Advanced Churn Analysis in E-Commerce: Harnessing Hybrid Machine Learning and Deep Learning Models

A PROJECT REPORT

Submitted by : Fathima Noorudheen - CB.SC.I5DAS20109

in partial fulfillment of the requirements for the award of the

degree of

INTEGRATED MASTER OF SCIENCE

IN

DATA SCIENCE

AMRITA VISHWA VIDYAPEETHAM

epartment of Mathematics

AMRITA SCHOOL OF PHYSICAL SCIENCES

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE 641112

MAY 2024

Contents

INTRODUCTION

E-commerce is rapidly expanding in today's financial landscape . It triggered a paradigm change that impacted both marketers and consumers. The COVID-19 crisis has also accelerated the expansion of online retailing . Customers can now purchase a wide range of goods from their own homes, and firms can continue to operate despite the constraints. Furthermore, the world is changing rapidly, and businesses are using technology to keep up with market needs. Corporate businesses strive to meet client needs while also generating a profit from their investments. Therefore, understanding the future trend plays a significant role.

Customer churn, or attrition, is one of the most critical concerns for any firm that directly sells or services customers. It is critical for telecom service providers, eCommerce, and SaaS enterprises to watch and assess how many consumers leave and how many continue with the platform, as well as the reasons for each. Understanding consumer behaviour can considerably improve decision-making processes and assist reduce churn to increase profitability.

It is feasible to spend five to twenty-five times more on acquiring new clients than on keeping existing ones. Increasing client retention by 5% over time can boost profitability by 25% to 95%. To anticipate customer attrition, I will create a machine learning model.

METHODS

We have performed a detailed churn analysis on the dataset(’ **E Commerce Dataset.xlsx’) where the data set belongs to a leading online E-Commerce company. An online retail (E commerce) company wants to know the customers who are going to churn, so accordingly they can approach customer to offer some promos. We have performed the process of developing a churn prediction model for the dataset using machine learning algorithms like Logistic Regression, SVM, Random Forest, XGBoost and Adaboost along with the deep learning sequential model with Dense, Embedding, and Flatten layers.**

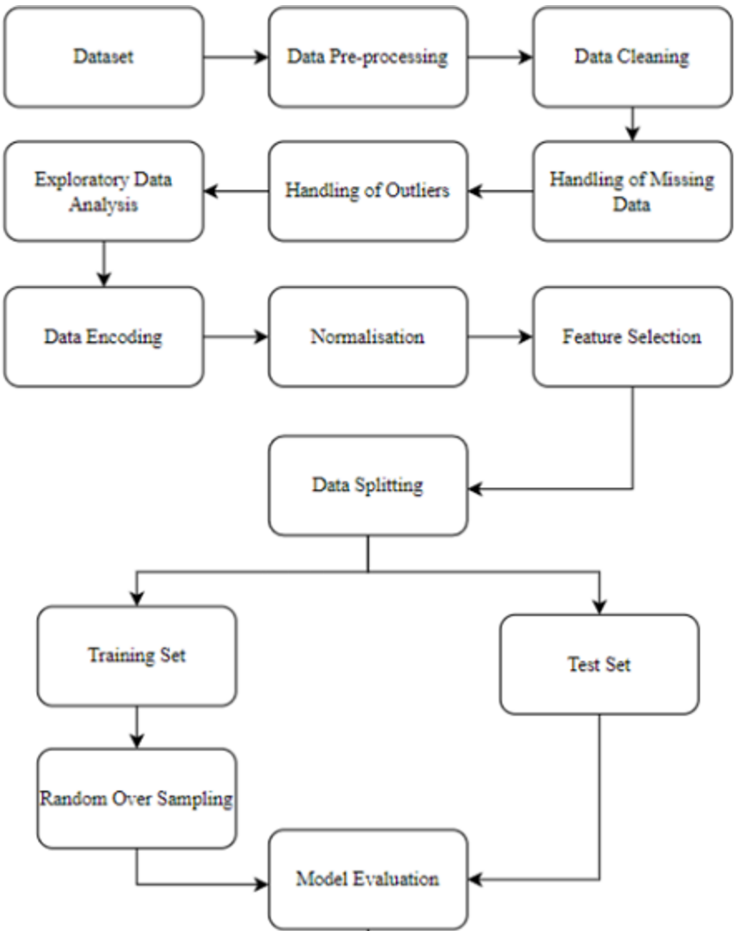
The goal is to build a predictive model that can accurately identify customers who are at risk of leaving the company (churn) based on the provided variables. This can help the company take proactive steps to retain these customers and reduce the rate of churn.

- Perform a thorough exploratory analysis of the provided customer data to gain insights into the behavior and characteristics of the customers. This includes analyzing patterns and trends in variables. This analysis can help the company understand its customers better and inform future decision-making.

In order to achieve this goal, the following steps will be taken:

- Data Preprocessing: The data will be cleaned, and any missing or incorrect values will be handled appropriately. This step is essential for ensuring the accuracy of the predictive model.
- Feature Engineering: New variables will be created based on the existing ones to add more depth to the analysis. Feature engineering can often improve the performance of the model.
- Model Development: Several machine learning and deep learning algorithms will be tested to find the most effective one for this specific task. The models will be trained using the processed data.
- Model Evaluation: The performance of the model will be evaluated using appropriate metrics. This will help determine the effectiveness of the model and identify any areas for improvement.
- Model Deployment: Once the model is finalized, it will be deployed to predict churn in real-time. The results can then be used to inform the company's customer retention strategies.

By following this approach, the company can leverage data to make informed decisions and take action to reduce customer churn. Ultimately, this could lead to increased customer loyalty and profitability.



UNDERSTANDING THE DATASET & PREPROCESSING

The dataset(' **E Commerce Dataset.xlsx')** **where the data set belongs to a leading online E-Commerce company. The features in the dataset include :**

- CustomerID: Unique customer ID
- Churn: Churn Flag
- Tenure: Tenure of customer in organization
- PreferredLoginDevice: Preferred login device of customer
- CityTier: City tier
- WarehouseToHome: Distance in between warehouse to home of customer
- PreferredPaymentMode: Preferred payment method of customer
- Gender: Gender of customer
- HourSpendOnApp: Number of hours spend on mobile application or website
- NumberOfDeviceRegistered: Total number of deceives is registered on particular customer
- PreferedOrderCat: Preferred order category of customer in last month
- SatisfactionScore: Satisfactory score of customer on service
- MaritalStatus: Marital status of customer
- NumberOfAddress: Total number of added added on particular customer
- OrderAmountHikeFromlastYear: Percentage increases in order from last year
- CouponUsed: Total number of coupon has been used in last month
- OrderCount: Total number of orders has been places in last month
- DaySinceLastOrder: Day Since last order by customer
- CashbackAmount: Average cashback in last month

Dataset Overview

Review the provided customer data to become acquainted with the variables and their structure.

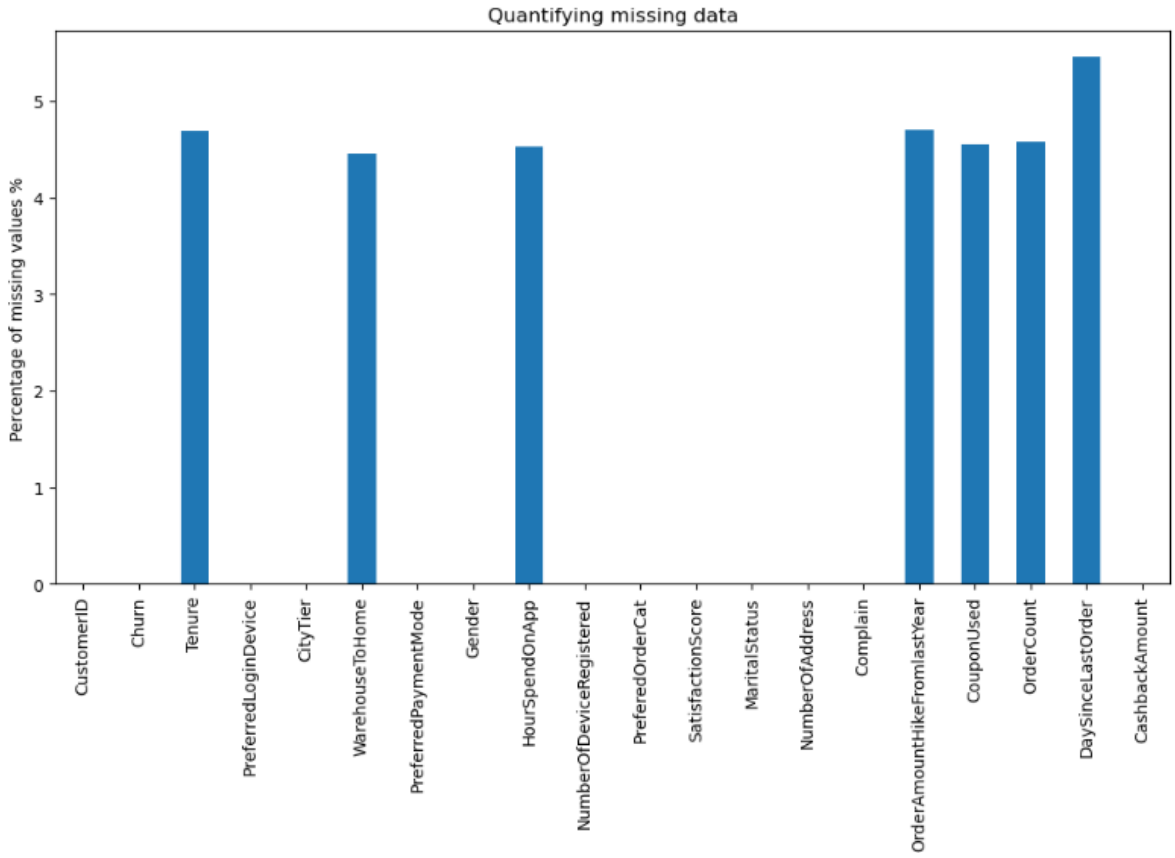
Check the data for quality, missing numbers, and probable errors.

