. Customers opting for Cash on delivery churn out the most (23.3%) followed by debit card and UPI, each at 19%. Customers paying via credit card Churn out the least (11.1%)

A



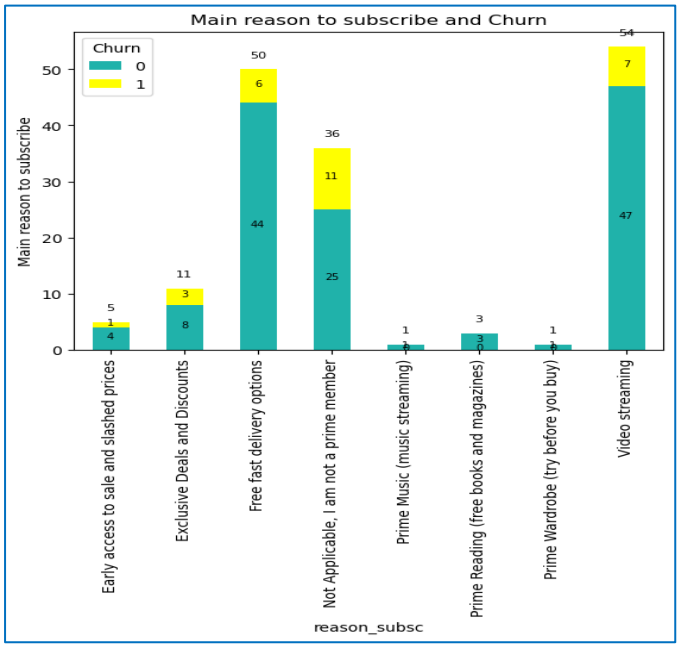
B



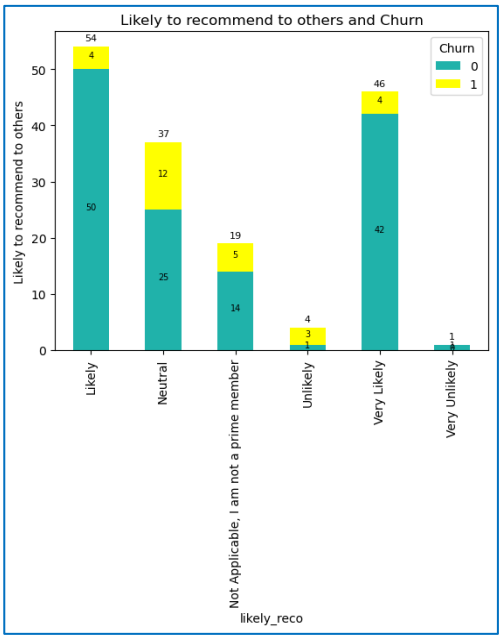
A. The churn rate increases with decrease in the usability rating. 41% for neutral followed by 50% for poor and 100 % for very poor.

B. Customers who have no preference for either website or app churn out the most 20.4%, while those who prefer website only or app only are 15.6% and 16.4% respectively.

A

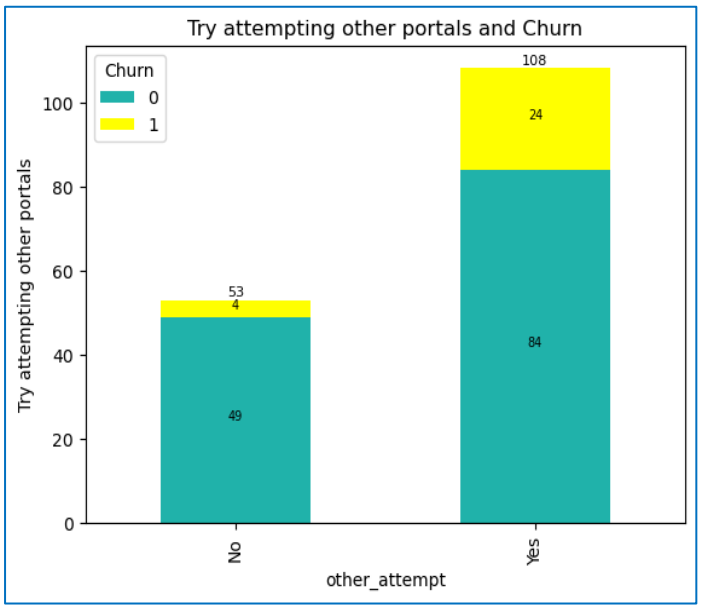


B

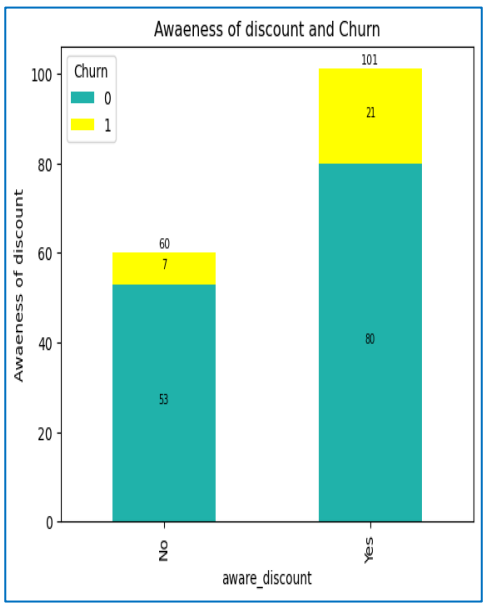


A. Customers who have subscribed for exclusive deals and discounts churn at a rate of 27.2 % followed by 20% who seek out early access to sales and slashed prices.
B. Those who are unlikely to recommend to others churn out the most with 75% rate.

A

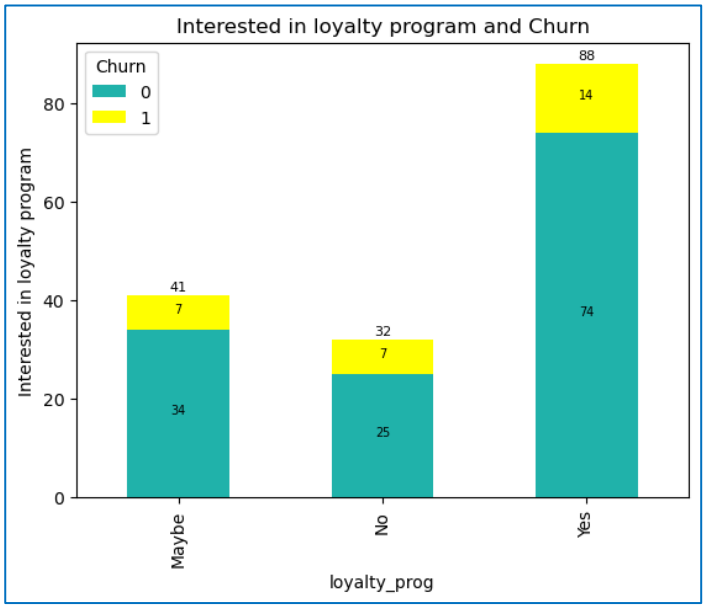


B

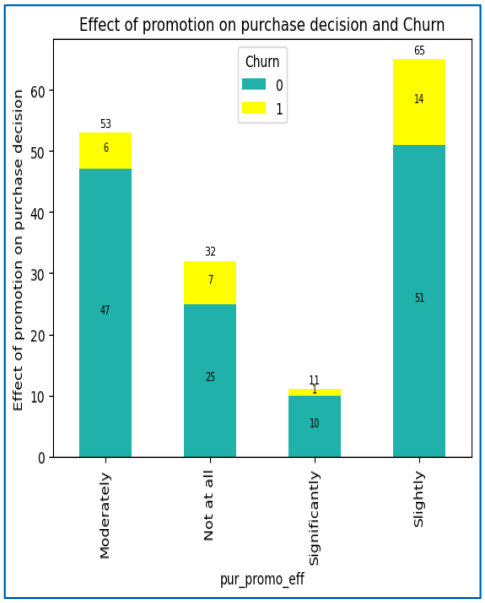


A. 22% of those who have attempted other portals churn out, while only 7% of those who didn't attempt other portals churn out.
B. Those customers who are aware of discounts churn out more than those who are not (20.7% v/s 11%)