Determine whether data pre-processing is necessary.

--------------------Data description----------------------------

	count	mean	std	min	25%	50%	75%	max
Churn	5630.0	0.168384	0.374240	0.0	0.00	0.00	0.0000	1.00
Tenure	5366.0	10.189899	8.557241	0.0	2.00	9.00	16.0000	61.00
PreferredLoginDevice	5630.0	0.928419	0.709822	0.0	0.00	1.00	1.0000	2.00
CityTier	5630.0	1.654707	0.915389	1.0	1.00	1.00	3.0000	3.00
WarehouseToHome	5379.0	15.639896	8.531475	5.0	9.00	14.00	20.0000	127.00
PreferredPaymentMode	5630.0	3.548135	1.389659	0.0	3.00	4.00	4.0000	6.00
Gender	5630.0	0.601066	0.489723	0.0	0.00	1.00	1.0000	1.00
HourSpendOnApp	5375.0	2.931535	0.721926	0.0	2.00	3.00	3.0000	5.00
NumberOfDeviceRegistered	5630.0	3.688988	1.023999	1.0	3.00	4.00	4.0000	6.00
PreferedOrderCat	5630.0	2.369627	1.411435	0.0	2.00	2.00	4.0000	5.00
SatisfactionScore	5630.0	3.066785	1.380194	1.0	2.00	3.00	4.0000	5.00
MaritalStatus	5630.0	1.168384	0.664344	0.0	1.00	1.00	2.0000	2.00
NumberOfAddress	5630.0	4.214032	2.583586	1.0	2.00	3.00	6.0000	22.00
Complain	5630.0	0.284902	0.451408	0.0	0.00	0.00	1.0000	1.00
OrderAmountHikeFromlastYear	5365.0	15.707922	3.675485	11.0	13.00	15.00	18.0000	26.00
CouponUsed	5374.0	1.751023	1.894621	0.0	1.00	1.00	2.0000	16.00
OrderCount	5372.0	3.008004	2.939680	1.0	1.00	2.00	3.0000	16.00
DaySinceLastOrder	5323.0	4.543491	3.654433	0.0	2.00	3.00	7.0000	46.00
CashbackAmount	5630.0	177.223030	49.207036	0.0	145.77	163.28	196.3925	324.99

The process of handling missing values is a critical component in data preprocessing. It is important not only to identify them but also to thoroughly analyze the missing values in the data. This is because they can greatly affect the results of your data analysis. Therefore, understanding why the data is missing and determining the best method to handle these missing values can significantly improve the overall quality of the dataset and lead to more accurate analyses and predictions.

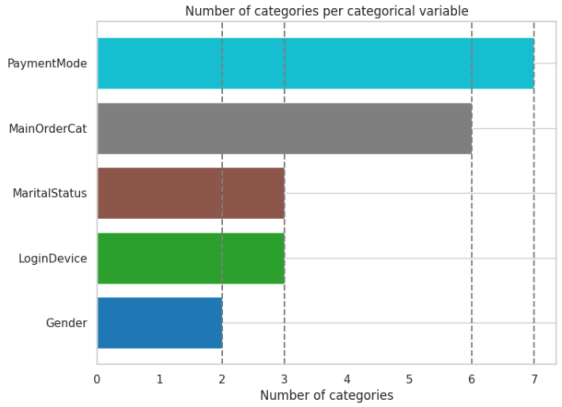


EXPLORATORY DATA ANALYSIS

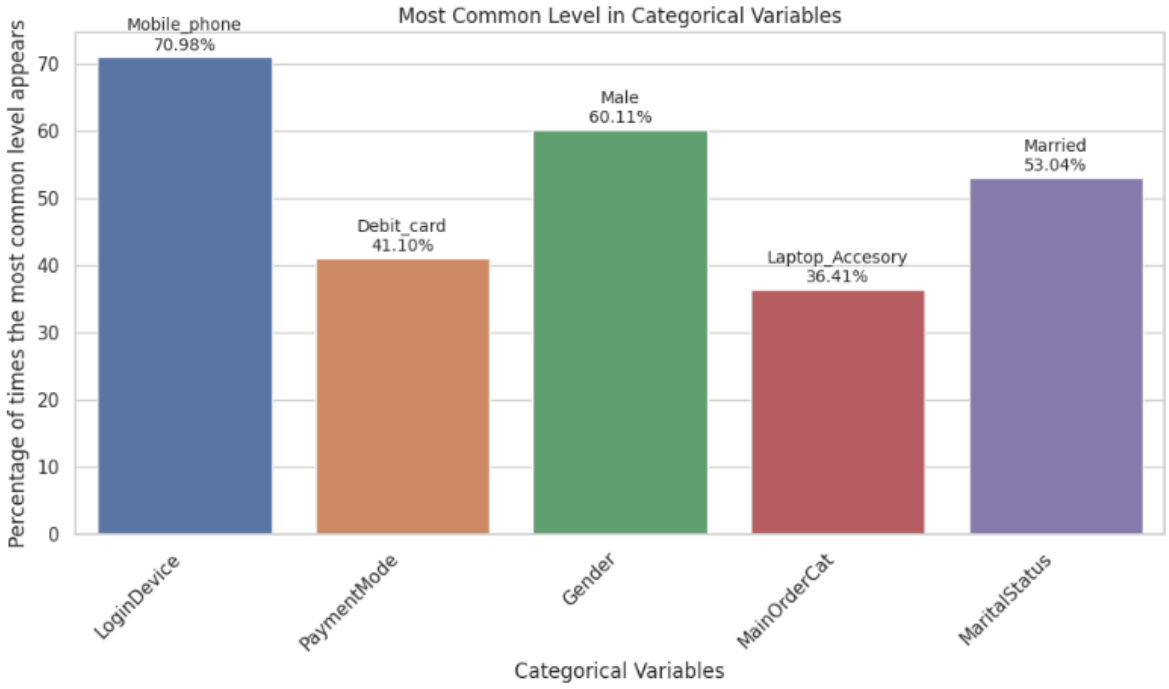
In the exploratory data analysis, we will examine the relationships between the different variables in the dataset. We will use various statistical techniques and data visualization methods to better understand the patterns and trends in the data. This will provide valuable insights that can guide our feature selection and model development process.

IDENTIFY VARIABLES WITH A LARGER NUMBER OF CATEGORIES COMPARED TO OTHERS.

This graph depicts the distribution of categories among categorical variables in your dataset, making it easier to find variables with more categories than others. It can help you comprehend the diversity and granularity of the categorical data in your dataset.



MOST COMMON LEVEL IN CATEGORICAL VARIABLES



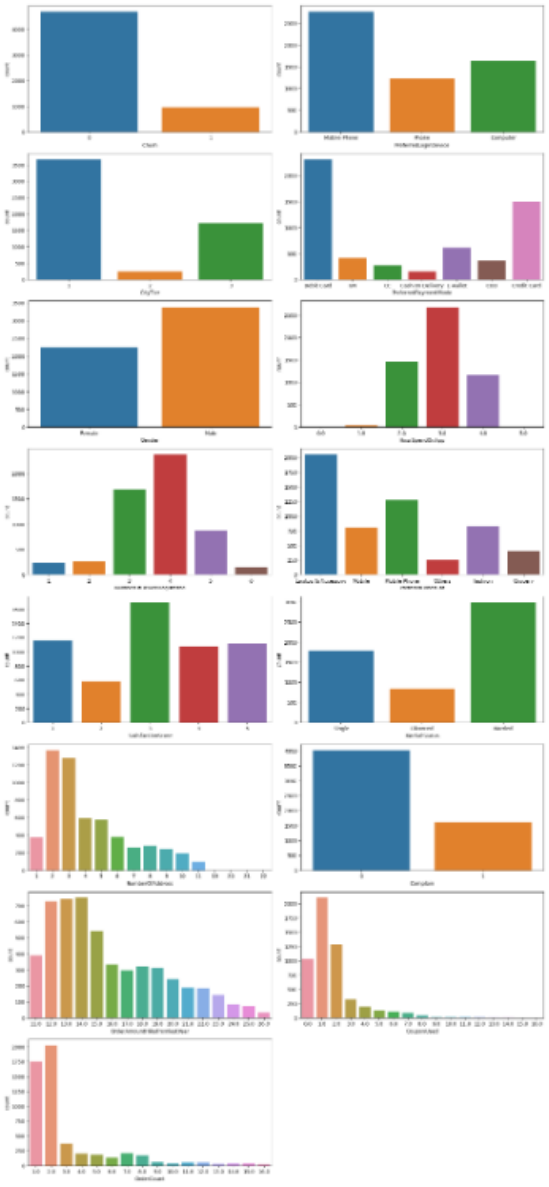
This graph indicates the frequency of the most common level for each categorical variable in the dataset. Each bar signifies a categorical variable, and its height represents the prevalence of the most common level. Taller bars indicate a higher proportion of the most common level. This helps identify variables with dominant levels and provides insights into the distribution of data. Variables with lower percentages may have more evenly distributed categories or a broader range of values.