A



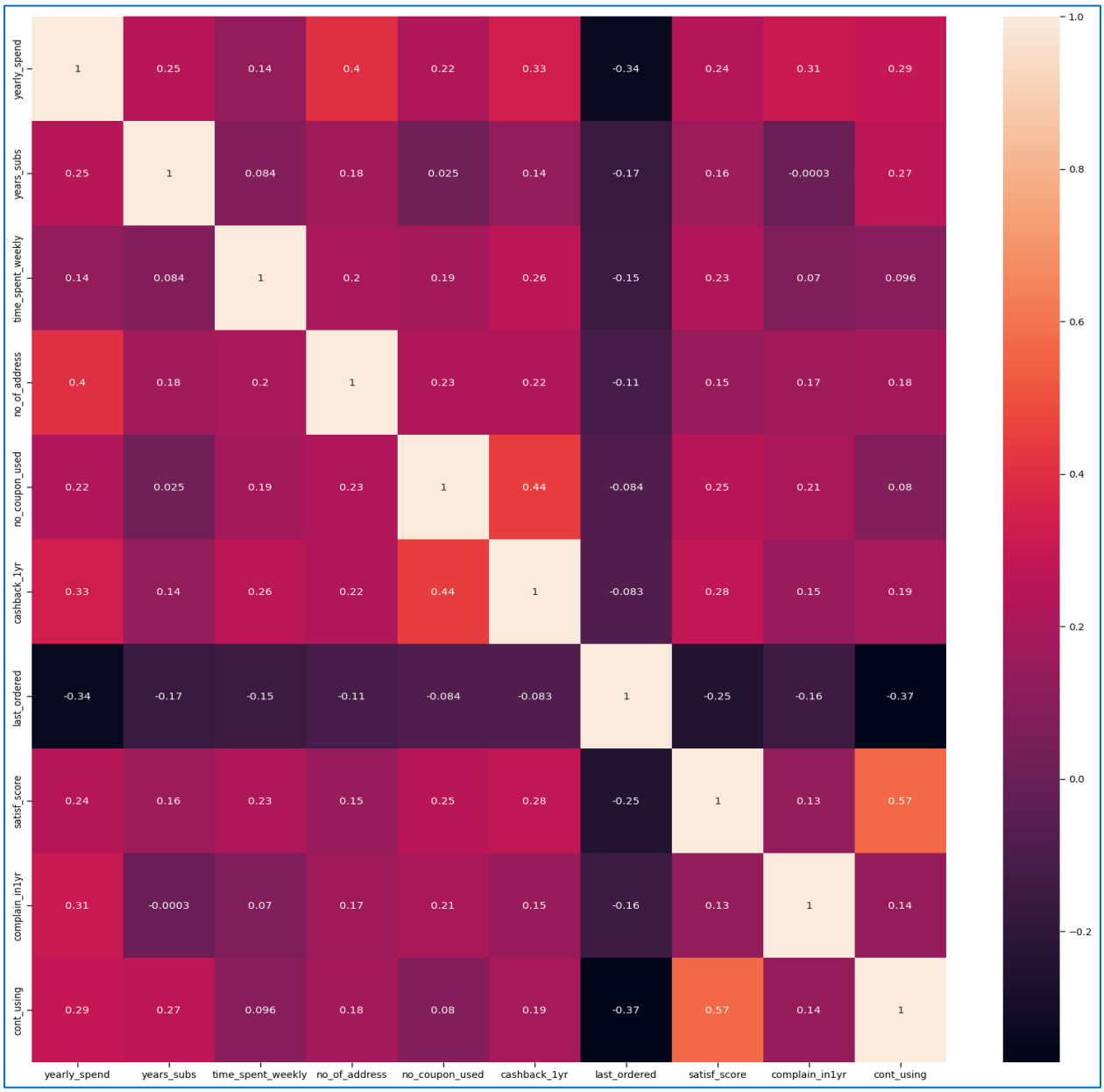
B



A. Respondents who are not interested in any kind of loyalty program churn out the most with 21%
B. Those who are not at all or slightly affected churn out at a rate of 21% each approximately, while those who are affected significantly churn only 9%

Multivariate Analysis

For Numerical Variables



NEGATIVE CORRELATION

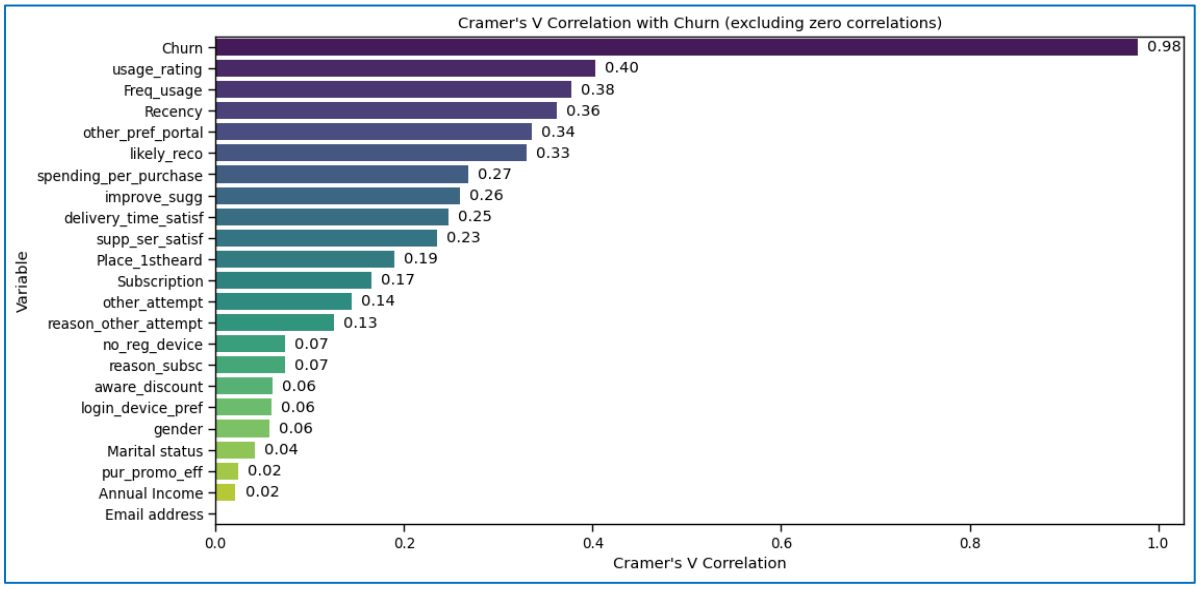
- A negative correlation exists (-0.25) between day last ordered and satisfaction score indicating that if the satisfaction score is higher then number of days from the last order placed is lower, resulting in more recent purchases.
- A higher negative correlation exists (-0.34) between day last ordered and amount spent yearly indicating that if the amount spent yearly is higher then number of days from the last order placed is lower, pointing towards greater feeling of reliability of the customers towards the portal.

- The highest negative correlation exists (-0.37) between day last ordered and willing to continue usage of portal indicating that if the customer is willing to continue using the portal then number of days from the last order placed is lower, pointing towards greater feeling of value provided by each purchase made via the portal.

POSITIVE CORRELATION

- A positive correlation of 0.5 exists between satisfaction score and willingness to continue using the portal. This indicates that the two factors are directly proportional to each other.
- A positive correlation of 0.44 exists between cashback amount received in past 1 year and number of coupons redeemed while using the portal. This indicates that as the number of coupons redeemed increases so does the cashback on the purchase.
- A positive correlation of 0.4 exists between number of addresses saved against an account and yearly spent while using the portal. This indicates that as the number of addresses increases so does the amount spent.

For categorical variables



Cramer's V ranges from 0 to 1. A value of 0 indicates no association between the variables, while a value of 1 implies a perfect association. The interpretation of Cramer's V is similar to other correlation coefficients. Common guidelines are:

a. Small association: $V<0.1$

b. Medium association: $0.1 \leq V<0.3$

c. Large association: $V \geq 0.3$

A large association with churn can be seen for the following variables:

1. Usability rating given by the respondent (0.4)
2. How frequently a respondent uses the portal (0.38)

3. Recency (0.36)
4. Likely to recommend (0.33)

Hypothesis Testing

A hypothesis is an assumption that is made based on some evidence. This is the initial point of any investigation that translates the research questions into predictions. It includes components like variables, population and the relation between the variables. A research hypothesis is a hypothesis that is used to test the relationship between two or more variables.

A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration. As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables. A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable. Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal. They cannot have a normal distribution since they can only have a few particular values.

1. Usability Rating and Churn

- Null Hypothesis (*H*0): There is no association between Usability Rating and Churn.
- Alternative Hypothesis (*H*1): There is a significant association between Usability Rating and Churn.

Chi-Square Value: 29.93
Degrees of Freedom: 4
P-value: 5.052393429485121e-06
There is a significant association between the variables.

2. Frequency of Portal Usage and Churn

- Null Hypothesis (H0): There is no association between the frequency of portal usage and Churn.
- Alternative Hypothesis (H1): There is a significant association between the frequency of portal usage and Churn.

Chi-Square Value: 26.86
Degrees of Freedom: 4
P-value: 2.118987601495148e-05
There is a significant association between the variables.

3. Recency and Churn

- Null Hypothesis (H0): There is no association between Recency and Churn.

- Alternative Hypothesis (H1): There is a significant association between Recency and Churn.

Chi-Square Value: 25.96
Degrees of Freedom: 5
P-value: 9.064396225653272e-05
There is a significant association between the variables.