FEATURE COUNT

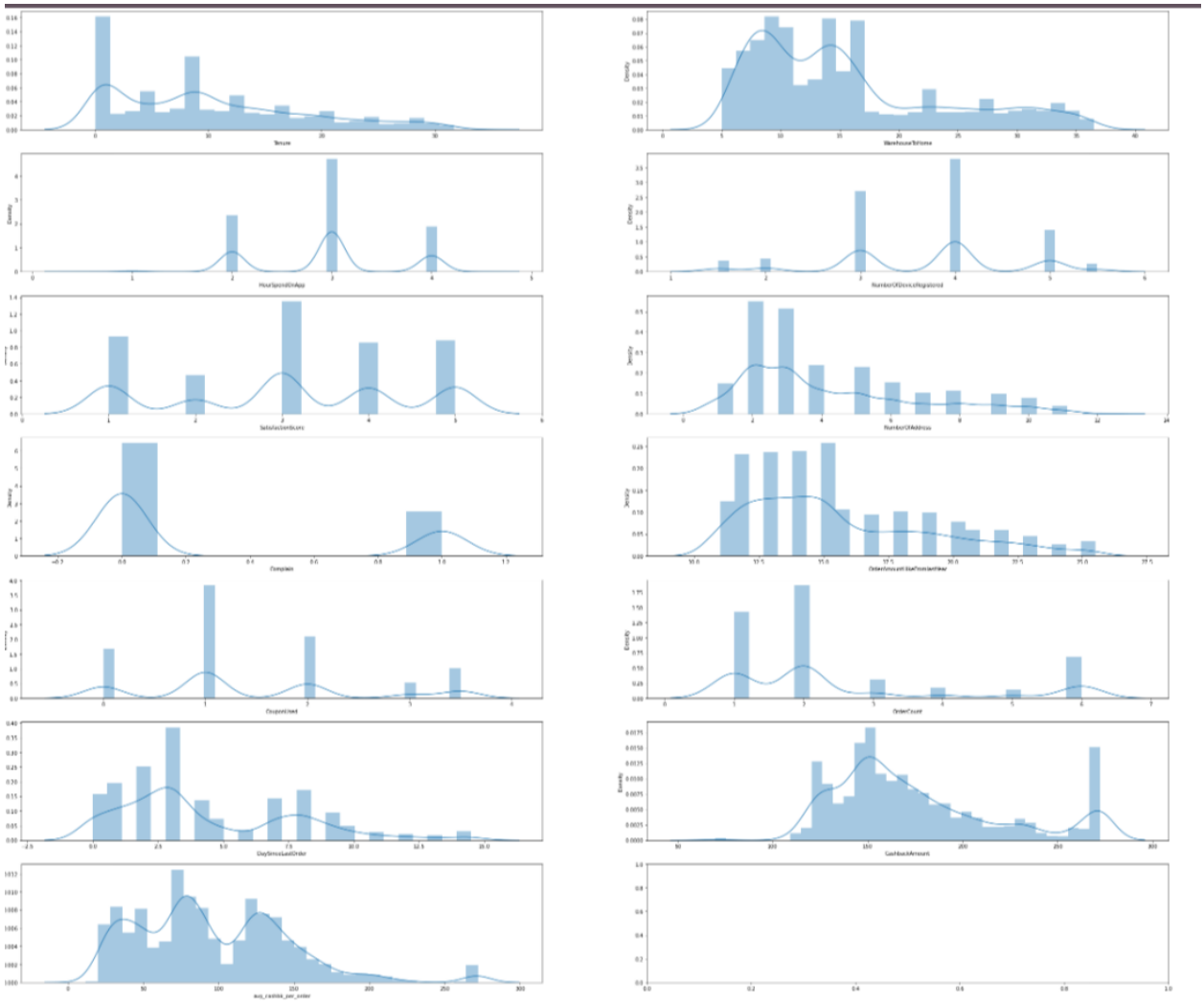
The "Count of All the Features" is a full inventory of all variables found in the dataset. This encompasses both the dependent and independent variables used in the study. In data science, a 'feature' is a specific quantitative aspect or characteristic of the phenomenon being studied.

Features of this E-commerce project's dataset could include information such as customer demographics, purchase history, product details, and more. A thorough count and analysis of these data is critical for determining which features are relevant and can contribute to the predictive model.

Furthermore, this deep study of features entails understanding the nature of the features - whether they are continuous or categorical, checking for missing values, and comprehending the distribution of values. This procedure is necessary for data preparation and feature engineering, which will ultimately affect the accuracy and efficiency of machine learning or deep learning models.



DISTIRBUITION OF FEATURES



The density distribution graph provides a visual representation of the distribution of all the features in the dataset. It is a crucial tool in exploratory data analysis as it allows us to understand the underlying structure of the data, identify outliers, and determine the type of distribution (normal, uniform, etc.).

The density distribution graph is carefully analysed and visualised. Each curve in this graph represents a feature from the dataset. The x-axis displays the range of each feature, while the y-axis depicts the density of data points within that range. Peaks on the graph show the most common values for each attribute, whilst valleys represent less common values. By analysing these graphs, we may obtain insight into the spread and skews of our data for each feature, which can be extremely useful when deciding on additional preprocessing steps or selecting appropriate models for analysis. If the feature is significantly skewed, we may want to use a transformation

to make it more normally distributed, which could improve the performance of some machine learning models.

CHURN ANALYIS

What is churn analysis?

Customer churn refers to the rate at which customers quit a platform or service. Customer churn analysis is a way for analysing the rate. There are typically two types of churn.

Voluntary Churn occurs when a client deliberately decides not to subscribe anymore, for example, because they found a better bargain elsewhere or had a negative experience.

Involuntary Churn occurs when a customer departs the platform involuntarily, such as when a payment fails because their credit card has been maxed out.

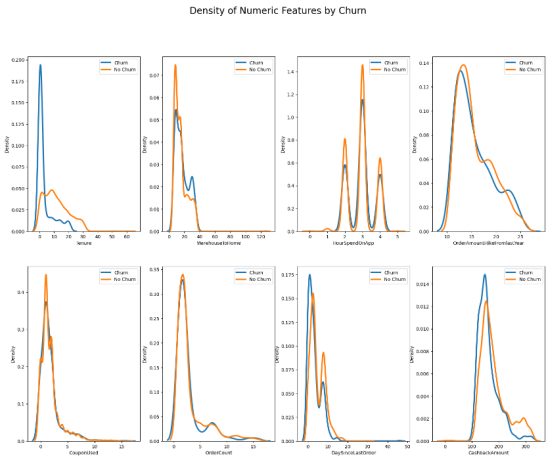
Some churn is to be expected, but severe turnover can have a significant impact on the company's bottom line. It also displays the company's net consumer perception, which is critical to the company's long-term growth and sustainability. It is rather obvious to determine whether or not there is considerable churn, but until we have crucial data points from which to derive meaningful knowledge, it is all guesswork. Several factors drive churn, and understanding them can help us figure out what to do next.

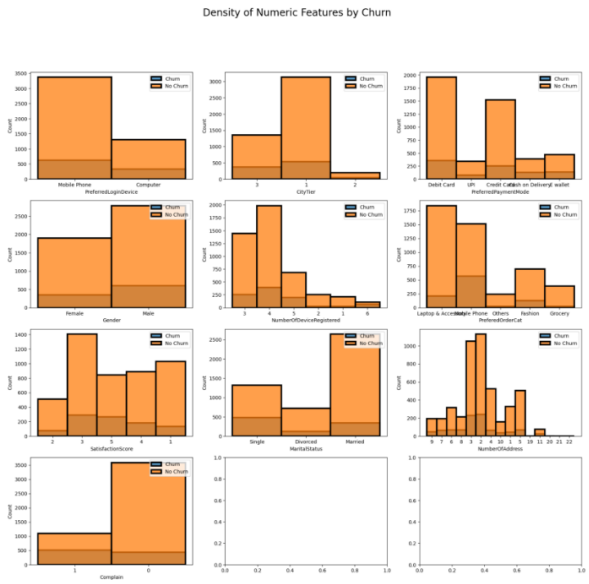
1. Defining Churn: Identifying what constitutes churn in an e-commerce business, which can vary depending on the business model and goals.
2. Identifying Churn Factors: Identifying the factors that contribute to churn, including customer behavior and demographics, satisfaction and experience, competitor actions, and external factors.
3. Analyzing Churn Trends: E-commerce organizations use churn trends to determine when and why customers leave. This includes customer segmentation, time-series analysis, and cohort analysis.
4. Predicting Churn: Utilizing predictive modeling techniques, such as machine learning algorithms, to predict which customers are most likely to leave in the future.
5. Reducing Churn: Implementing strategies based on churn data to reduce churn and increase customer retention. This could include personalized marketing and communication, improved customer experience, loyalty programs, and ongoing monitoring and adaptation.

In summary, churn analysis in e-commerce is a continuous process that helps companies understand consumer behavior, identify potential churn risks, and implement proactive measures to retain customers and build long-lasting relationships.

Distributions Insights Of the Numeric Features by churn

Customer churn is influenced by various factors. Longer tenure, faster deliveries, increased app usage, having more registered devices, and high satisfaction scores all reduce churn rates. Moreover, customers with more registered addresses, those who used more coupons, and placed more orders show lower churn. However, more customer complaints and longer periods since the last order correlate with higher churn. Surprisingly, the churn rate doesn't vary significantly across different city tiers.



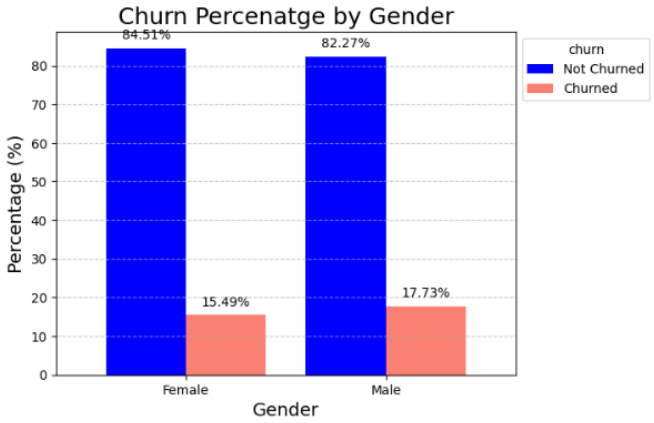


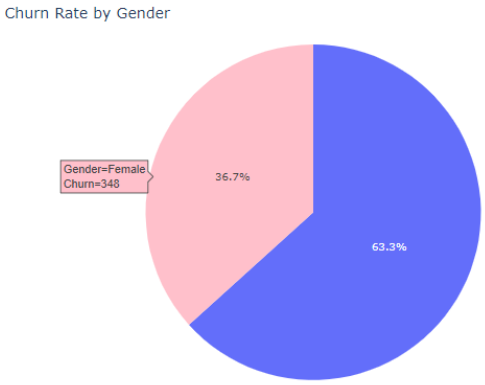
UNIVARIATE ANALYSIS - Analyzing the Churn by each Variable

The term Univariate Analysis refers to the analysis of a single variable. The goal is to investigate and identify patterns pertaining to a single variable.

CHURN BY GENDER

The Churn Rate by Gender refers to the percentage of customers, categorized by gender, who stop using a product or service over a given period of time. This metric is significant as it can help businesses understand if there's a difference in churn based on gender. Depending on the results, different customer retention strategies may be implemented for different genders. For instance, if one gender has a higher churn rate, the company might want to investigate the reasons behind this and devise strategies to address the issue and reduce the churn rate. Conversely, if there's a low churn rate for a certain gender, the company might want to analyze what they are doing right for that demographic and see if it can be applied to others.



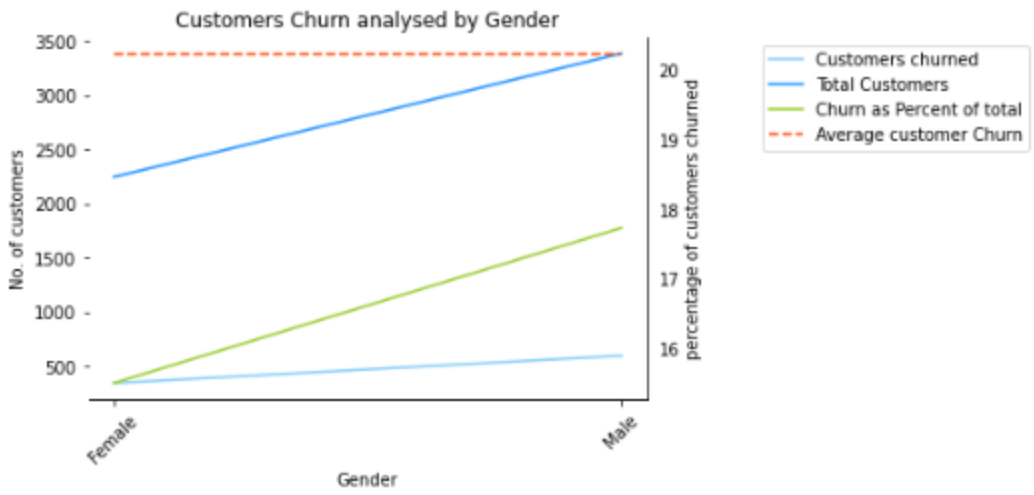


As we see the males are more likely to churn as we have 63.3 % churned males from the app may be the company should consider increasing the products that gap the males interest and so on.. we are going to see if there is another factors that makes the highest segment of churned customers are males.

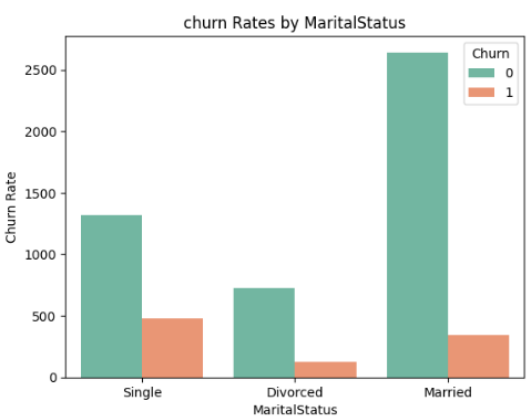
These percentages indicate that there are indeed gender-based patterns in customer churn:

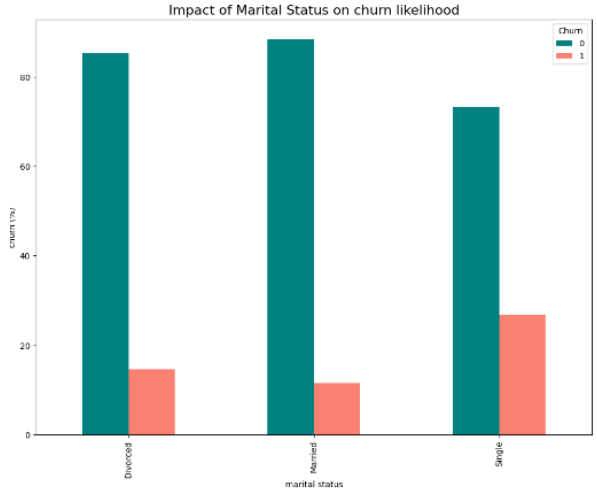
- Male customers have a slightly higher churn rate (17.73%) compared to female customers (15.49%).
- Female customers exhibit a higher retention rate (84.51%) than male customers (82.27%).

This suggests that gender may play a role in customer churn, with different churn rates observed between male and female customers. This insight can be important for developing gender-specific customer engagement and retention strategies.



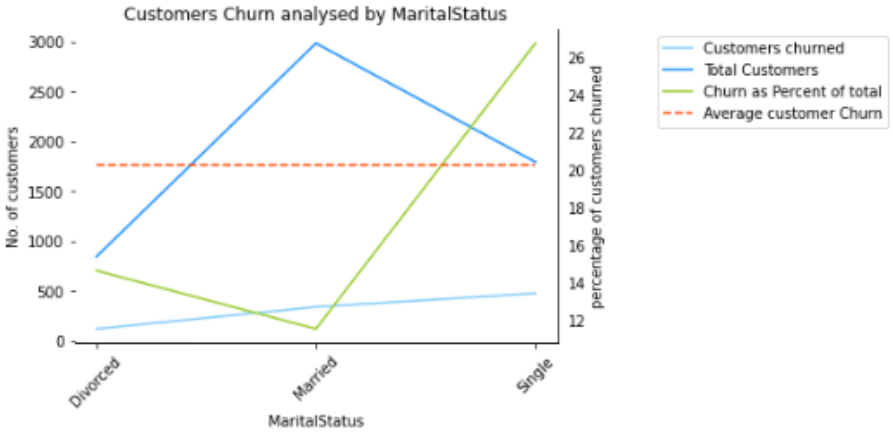
CHURN RATES BY MARITAL STATUS



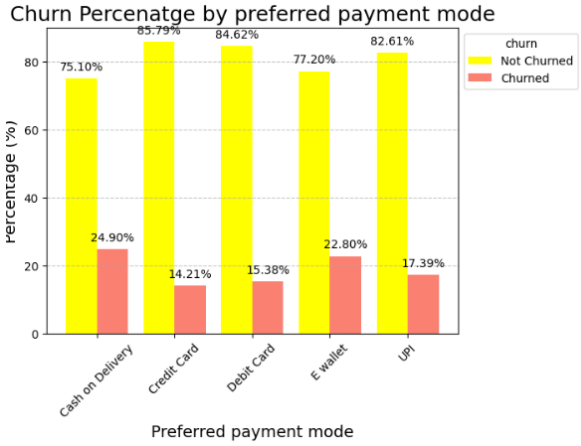


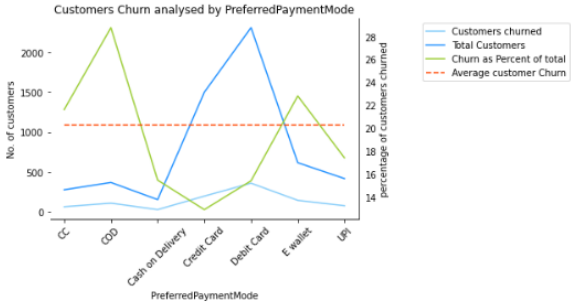
- Single customers have a significantly higher churn rate (26.73%) compared to married (11.52%) and divorced (14.62%) customers.
- Married customers exhibit the highest retention rate (88.48%).

This suggests that marital status is a relevant factor in customer churn, with different marital statuses showing distinct churn behaviors.



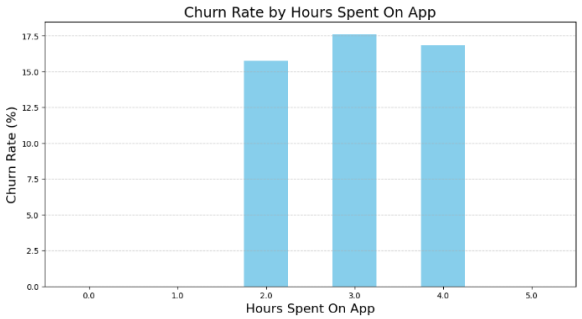
CHURN BY PREFERRED PAYMENT MODE

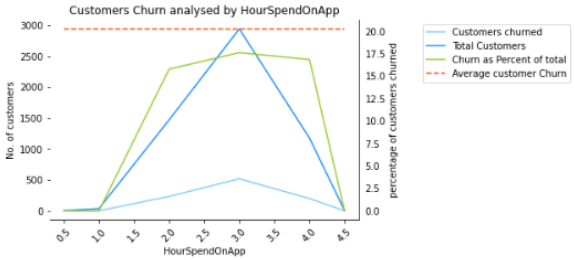




- Cash on Delivery users have the highest churn rate (24.90%), followed by E wallet users (22.80%), and UPI users (17.39%).
- Credit Card and Debit Card users exhibit lower churn rates, at 14.21% and 15.38%, respectively.