4. <u>Other Preferred Portal and Churn</u>

- Null Hypothesis (*H*0): There is no association between having another preferred portal and Churn.
- Alternative Hypothesis (*H*1): There is a significant association between having another preferred portal and Churn.

Chi-Square Value: 76.37
Degrees of Freedom: 58
P-value: 0.053250726975993826
There is no significant association between the variables.

5. <u>Likelihood to Recommend and Churn</u>

- Null Hypothesis (*H*0): There is no association between Likelihood to Recommend and Churn.
- Alternative Hypothesis (*H*1): There is a significant association between Likelihood to Recommend and Churn.

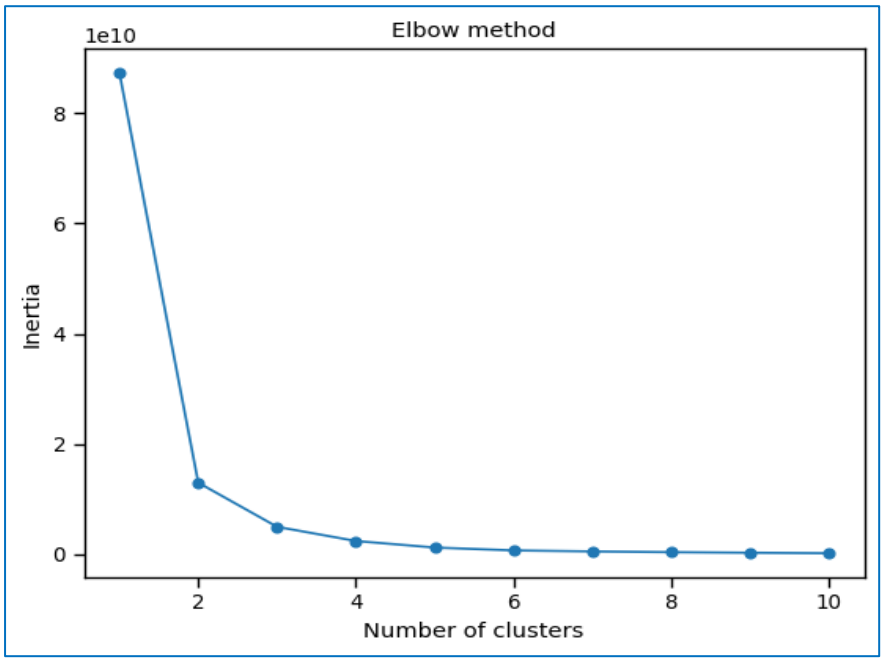
Chi-Square Value: 22.49
Degrees of Freedom: 5
P-value: 0.00042089976001048507
There is a significant association between the variables.

From this, we can conclude that usability rating given by user, portal usage, recency of portal usage, likelihood to recommend other with Churn, show significant association at 95% confidence.

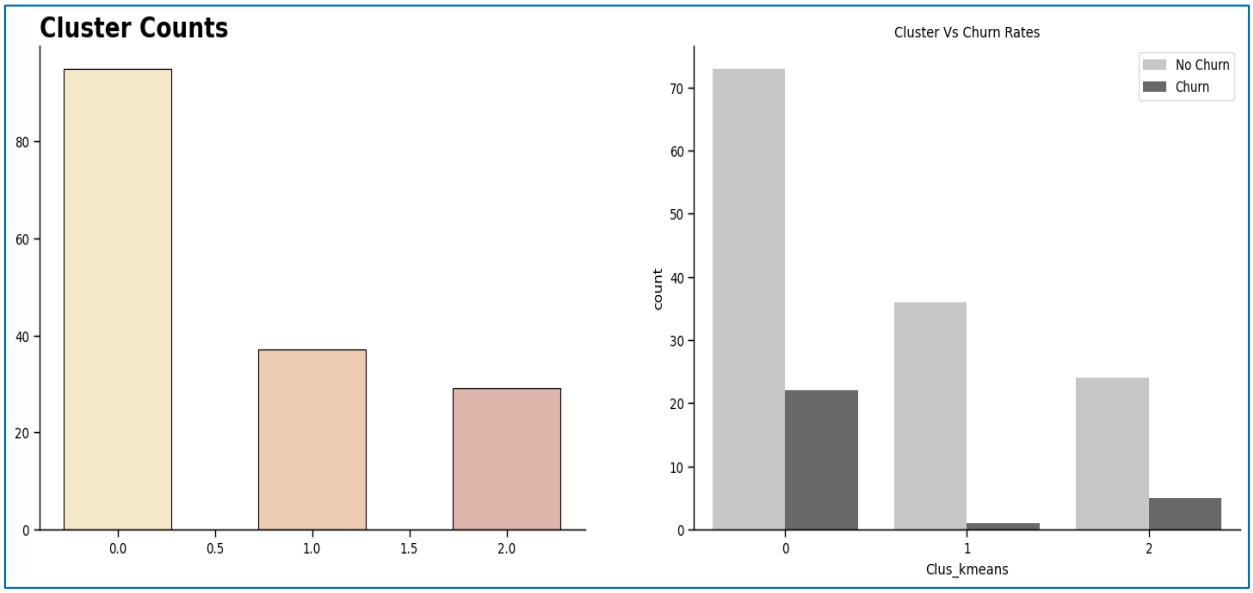
<u>Outlier treatment</u>

Some variables like amount spent per year, amount spent per purchase, number of years of active subscription and number of devices registered had outliers present in them. Most of the outliers have been treated by using median imputation method. Substituting the outlier with a median is preferable as that compared to mean imputation because mean leads to be affected more by the presence of outlier. Post treating outlier, data imbalance was handled by SMOTE method.

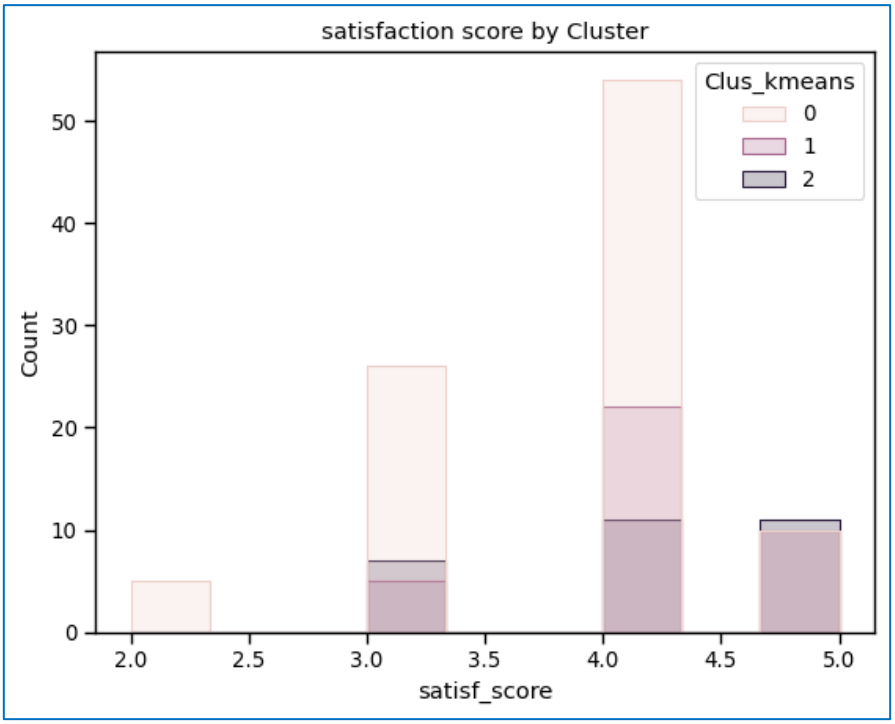
K means



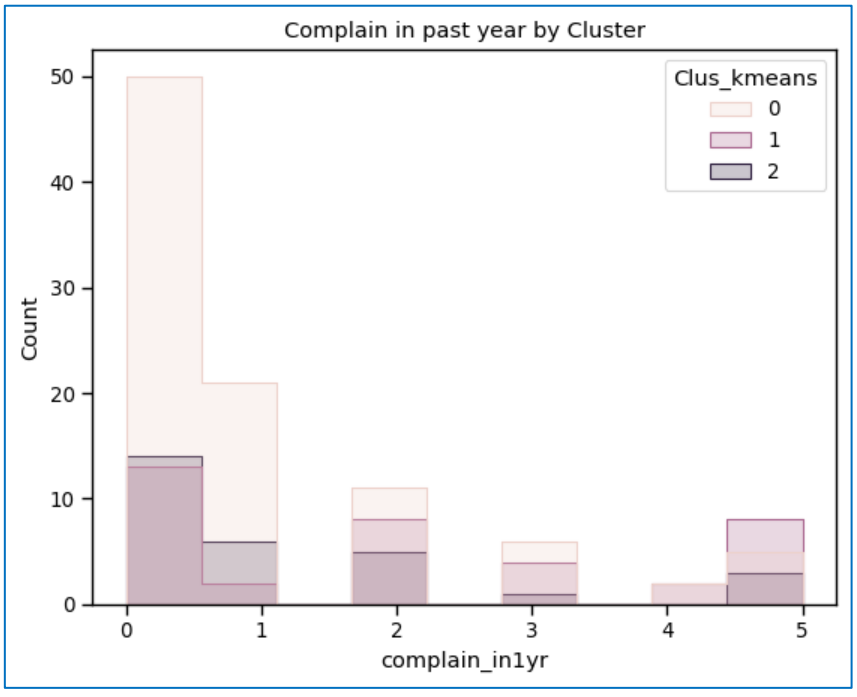
The number of optimum clusters are 3 as seen from the elbow graph above.