CHURN RATE BY HOURS SPENT ON APP



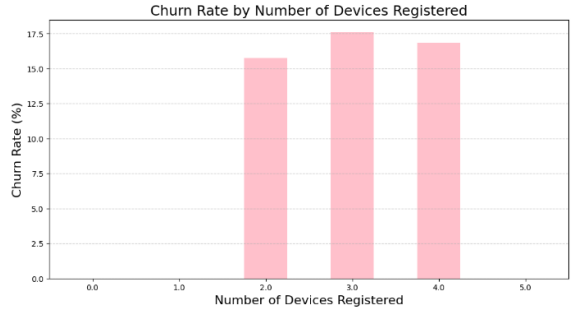


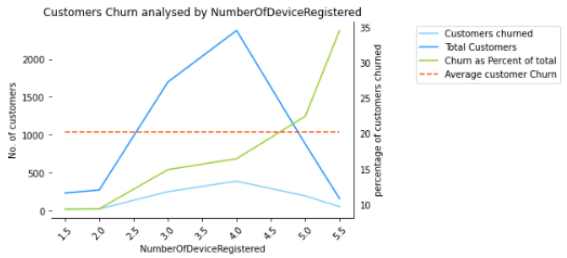
- 0, 1, and 5 Hours: There is a 0% churn rate for customers who spend 0, 1, and 5 hours on the app. This could indicate either a very small number of users in these categories or exceptionally high retention for these groups.
- 2 Hours: The churn rate is 15.77% for users who spend 2 hours on the app.
- 3 Hours: Users who spend 3 hours on the app have a slightly higher churn rate of 17.61%.
- 4 Hours: The churn rate for users spending 4 hours on the app is 16.84%.

These percentages suggest that there is a relationship between the time spent on the app and churn rate, with a noticeable increase in churn for those spending 2 to 4 hours on the app. It's important to consider that the 0% churn rate for 0, 1, and 5 hours might be due to specific user behavior in these groups.

This insight can be valuable for understanding user engagement and developing strategies to enhance app experiences, particularly for users in the 2–4 hour range where churn rates are higher.

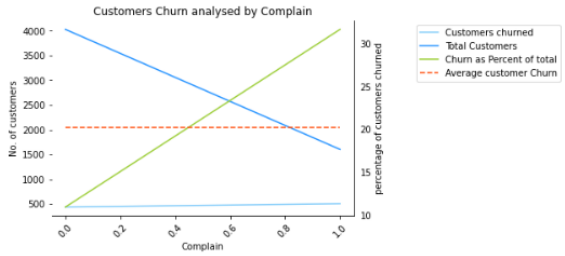
CHURN RATE BY NUMBER OF DEVICES REGISTERED

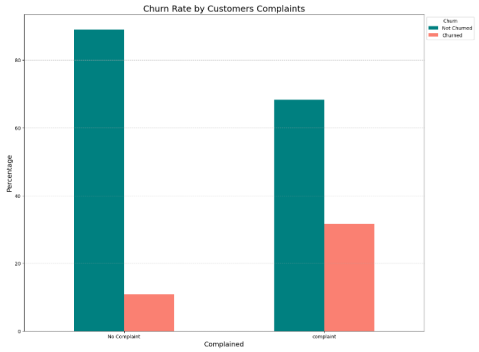




- 1 to 2 Devices: Customers with 1 or 2 devices registered have the lowest churn rates, at 9.36% and 9.42% respectively.
- 3 to 4 Devices: There is a noticeable increase in churn rates for customers with 3 or 4 devices, with rates of 14.95% and 16.49%.
- 5 to 6 Devices: The churn rate rises significantly for customers with 5 and 6 devices registered, reaching 22.47% and 34.57%.

CHURN RATE BY CUSTOMER COMPLAINTS



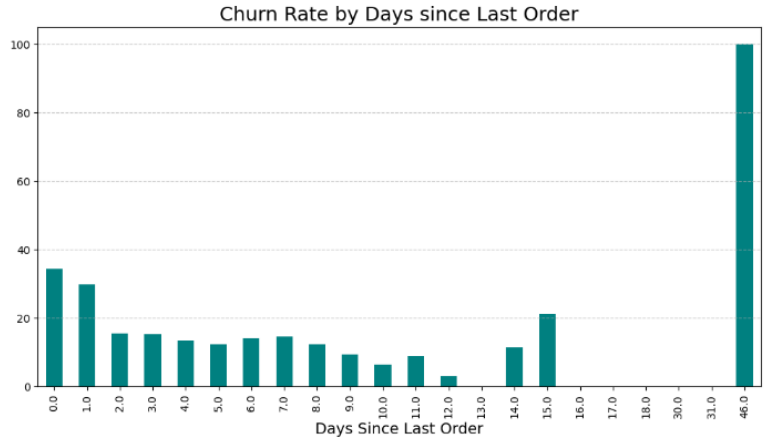


- Customers who have lodged complaints have a significantly higher churn rate (31.67%) compared to those who have not complained (10.93%).
- This suggests that customer complaints are a strong indicator of churn risk and highlight the importance of addressing customer issues effectively to improve retention.

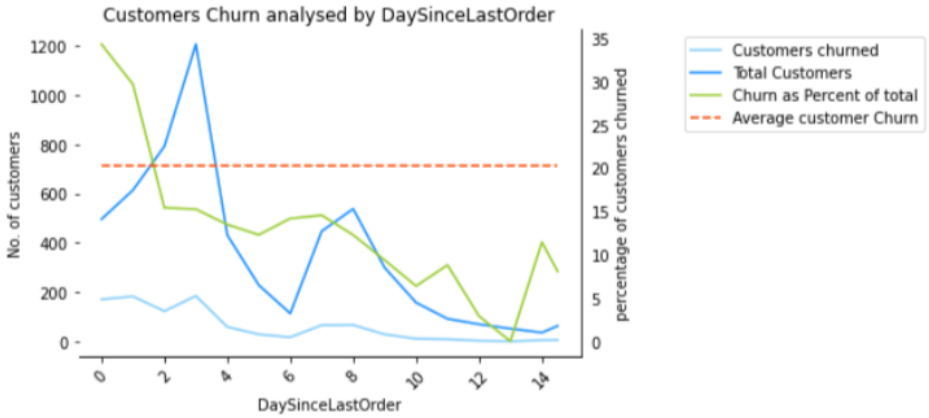
CHURN RATE BY DAYS SINCE LAST ORDER

- There is notable variation in churn rate based on the duration since the last order.

- High Churn Rates for Short durations: A significantly high churn rate is observed for customers who have just recently made an order (0 and 1 day ago) with rates of 34.27% and 29.64% respectively.
- Deceasing trend: the churn rate generally decrease as the duration since the last order increase reaching lower rates for durations of 4 to 10 days.
- variablity in longer durations: for durations longer than 10 days the churn rate shows variability with some days experiencing very low or zero churn rates. this couldd be influenced by a smaller sample size for these categories.
- anomaly at 46 days: A 100% churn rate is observed at 46 days since the last order which likely indicates an anomaly or very small number of customers in the category.



These observations suggest that the duration since a customer's last order can be predict of churn particularly in the immediate days following an order. the trend indicates that customers are more likely to churn shortly after placing an order with the likelihood decreasing as more time passes. the insight can inform strategies for engaging customers at critical times to reduce the likelihood of churn.

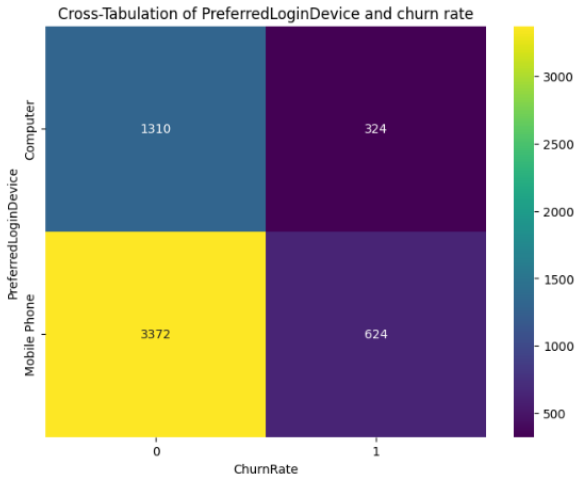


BIVARIATE ANALYSIS - Analyzing the Churn by multiple Variables.

In bivariate analysis, we will use two variables to determine their relationship. We'll utilise the correlation coefficient to determine the link between variables. A correlation number close to 1 suggests a positive correlation, a correlation near -1 shows a negative relationship, and a correlation around zero implies neutrality.

Relationship between the Preferred Login Device and Churn Rate

Churn	0	1
PreferredLoginDevice		
Computer	80.17	19.83
Mobile Phone	84.38	15.62

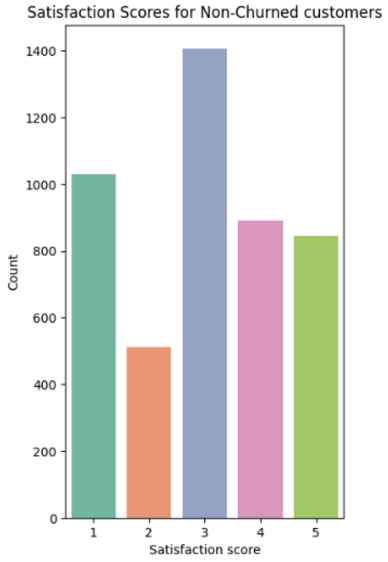


These percentages suggest that:

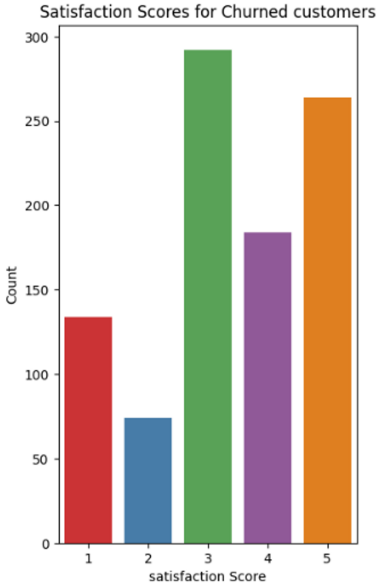
- Customers who prefer using a computer for login have a slightly higher churn rate (19.83%) compared to those who prefer mobile phones (15.62%).
- Mobile users have a higher retention rate.

This analysis reveals that there is indeed a relationship between the preferred login device and churn rate, with the type of device potentially influencing the likelihood of a customer churning.

Analyzing the significant difference in satisfaction scores between customers who churn and those who don't




```
(Churn
0    3.00
1    3.39
Name: SatisfactionScore, dtype: float64,
2.105157407388599e-15)
```



- Customers who did not churn (Churn = 0): The average satisfaction score is 3.00.
- Customers who churned (Churn = 1): The average satisfaction score is 3.39.

- Statistical Significance: The t-test for the difference in satisfaction scores between the two groups yields a p-value of approximately 2.11×10−152.11×10−15, which is significantly less than the conventional threshold of 0.05.

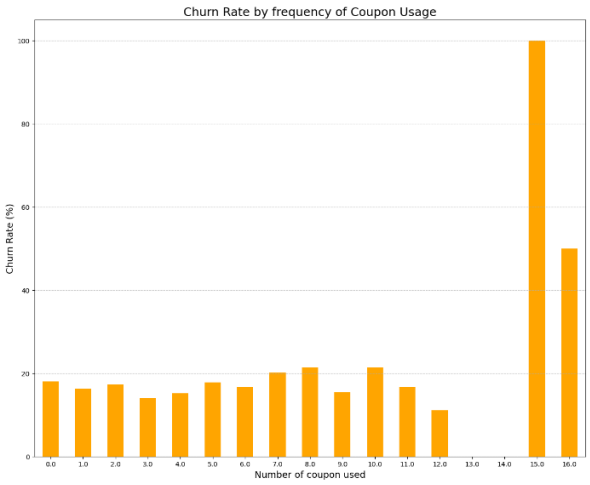
These results indicate that there is a statistically significant difference in the satisfaction scores between customers who churn and those who don’t. Surprisingly, the average satisfaction score is higher for customers who churned compared to those who didn’t. This might suggest that

factors other than satisfaction scores are influencing the decision to churn, or that the satisfaction scores may not fully capture the customer's experience or likelihood to remain with the service.

Relationship between the frequency of coupon usage and churn rate

- The churn rate varies with the number of coupons used, showing a non-linear relationship.

- Customers who used 0 to 2 coupons show churn rates fluctuating between approximately 16% and 18%.
- A notable decrease in churn rate is observed for customers who used 3 to 4 coupons, with rates around 14% to 15%.
- the churn rate increases again for customers using 5 to 8 coupons, reaching over 20% in some cases.

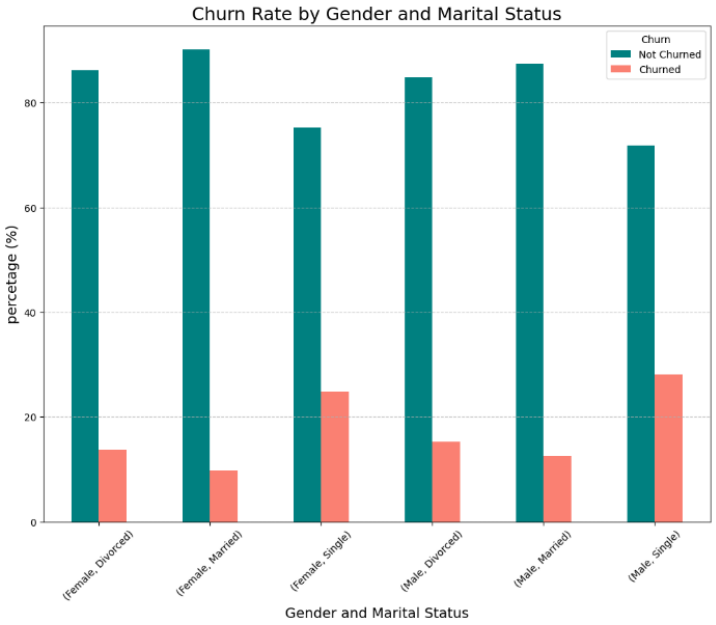


- Interestingly, for 12 to 14 coupons, the churn rate drops significantly, even to 0% for 13 and 14 coupons, which could be due to a small sample size in these categories.
- An exceptionally high churn rate (100%) is observed for customers who used 15 coupons, which might indicate an anomaly or a very small sample size for this group.

These observations suggest that while there is a relationship between coupon usage and churn, it is not straightforward. The churn rate does not consistently increase or decrease with the frequency of coupon usage, indicating that other factors might also play a significant role in determining churn.

Combination of gender and marital status affect churn rates

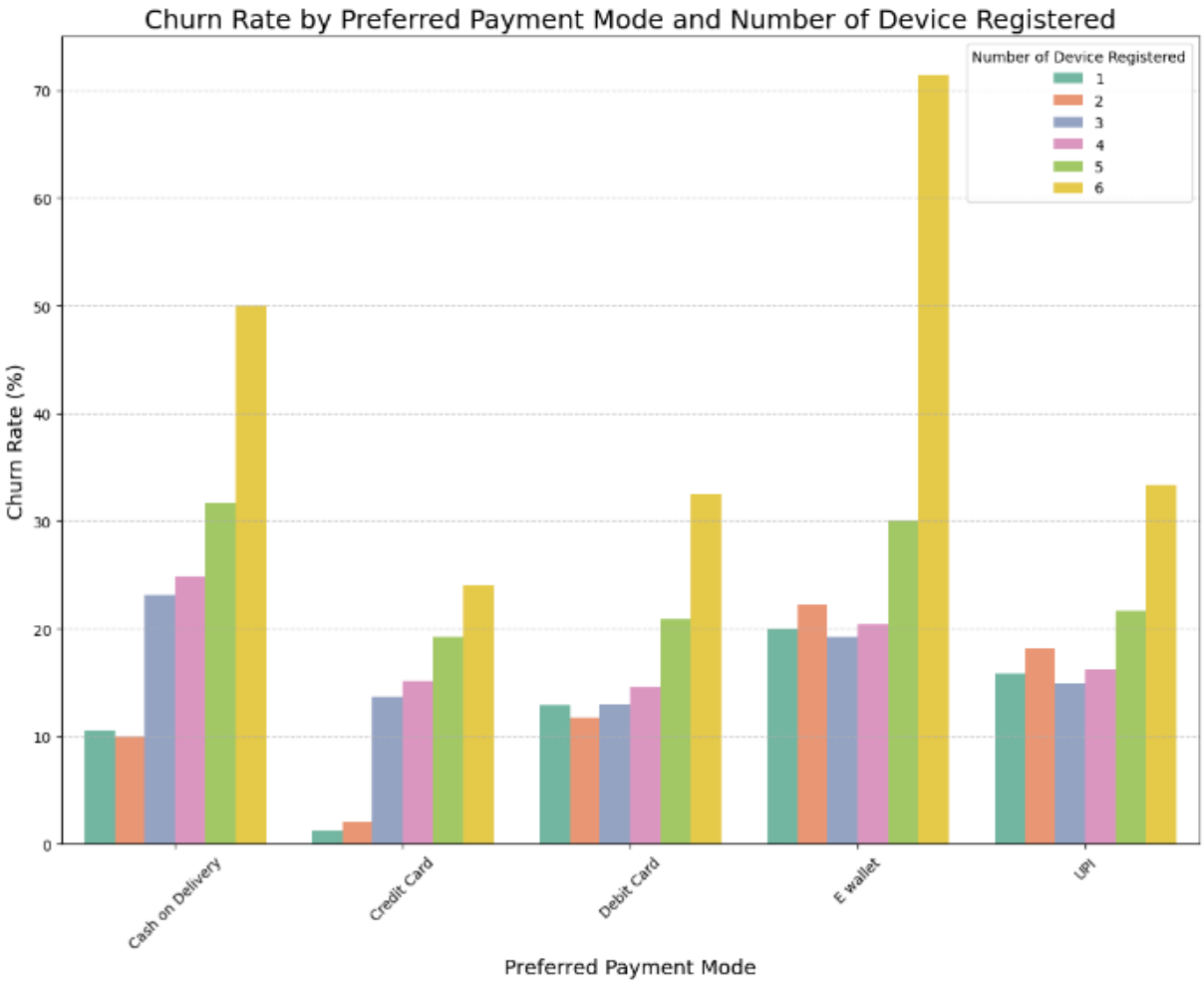
	Churn	0	1
Gender	MaritalStatus		
Female	Divorced	86.21	13.79
	Married	90.18	9.82
	Single	75.20	24.80
Male	Divorced	84.80	15.20
	Married	87.43	12.57
	Single	71.87	28.13



- Married female customers have the highest retention rate (90.18%) and the lowest churn rate (9.82%).
- Single male customers exhibit the highest churn rate (28.13%) among all groups.

- single customers (both male and female) have higher churn rates compared to married and divorced customers.
- There is a noticeable difference in churn rates between genders within the same marital status category, particularly among single customers.