The three clusters along with their churn count can be seen from the above visualization. Most of the customers belonging to the 1st cluster (cluster 0) have churned out at a higher rate, followed by the 3rd cluster (cluster 2). The least churn amongst the three is seen in the 2nd cluster (cluster 1).



Analyzing prevalent features at a cluster level. Since the churn out rate for cluster 0 is most, no respondent has given a satisfaction score of 5. The churn percent of cluster 1 is the least hence all the respondents have rated satisfaction score as 3 or above. This indicates a neutral mindset (a rating of 3 on 5) to positive mindset (rating a 5 on 5) of the respondents belonging to this cluster. Cluster 2 respondents are either neutral (rating a 3 on 5) or extremely satisfied (rating a 5 on 5).



The respondents belonging to cluster 0 are maximum in number. The number of complaints reported by them in the past year is maximum for all the categories (0 complaints per year, 1 complaint per year, 2 complaints per year and 3 complaints per year and 4 complaints per year), except for the category 5 complaints per year. In the category of 5 complaints per year respondents of cluster 1 top amongst other clusters. This indicates customers belonging to cluster 0 are higher in count hence fetch place in each category. The higher churn rate can be contributed by the customers who file higher number of complaints per year (2 or more complaints per year).

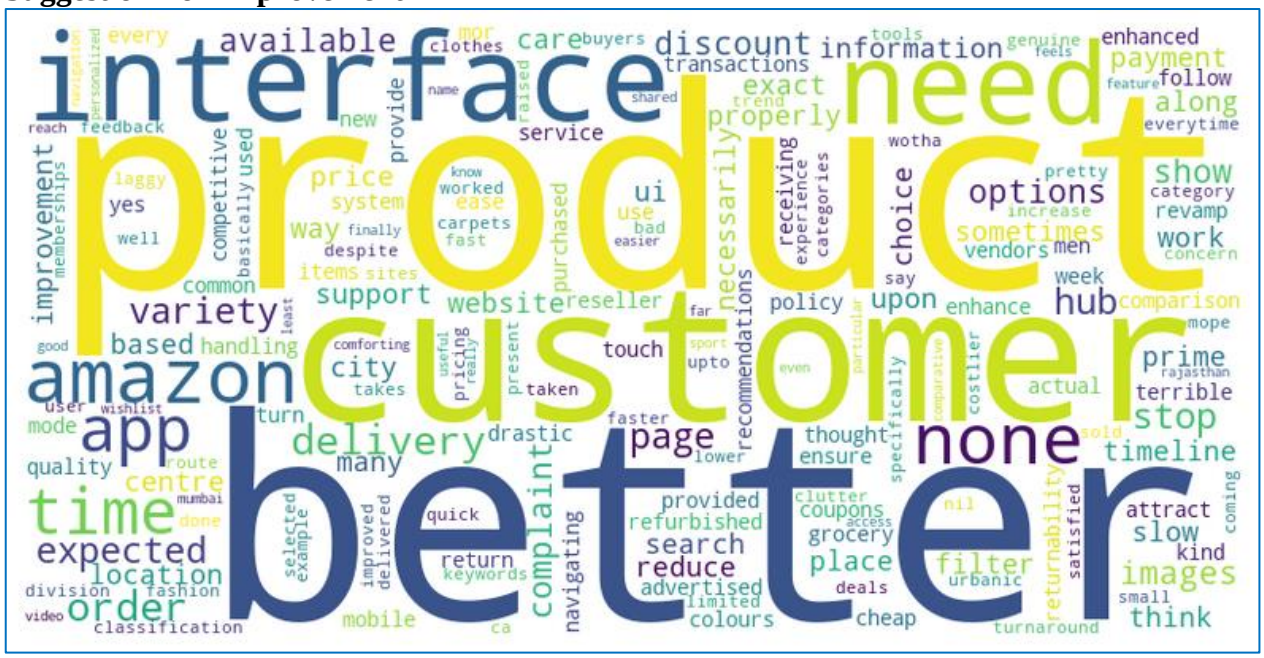
The customers belonging to cluster 2 mimics the pattern of cluster 0 but with a lower percent in each category. Yet the churn behavior of the customers belonging to cluster 2 varies from that of cluster 0, this could be due to the fact that these customers do not give similar weightage to the number of complaints in the past year and weigh other factors higher while deciding to churning out.

Qualitative Analysis

Reason To Attempt Other Portals

Most of the respondents have opted for other portals because of better aspects of price and availability. Product variety and better delivery have been areas of key concern that have led to trial of other portals. Flipkart has emerged as the most sought out alternative. The amazon interface was also an area of key concern.

Suggestion For Improvement

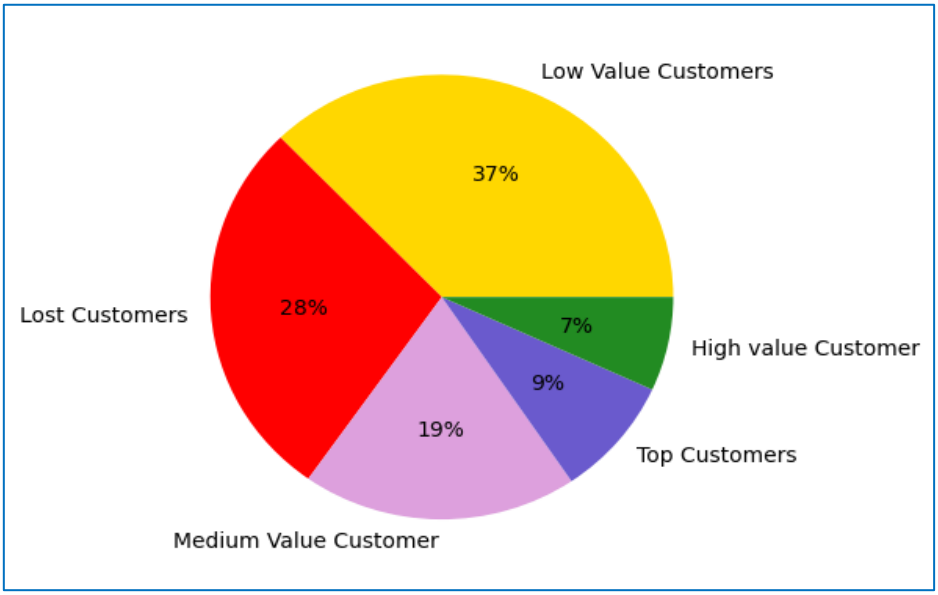


The key areas of improvement include better product range, improved portal interface along with greater understanding of customer need.

RFM Segmentation

The RFM score is calculated on a scale of 5. The weightage given to Recency is 0.15, Frequency is 0.28 and that of Monetary is 0.57. The criteria for customer segmentation are as follows:

- RFM score >4.5: Top Customer
- 4.5 > RFM score > 4: High Value Customer
- 4> RFM score >3: Medium value customer
- 3> RFM score>1.6: Low-value customer
- RFM score<1.6: Lost Customer



From the above graph it can be concluded that more than one-third (37%) of the customers can be classified as Low value customers, followed by Lost customers who account for 28% of the sample size.
Only 7% of respondents are categorised as high value customers along with 9% of them being the top value customers. Individual strategies and scheme need to be devised for each segment in order to cater to their needs in a better manner.
Customer segmentation based on RFM (Recency, Frequency, Monetary) scores is a common practice in marketing. Once customers are categorized into segments, tailored retention strategies can be implemented. Here are suggested retention policies for each segment:

Top Customer (RFM score > 4.5):

- Policy: Offer exclusive loyalty programs and VIP treatment.
- Action: Provide personalized offers, early access to new products/services, and dedicated customer support. Acknowledge their loyalty with special rewards.

High-Value Customer (4.5 > RFM score > 4):

- Policy: Encourage continued spending and engagement.

- Action: Offer tiered rewards, special promotions, and early access to sales. Provide personalized recommendations based on their past purchases to increase the average transaction value.

Medium-Value Customer (4 > RFM score > 3):

- Policy: Incentivize increased frequency and spending.
- Action: Send targeted promotions to encourage more frequent purchases. Implement loyalty programs with achievable milestones and rewards. Provide bundle deals to increase the average order value.

Low-Value Customer (3 > RFM score > 1.6):

- Policy: Nurture and convert into higher segments.
- Action: Send re-engagement campaigns, providing incentives for a return visit or purchase. Offer discounts or promotions to encourage repeat business. Collect feedback to understand and address any issues.

Lost Customer (RFM score < 1.6):

- Policy: Attempt to re-engage and understand reasons for disengagement.
- Action: Send targeted win-back campaigns with special offers or discounts. Request feedback to identify and address any concerns that led to disengagement. Consider personalized communication to reconnect.

It is important to continuously analyse the effectiveness of these retention strategies and adjust them based on customer feedback and evolving market conditions. Regularly update customer segmentation and RFM scoring to ensure the strategies remain relevant. Additionally, consider incorporating other data sources and customer behaviour metrics for a more comprehensive understanding of your customer base.

Predictive Modelling

1. Area Under the Receiver Operating Characteristic Curve (AUC-ROC):

- Interpretation:
    - AUC-ROC measures the area under the curve of the receiver operating characteristic (ROC) plot. It provides a summary of the model's ability to distinguish between classes across various probability thresholds.
- Interpretation Guidelines:
    - AUC-ROC values range from 0 to 1, where higher values indicate better performance.
    - A model with an AUC-ROC close to 1 suggests good discrimination between positive and negative classes.

2. F1 Score:

- Interpretation:
    - F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall, making it suitable for imbalanced datasets.
- Interpretation Guidelines:
    - F1 score values range from 0 to 1, with higher values indicating better performance.
    - A high F1 score suggests a good balance between precision and recall, meaning the model is providing accurate predictions while capturing a significant portion of the positive class.

3. Precision:

- Interpretation:
    - Precision is the ratio of true positive predictions to the total number of positive predictions made by the model. It focuses on the accuracy of positive predictions.
- Interpretation Guidelines:
    - Precision values range from 0 to 1, with higher values indicating better precision.
    - High precision suggests that when the model predicts the positive class, it is likely to be correct.

4. Recall (Sensitivity or True Positive Rate):

- Interpretation:

- Recall is the ratio of true positive predictions to the total number of actual positive instances in the dataset. It focuses on the model's ability to capture all positive instances.

- Interpretation Guidelines:
  - Recall values range from 0 to 1, with higher values indicating better recall.
  - High recall suggests that the model is effective at identifying a large proportion of actual positive instances.

Practical Considerations:

Trade-off between Precision and Recall: There is often a trade-off between precision and recall. Increasing one metric may lead to a decrease in the other. The F1 score is useful in finding a balance between the two.

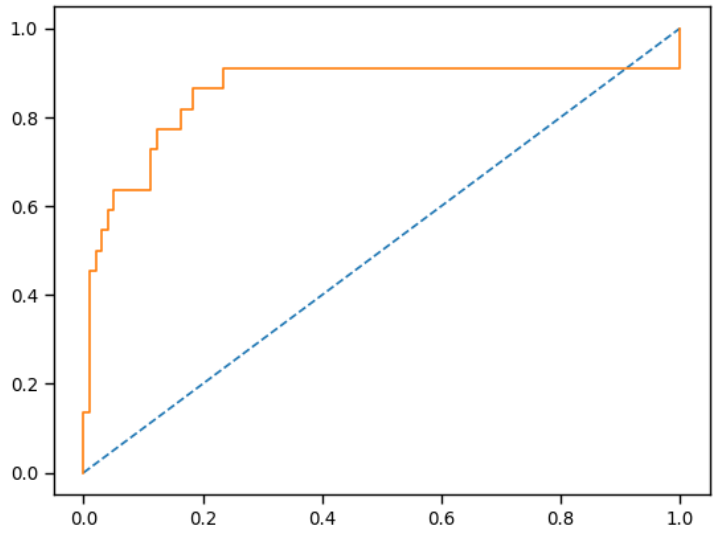
Imbalanced Datasets: In imbalanced datasets where one class is much more prevalent than the other, precision and recall become crucial metrics. A model can achieve high accuracy by simply predicting the majority class, but precision and recall provide insights into its ability to correctly identify the minority class.

Application Context: The choice of evaluation metric depends on the specific goals and requirements of the application. For example, in a medical diagnosis scenario, high recall (minimizing false negatives) might be more critical than high precision.

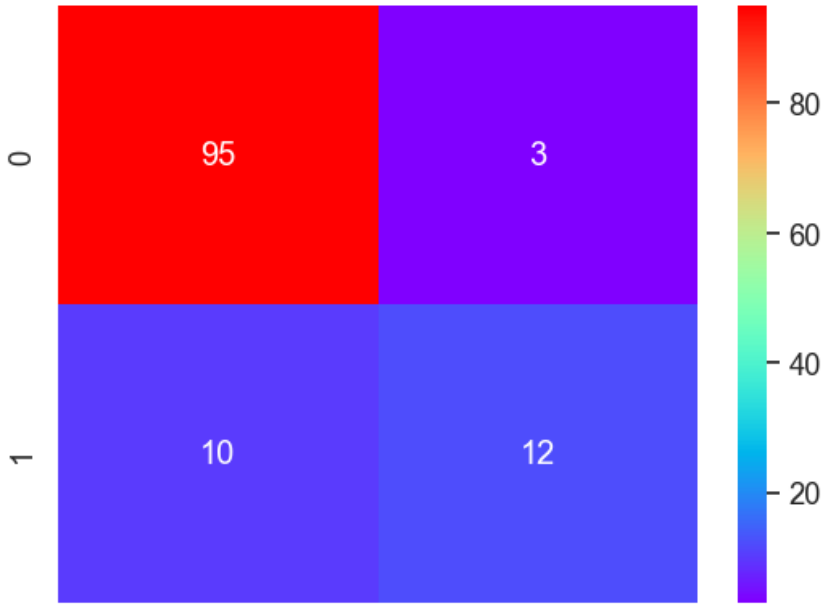
1. LOGISTIC REGRESSION

The AUC-ROC (Area Under the Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classification model at various classification thresholds. The AUC (Area Under the Curve) is a single metric that summarizes the overall performance of the model across different threshold values. The AUC value ranges from 0 to 1, with higher values indicating better model performance.

ROC curve of training data



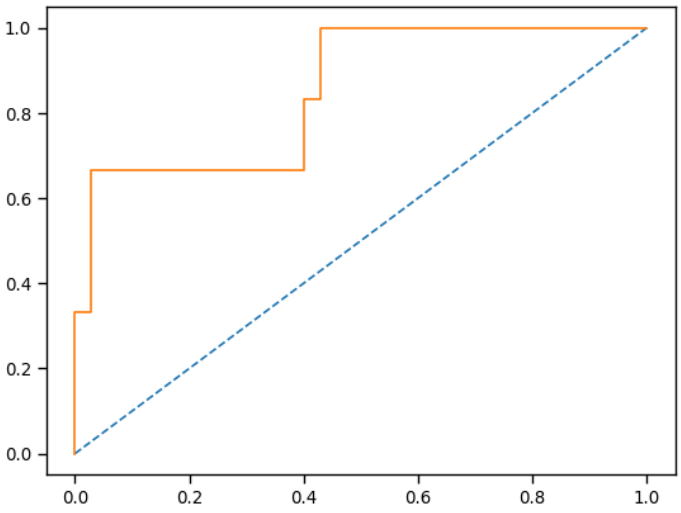
The AUC of the training data is 0.857 which indicates a strong likelihood that the model will assign a higher predicted probability to a randomly chosen positive instance than to a randomly chosen negative instance. This signifies that our model is trained well.



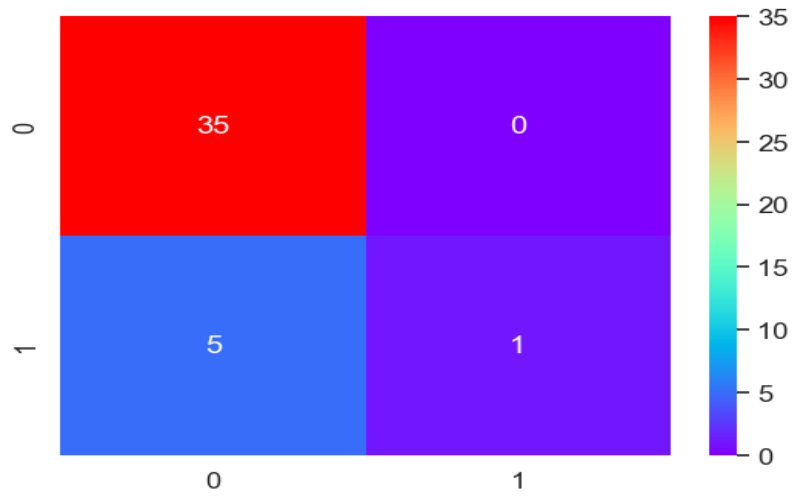
	0	1		
	precision	recall	f1-score	support
0	0.90	0.97	0.94	98
1	0.80	0.55	0.65	22
accuracy			0.89	120
macro avg	0.85	0.76	0.79	120
weighted avg	0.89	0.89	0.88	120

Accuracy of 89% suggests that the model is well trained making correct predictions for both positive and negative classes.