The preferred payment mode combined with the number of devices registered impact churn



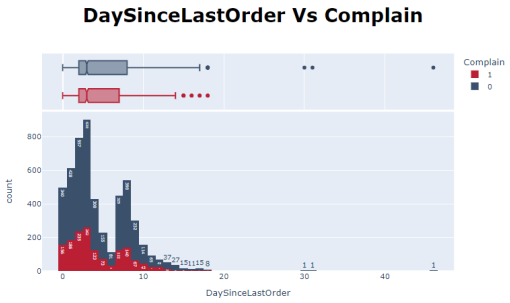
- Customer using the E wallet Payment method with 6 devices registered have the highest churn rate across all combinations.
- the next highest churn rate is observed for customers using cash on delivery with 6 device registered.
- the lowest churn rate is found amouing customers using a credit card ith only 1 device registered.

The duration since the last order and the number of complaints lodged have a combined effect on churn

- immediate post order period: On the same day as the last order(day 0),customer who have lodged a complaint show a significant higher churn rate (55.13 %) compared to those who have not complained (24.71 %).
- day following the order: In the days following an order (Day 1 to day 18) the churn rate for customers with complaints generally remains higher than for those without complaints.the difference is particularly notable in the first few days after the order.
- Extended Periods without order: for longer duration without orders (30 days and beyond) the churn rate drops to 0 % for customers without complaints interestingly a 100% churn rate is observed at 46 days since the last order but this could be due to very small sample size for this category.
- No complaints with Zero Churn: there are instance where the churn rate is 0% for customers without complaints (days 12,13,17,18,30 and 31).this suggests strong retention in these groups.
- High Churn at extended durations with complaints: for some extended duration (days15) the churn rate for customers with complaints is notable high (50.00 %). These finding indicate that the combination of recent interaction (last order) and customer disatisfaction (complaints) has a significant impact on churn. Immediate post order periods combined with

	Churn	0	1
DaySinceLastOrder	Complain		
0.0	0	75.29	24.71
	1	44.87	55.13
1.0	0	78.97	21.03
	1	50.54	49.46
2.0	0	91.38	8.62
	1	68.51	31.49
3.0	0	89.79	10.21
	1	72.17	27.83
4.0	0	91.56	8.44
	1	73.98	26.02
5.0	0	92.26	7.74
	1	78.08	21.92
6.0	0	87.65	12.35
	1	81.25	18.75
7.0	0	92.31	7.69
	1	67.21	32.79
8.0	0	93.47	6.53
	1	71.43	28.57
9.0	0	93.10	6.90
	1	82.09	17.91
10.0	0	94.74	5.26
	1	90.70	9.30
11.0	0	93.85	6.15
	1	84.62	15.38
12.0	0	100.00	0.00
	1	91.67	8.33
13.0	0	100.00	0.00
	1	100.00	0.00
14.0	0	92.59	7.41
	1	75.00	25.00
15.0	0	86.67	13.33
	1	50.00	50.00
16.0	0	100.00	0.00
	1	100.00	0.00
17.0	0	100.00	0.00
	1	100.00	0.00
18.0	0	100.00	0.00
	1	100.00	0.00
30.0	0	100.00	0.00
31.0	0	100.00	0.00
46.0	0	0.00	100.00

Relation between Complain and DaySinceLastOrder for churned customers



	DaySinceLastOrder
Complain	
0	1313.0
1	1580.0

customers who didn't made complain has higher DaySinceLastOrder , however it's only one customer so its an outlier if we remove it we will customers with no complain has lower DaySinceLastOrder

Statistical methods

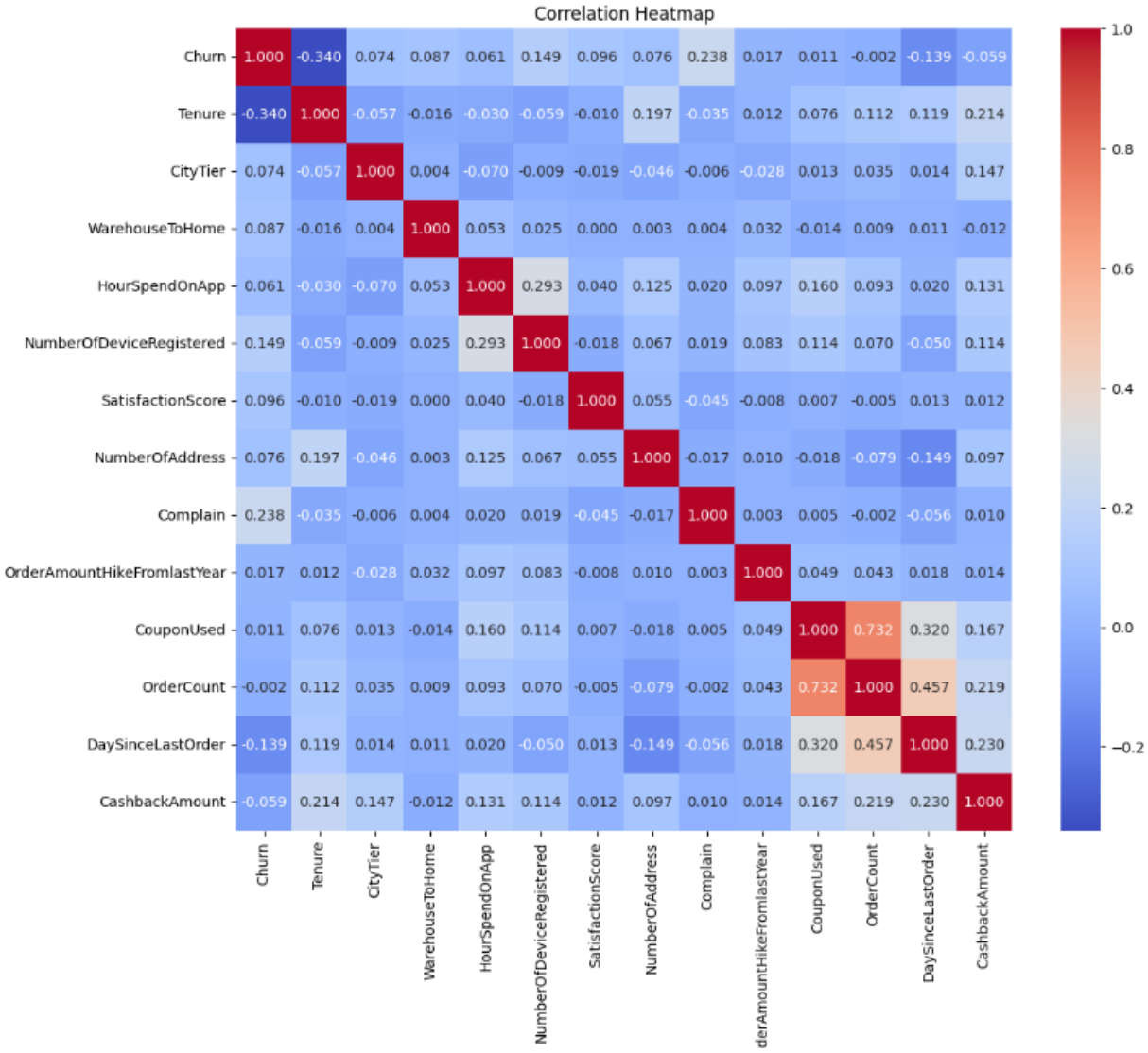
It refers to a range of techniques and procedures employed to analyze data, draw conclusions, and make informed predictions about future outcomes. These methods can be divided into two broad categories: descriptive statistics and inferential statistics.

Descriptive statistics summarize and organize data from a sample. Common techniques include measures of central tendency (mean, median, mode), measures of variability (range, variance, standard deviation), and measures of relationship (correlation, covariance).

Inferential statistics, on the other hand, enable predictions or inferences about a population based on a sample of data. It includes hypothesis testing, regression analysis, and analysis of variance (ANOVA).

Here we are using the statistical methods for exploratory data analysis (EDA) to gain insights into the E-commerce data and determine which machine learning or deep learning models to use for predicting customer churn and understand different relational analysis between the features.

Correlation Heatmap



The heatmap in this data analysis is used to indicate the correlation between different variables. Darker colors often represent stronger correlations. Positive values typically indicate a positive correlation, meaning as one variable increases, the other does as well. Negative values, on the other hand, indicate an inverse correlation: as one variable increases, the other decreases.

Positive correlation:

HoursOnApp and CustomerID show a correlation of 0.58, indicating that customers with higher IDs (most likely newer clients) spend more time on the app.

CouponUsed and OrderCount have a 0.64 connection, showing that customers who place more orders utilise more coupons. It suggests that promotions and discounts can help increase purchasing frequency.

CashbackAmount has a rather strong positive correlation with Churn (0.51), suggesting that consumers who receive more cashback may be more likely to churn. This suggests that, while cashback incentives are used, they may not be successful in long-term customer retention and may be connected with one-time transactions or customers who are not engaged beyond the transactional advantage.

The correlation between churn and tenure is -0.35, indicating that customers with longer tenure are less likely to churn, implying that efforts to keep consumers engaged over time may lower customer turnover.

Many variables have little to no correlation with one another, implying that there is no linear link or that any relationship is complex and cannot be captured just by correlation. This indicates that client retention initiatives should be diversified. While promotions may increase sales, they do not guarantee long-term commitment.

The data acquired from the relationship between the numerical factors appears to imply that it would be useful to focus on personalised engagement techniques that enhance tenure, as well as to re-evaluate the cashback incentive programmes to ensure that they are contributing positively to long-term customer loyalty.

Finding a correlation between SatisfactionScore and HourSpendOnApp?

The correlation between SatisfactionScore and HourSpendOnApp is an important aspect to consider for customer retention.