ROC curve of test data



The AUC of the test data is 0.852 which indicates a strong likelihood that the model will assign a higher predicted probability to a randomly chosen positive instance than to a randomly chosen negative instance.

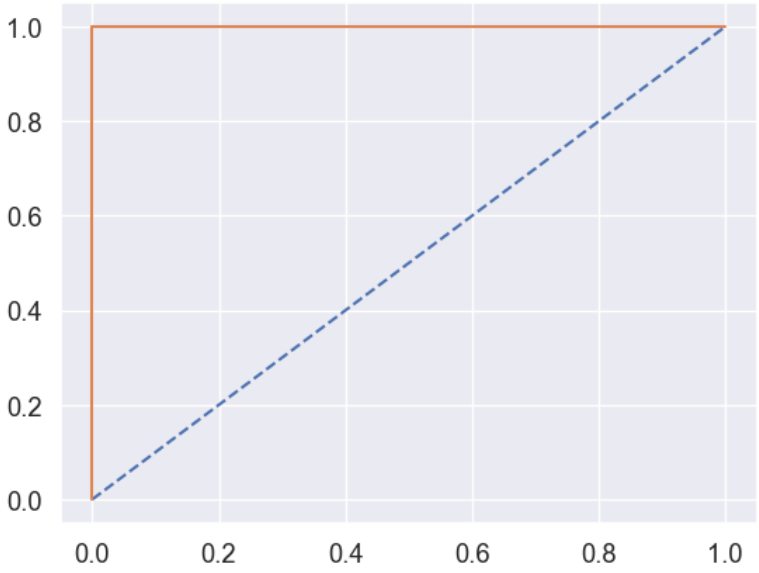


	precision	recall	f1-score	support
0	0.88	1.00	0.93	35
1	1.00	0.17	0.29	6
accuracy			0.88	41
macro avg	0.94	0.58	0.61	41
weighted avg	0.89	0.88	0.84	41

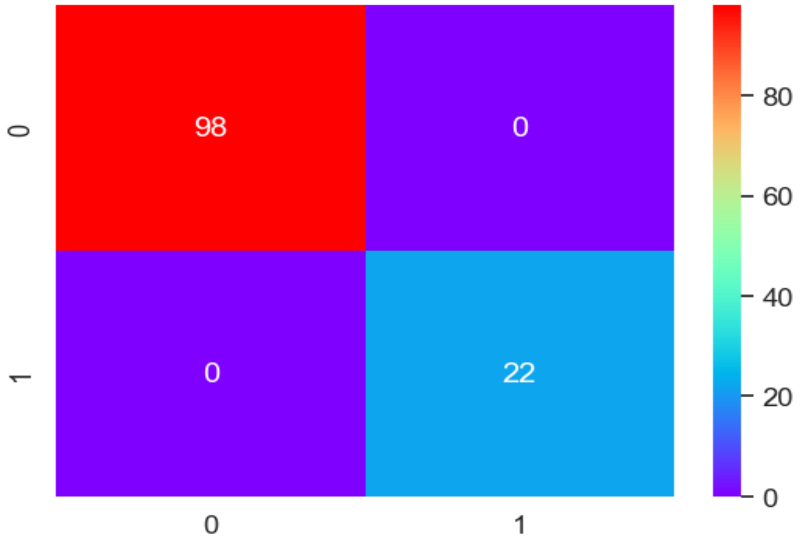
Accuracy of 88% suggests that the model ismaking correct predictions for both positive and negative classes in the test data.

RANDOM FOREST

Train data



The AUC of the training data is 1 which indicates a strong likelihood that the model will assign a higher predicted probability to a randomly chosen positive instance than to a randomly chosen negative instance. This signifies that our model is trained well. We have also performed cross validation in order to check for overfitting if any.

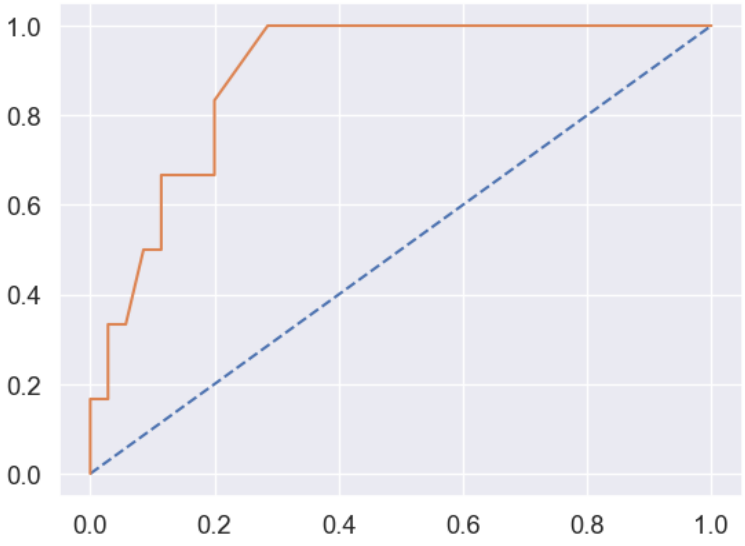


1.0

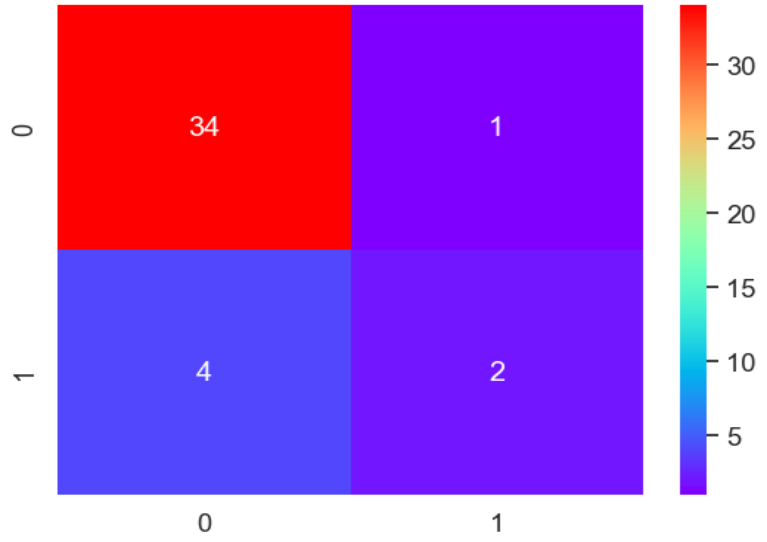
	precision	recall	f1-score	support
0	1.00	1.00	1.00	98
1	1.00	1.00	1.00	22
accuracy			1.00	120
macro avg	1.00	1.00	1.00	120
weighted avg	1.00	1.00	1.00	120

Accuracy of 100% suggests that the model is well trained making correct predictions for both positive and negative classes.

Test data



The AUC of the training data is 0.89 which indicates a strong likelihood that the model will assign a higher predicted probability to a randomly chosen positive instance than to a randomly chosen negative instance. This signifies that our model is trained well. We have also performed cross validation in order to check for overfitting if any.




```
0.8780487804878049
              precision    recall  f1-score   support

           0       0.89      0.97      0.93        35
           1       0.67      0.33      0.44         6

    accuracy                           0.88        41
   macro avg       0.78      0.65      0.69        41
weighted avg       0.86      0.88      0.86        41
```

Accuracy of 88% suggests that the model is well trained making correct predictions for both positive and negative classes

From the above we can see that the Random Forest predicts better though the accuracy for both the logistic regression model and this is same.(88%).

However, in the Random Forest Model the recall and F1 score is better than that of the Logistic Regression. Hence the model used for churn prediction shall be random forest as it handles the imbalanced data set in a better manner.This model shall predict each case with an accuracy of 88 % i.e., there are 88% chance of a customer being categorised correctly by this model.

Findings

- **Location Influence on Churn:** Customers in the suburb exhibit the highest churn rate at 27.2%, surpassing the city's rate of 17%.
- **Marital Status Impact:** Married customers have the lowest churn rate at 12.6%, while singles and divorcees have rates of 20% and 33%, respectively.
- **Educational Background Effect:** Customers with a Bachelor's degree show the highest churn at 23%, while those with education below high school levels exhibit minimal churn.
- **Employment Status and Churn:** Unemployed customers have the highest churn rate at 27.2%, while other employment categories show an even rate around 16-17%.
- **Income Disclosure and Churn:** Customers not disclosing their income have the highest churn rate at 28.5%, with increasing rates for higher income categories (12% to 15.6%).
- **Usage Patterns and Churn:** Rare (26.1%) and occasional (21%) users have the highest churn, contrasting with frequent users who exhibit the least churn at 3.1%.
- **Spending Per Purchase Impact:** Customers spending less than 500 INR per purchase have a notably high churn rate of 40%, compared to other spending categories (12.5% to 16.6%).
- **Subscription Status and Churn:** Non-subscribers have the highest churn rate at 29.5%.
- **Login Device Influence:** Customers using tablets for login show the highest churn at 40%, followed by mobile (17.9%) and laptop (14.2%).
- **Payment Method and Churn:** Cash on delivery customers exhibit the highest churn rate (23.3%), while credit card users have the lowest (11.1%).
- **Usability Rating and Churn:** Churn rate increases with decreasing usability rating, with 41% for neutral, 50% for poor, and 100% for very poor ratings.
- **Web or App Preference and Churn:** Customers with no preference churn the most at 20.4%, compared to those preferring either the website or app.
- **Subscription Preferences and Churn:** Subscribers for exclusive deals and discounts have a churn rate of 27.2%, while those seeking early access to sales and discounts churn at 20%.
- **Likelihood to Recommend and Churn:** Respondents unlikely to recommend churn the most with a rate of 75%.
- **Other Portal Attempts and Churn:** Customers who attempted other portals exhibit a higher churn rate (22%) compared to those who didn't (7%).
- **Awareness of Discounts and Churn:** Customers aware of discounts churn more than those who are not (20.7% vs. 11%).
- **Interest in Loyalty Programs and Churn:** Respondents not interested in any loyalty program have the highest churn rate at 21%.
- **Impact of Affected Sentiment on Churn:** Customers not affected or slightly affected churn at a rate of approximately 21%, whereas significantly affected customers churn only at 9%.
- more than one-third (37%) of the customers can be classified as Low value customers, followed by Lost customers who account for 28% of the sample size. Only 7% of respondents are categorised as high value customers along with 9% of them being the top value customers. Individual strategies and scheme need to be devised for each segment in order to cater to their needs in a better manner.

Suggestions/Recommendations

- Targeted Retention Strategies: Implement targeted retention strategies for customers in the suburb, singles, and those with a Bachelor's degree to mitigate higher churn rates in these segments.
- Improve Usability: Enhance the usability of the platform, especially for customers providing neutral or poor usability ratings, to reduce churn associated with dissatisfaction.
- Incentivize Subscriptions: Encourage subscriptions by offering exclusive deals and early access to sales, potentially reducing churn among subscribers.
- Diversify Payment Methods: Provide incentives or improvements for customers using less common payment methods to balance churn rates among different payment options.
- Personalized Engagement: Implement personalized engagement strategies for rare and occasional users to increase their loyalty and reduce churn.
- Income Disclosure Encouragement: Encourage customers to disclose their income by providing benefits or discounts, potentially reducing the churn rate among this group.
- Loyalty Program Optimization: Optimize the loyalty program based on customer preferences, ensuring it aligns with the interests of the majority to reduce churn.
- Communication for Recommendations: Improve communication and engagement with customers unlikely to recommend, addressing concerns and enhancing their experience to minimize churn.
- Competitor Analysis: Analyze the experiences of customers who attempted other portals to identify areas for improvement and implement measures to retain such customers.
- Marketing Campaigns: Leverage marketing campaigns and incentives to increase awareness of discounts, potentially reducing churn among customers who are currently unaware.
- Sentiment-Based Strategies: Tailor strategies for customers affected significantly, focusing on enhancing their experience and satisfaction to reduce churn in this group.
- This strategic plan outlines customer segmentation based on RFM scores and corresponding retention policies for an e-commerce platform. The segmentation includes:
  - **Top Customers (RFM score > 4.5):**
    - **Policy:** Offer exclusive loyalty programs and VIP treatment.
    - **Action:** Provide personalized offers, early access, and dedicated support. Acknowledge loyalty with special rewards.
  - **High-Value Customers (4.5 > RFM score > 4):**
    - **Policy:** Encourage continued spending and engagement.
    - **Action:** Offer tiered rewards, promotions, and personalized recommendations.
  - **Medium-Value Customers (4 > RFM score > 3):**
    - **Policy:** Incentivize increased frequency and spending.
    - **Action:** Send targeted promotions, implement achievable loyalty milestones, and offer bundle deals.

- **Low-Value Customers (3 > RFM score > 1.6):**
  - **Policy:** Nurture and convert into higher segments.
  - **Action:** Send re-engagement campaigns, offer incentives, and collect feedback to address issues.
- **Lost Customers (RFM score < 1.6):**
  - **Policy:** Attempt to re-engage and understand reasons for disengagement.
  - **Action:** Send win-back campaigns with special offers, request feedback, and consider personalized communication.

Conclusion

In conclusion, the analysis of customer churn based on a sample of 161 respondents has unveiled valuable insights that lay the groundwork for strategic retention initiatives. The study delves into diverse factors, ranging from demographic indicators like employment status and income disclosure to behavioral patterns such as usage frequency and spending habits. These findings not only shed light on the nuanced dynamics of customer engagement but also pave the way for targeted strategies to address specific pain points and capitalize on opportunities for satisfaction enhancement.

The identified correlations between churn rates and variables such as employment status, income disclosure, usage patterns, spending per purchase, subscription status, login device, usability ratings, and other key factors offer a detailed panorama of customer behavior. Armed with this knowledge, the recommended targeted retention strategies aim to address specific challenges and leverage opportunities across different customer segments.

The importance of suburb residents, singles, and Bachelor's degree holders as distinct segments requiring tailored attention underscores the need for a nuanced approach in customer engagement. Similarly, the emphasis on enhancing platform usability, incentivizing subscriptions, diversifying payment methods, and deploying personalized engagement strategies underscores a commitment to providing a seamless and rewarding customer experience.

With an accuracy of 88%, the random forest predictive model can predict the new cases as churned out or stays in the Amazon ecosystem. The model can fairly predict the new cases.

The proposed strategies are not only rooted in addressing churn but also in fostering loyalty and satisfaction. By acknowledging and responding to the unique needs and preferences of each segment, businesses can create a more personalized and engaging experience, ultimately contributing to long-term customer retention. From encouraging income disclosure to optimizing loyalty programs, these recommendations reflect a holistic understanding of customer dynamics.

In essence, the findings and recommendations presented here provide a roadmap for businesses to navigate the complex landscape of customer churn. Implementation of these strategies has the potential to not only mitigate churn rates but also elevate overall customer satisfaction. As businesses move forward, incorporating these insights into their retention strategies can contribute to a more resilient and loyal customer base, driving sustained success in a competitive market.

Bibliography

https://www.researchgate.net/publication/359739936_User_churn_model_in_e-commerce_retail
https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12319&context=theses
https://jati.sites.apiit.edu.my/files/2018/07/2017_Issue2_Paper3.pdf
https://www.geeksforgeeks.org/k-means-clustering-introduction/
https://medium.com/analytics-vidhya/univariate-bivariate-and-multivariate-analysis-8b4fc3d8202c
https://www.jmp.com/en_in/statistics-knowledge-portal/what-is-correlation.html
[1] Nigel Williams, Sebastian Zander, Grenville Armitage: A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. www.researchgate.net. (2006)
https://www.gartner.com/en/information-technology/glossary/predictive-modeling#:~:text=Predictive%20modeling%20is%20a%20commonly,to%20help%20predict%20future%20outcomes.
https://online.stat.psu.edu/stat462/node/207/#:~:text=For%20binary%20logistic%20regression%2C%20the,%3Dexp(X%CE%B2).&text=1%20%E2%88%92%20%CF%80%20%3D%20exp-,%E2%81%A1,that%20predictor%20and%20the%20response.
https://www.ibm.com/topics/logistic-regression
https://www.geeksforgeeks.org/understanding-hypothesis-testing/
https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
https://www.jmp.com/en_in/statistics-knowledge-portal/what-is-correlation.html#404f1893-ae56-43ed-b84c-f6c99f313eca

Annexure

Survey Questionnaire using Google form

Churn Prediction for Amazon

Dear Respondent,
You are invited to participate in the survey designed to study the churn prediction analysis for Amazon. Your responses will help us enhance customer satisfaction and ensure customer retention .*Your participation is entirely voluntary, and your responses will be kept confidential. Please take a few minutes to complete this survey. This survey is estimated to take approximately 5-6 minutes to complete.*Thank you for your time and contribution.

* Indicates required question

1. Email *

Basic demographical information
This section gives insight into respondents information.The data is private and shall be used anonymously to ensure your confidentiality

2. Name of the respondent *

3. Please select your age group under which you fall *

Mark only one oval.

- under 15
- 15-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65 and above

4. the gender with which you associate yourself *

Mark only one oval.

- ◯ Male
- ◯ Female
- ◯ Other: _____

5. Which area are you located in? *

Mark only one oval.

- ◯ City
- ◯ Suburb
- ◯ Rural area
- ◯ International(Outside India)

6. Marital status *

Mark only one oval.

- ◯ Married
- ◯ Single
- ◯ Divorced

7. Please select your highest current qualification *

Mark only one oval.

- ◯ Less than High School or equivalent
- ◯ High School degree or equivalent
- ◯ Bachelor's degree
- ◯ Master's degree
- ◯ Doctorate

8. Employment status *

Mark only one oval.

- ◯ Student
- ◯ Employed
- ◯ Self employed
- ◯ Unemployed

9. Annual Income (**in INR**) *

Mark only one oval.

- ◯ <5 lacs per annum
- ◯ 5-10 lacs per annum
- ◯ >10 lacs per annum
- ◯ I do not wish to disclose
- ◯ Not applicable

Purchase Behaviour

This section tries to capture your past purchase behaviour

10. How **frequently** do you use Amazon platform to make purchases? *

Mark only one oval.

- ◯ Very Frequently (Several times a week)
- ◯ Frequently (About once a week)
- ◯ Occasionally (A few times a month)
- ◯ Rarely (Once a month or less)
- ◯ Never (I don't use e-commerce platforms for purchases)

11. When was the **last time you made a purchase** on Amazon platform? *

Mark only one oval.

- ◯ Within the last week
- ◯ Within the last month
- ◯ Within the last three months
- ◯ Within the last six months
- ◯ More than six months ago
- ◯ I have never made a purchase on an e-commerce platform

12. On average, how much do you **spend per purchase** using Amazon portal? *
*(**inclusive of both** website and application)*

Mark only one oval.

- ◯ less than 500 INR
- ◯ 500 INR - 1,000 INR
- ◯ 1,000 INR - 5,000 INR
- ◯ 5,000 INR - 10,000 INR
- ◯ > 10,000 INR
- ◯ I have never made any purchase

13. Approximately, how much have you spent on the Amazon portal **in the past 12 months**? *
*(Please give **numeric input, in INR**)*

_____

14. Have you subscribed to Amazon prime ? *

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ May consider in future
- ◯ Not interested

15. **For how many years** have you been an active subscriber of the Amazon? *
*(Please **input 0 for no active subscription** and **1 for one year or less than an year of subscription**.Kindly input **numeric value**)*

_____

16. If not for Amazon, which other portals do you prefer for online purchases? *
*(You **can select multiple portals, if applicable**)*

Tick all that apply.

- ☐ Flipkart
- ☐ Myntra
- ☐ Ajio
- ☐ Meesho
- ☐ Snapdeal
- ☐ Paytm Mall
- ☐ Jabong
- ☐ BigBasket
- ☐ Grofers
- ☐ Nykaa
- ☐ Pepperfry
- ☐ FirstCry
- ☐ None of the above

17. Your preferred mode of login device *

Mark only one oval.

- ◯ Mobile Phone
- ◯ Computer
- ◯ Laptop
- ◯ Tablet

18. Which categories do you purchase the most, while ordering via the Amazon platform? *
*(Note : You can **select multiple categories, if applicable**)*

Tick all that apply.

- ☐ Electronics and Gadgets
- ☐ Fashion and Apparel
- ☐ Home and Kitchen
- ☐ Books and Media
- ☐ Health and Personal Care
- ☐ Toys and Entertainment
- ☐ Automotive and Tools
- ☐ Groceries and Food
- ☐ Other

19. Your preferred mode of payment while making purchase *

Mark only one oval.

- ◯ Debit Card
- ◯ Credit Card
- ◯ E wallet
- ◯ Cash on Delivery
- ◯ UPI

20. How many devices are registered using a particular Amazon account? *

Mark only one oval.

- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5
- ◯ more than 5

21. Approximate duration of time spent on the Amazon page/App **on a weekly basis**?
*(Please provide **numeric input**, **In hours**)*

_____

22. Number of addresses saved against one Amazon account *
*(Kindly provide the **count** in **numeric value**)*

_____

23. Number of coupon used/ redeemed **in the past one year**? *
*(Please provide **numeric input**)*

_____

24. How much cash back have you received **in past one year**? *
*(Please provide a **numeric input in INR**)*

_____

25. Approximate **number of days** since the last order *
*(Please provide **numeric input**)*

_____

Consumer Satisfaction

This section tries to capture the satisfaction metrics of the customer

26. On a scale of 1 to 5, how satisfied are you with Amazon platform? *
*(Where **1** stands for **extremely dissatisfied** , **3** for **neutral experience** and **5** for **extremely satisfied**.)*

Mark only one oval.

	1	2	3	4	5	
Extr	◯	◯	◯	◯	◯	Extremely satisfied

27. How would you rate the usability of Amazon portal? *

Mark only one oval.

- ◯ Very Poor
- ◯ Poor
- ◯ Neutral
- ◯ Good
- ◯ Excellent

28. Are you satisfied with Amazon's customer support service? *

Mark only one oval.

- ◯ Very Dissatisfied
- ◯ Dissatisfied
- ◯ Neutral
- ◯ Satisfied
- ◯ Very Satisfied

29. How satisfied are you with the delivery time of your orders? *

Mark only one oval.

- ◯ Very Dissatisfied
- ◯ Dissatisfied
- ◯ Neutral
- ◯ Satisfied
- ◯ Very Satisfied

30. Number of complaints raised on the Amazon portal ? *
*(Within a **span of one year**. Please provide **numeric input**)*

_____

31. Which portal do you prefer over the other? *
*(Amazon website **or** the app)*

Mark only one oval.

- ◯ I prefer website interface over the application
- ◯ I prefer using application over website
- ◯ I have no preference

32. Where did you **first hear** about the portal? *

Mark only one oval.

- ◯ Social Media (Specify the platform, e.g., Facebook, Twitter, Instagram)
- ◯ Online Search (e.g., Google or other search engines)
- ◯ Word of Mouth (e.g., from a friend or family member)
- ◯ Online Advertisement (e.g., banner ads, display ads)
- ◯ Email Newsletter
- ◯ Other Website (Specify the website)
- ◯ Print Advertising (e.g., magazine or newspaper)
- ◯ Event or Conference
- ◯ Other: _____

Sentiments towards the portal

This section deals with the sentiment of the user towards the portal. It also touches upon future consideration regarding the portal.

33. Which are the most appealing features of the Amazon portal? *
*(Note : You can **select multiple attributes, if applicable**)*

Tick all that apply.

- ☐ User-Friendly Interface:The easy-to-use and intuitive design.
- ☐ Content Quality: The high-quality content and information provided.
- ☐ Personalization: The ability to customize the experience to my preferences.
- ☐ Speed and Performance: The portal's fast loading times and performance.
- ☐ Customer Support: The helpfulness and responsiveness of customer support.
- ☐ Features and Functionality: The range of features and tools available.
- ☐ Pricing and Value: The affordability and value for the price.
- ☐ Community and Interaction: The ability to engage with a community of users.

34. What is the main reason to subscribe to prime membership? *

Mark only one oval.

- ◯ Free fast delivery options
- ◯ Video streaming
- ◯ Early access to sale and slashed prices
- ◯ Exclusive Deals and Discounts
- ◯ Prime Music (music streaming)
- ◯ Kindle Owners' Lending Library (free e-book borrowing)
- ◯ Prime Wardrobe (try before you buy)
- ◯ Prime Reading (free books and magazines)
- ◯ Not Applicable, I am not a prime member

35. How likely are you to recommend Amazon prime to others? *

Mark only one oval.

- ◯ Very Likely
- ◯ Likely
- ◯ Neutral
- ◯ Unlikely
- ◯ Very Unlikely
- ◯ Not Applicable, I am not a prime member

36. How likely are you to continue using Amazon portal in the **next 12 months**? *

Mark only one oval.

	1	2	3	4	5	
Very	◯	◯	◯	◯	◯	Very Likely

37. Have you considered or attempted to use other e-commerce websites **in the last 12 months**? *

Mark only one oval.

- Yes
- No

38. If yes, what were the reasons for considering other e-commerce websites?

39. Are there any specific improvements or features you would like to see on our website to enhance your experience?

40. Are you aware of Amazon's promotional offers and discounts? *

Mark only one oval.

- Yes
- No

41. Would you be willing to participate in Amazon loyalty program to earn rewards and discounts? *

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Maybe

42. To what extent do Amazon's promotions influence your purchase decisions? *

Mark only one oval.

- ◯ Not at all
- ◯ Slightly
- ◯ Moderately
- ◯ Significantly