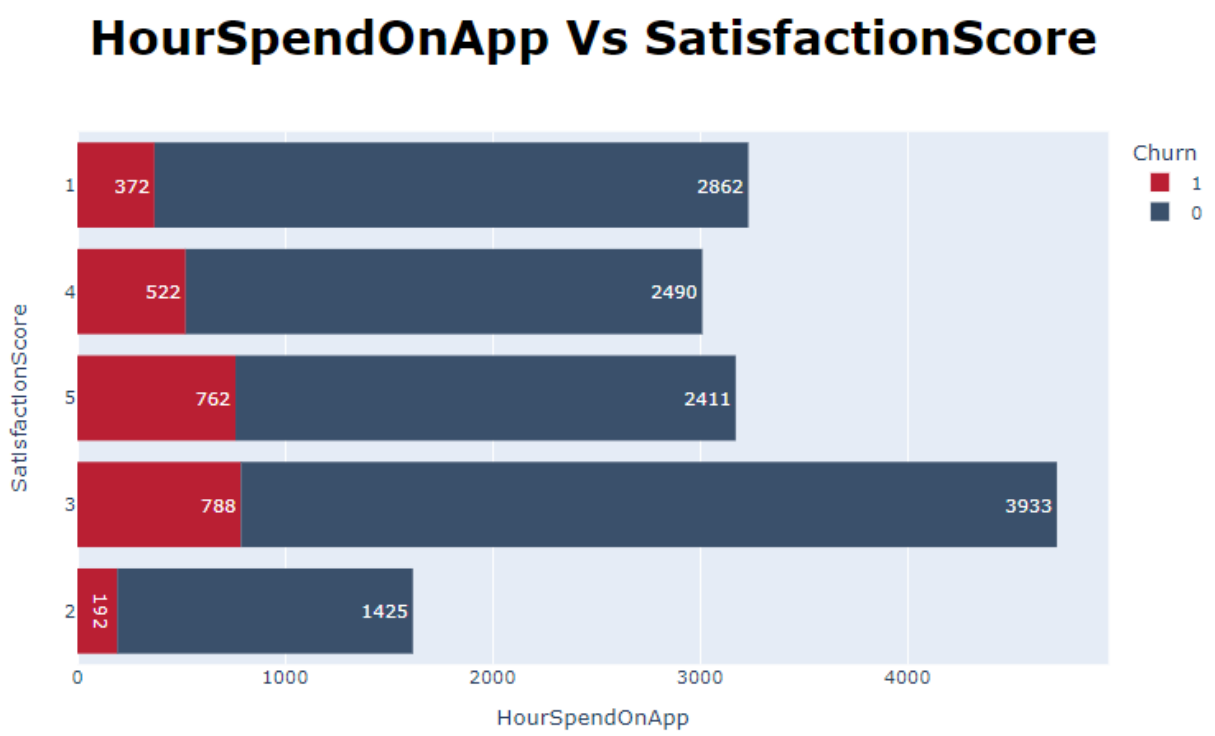
SatisfactionScore is a measure of how satisfied a customer is with the service or product. This could be determined through surveys or customer feedback. On the other hand, HourSpendOnApp refers to the amount of time a customer spends on the app.

If there is a positive correlation, it means that as the hours spent on the app increase, the customer satisfaction also increases. This could be due to the app's user-friendly nature, good customer service, or high-quality products.

On the contrary, if there is a negative correlation, it implies that as the hours spent on the app increase, the customer satisfaction decreases. This could be due to reasons such as poor app performance, unsatisfactory customer service, or low-quality products.

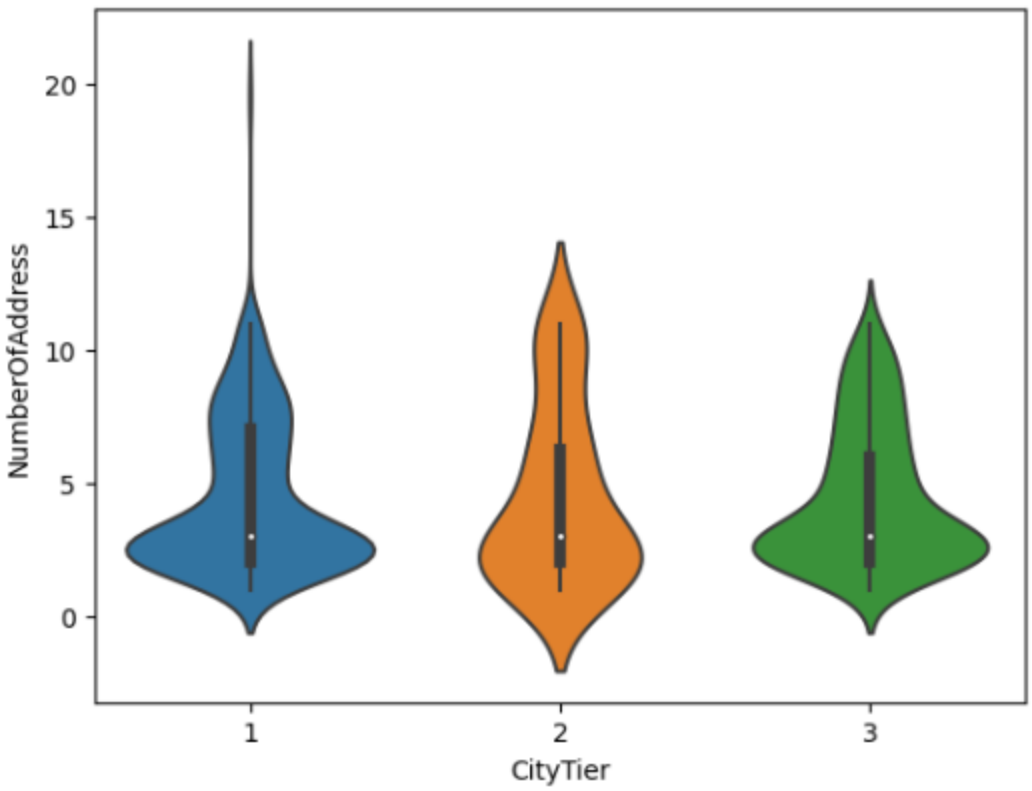
Analyzing the correlation between these two factors can help the company identify potential areas of improvement in their app or service to increase

overall customer satisfaction.



as we see people with less satisfaction score spend less time on the app than the people of satisfaction score 5 but also i do not think there is any relation between the satisfaction score and people's spent time on the app

Finding the the relation between NumberOfAddress and CityTier within the churn segment



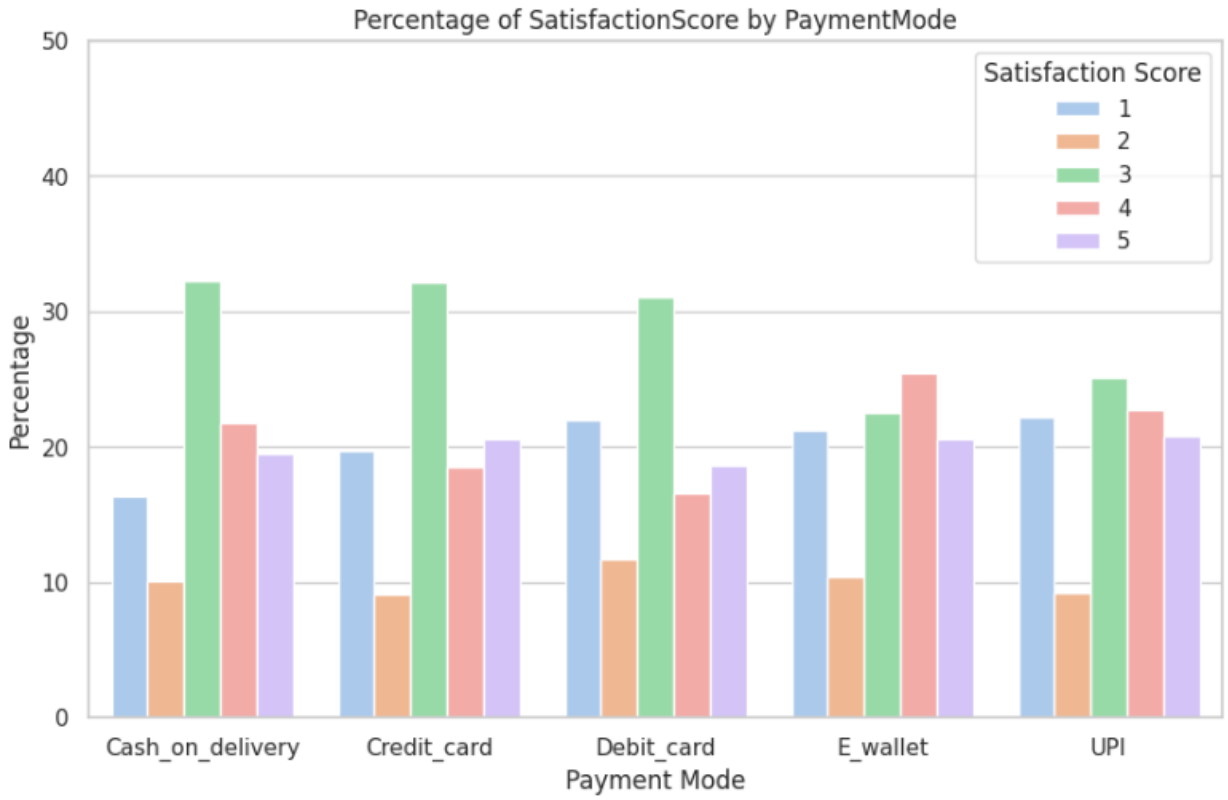
CityTier has a negative connection with NumberOfAddress. Higher city tiers are related with a lower average number of addresses and a more concentrated distribution. Customers in larger cities (CityTier 1) have more addresses on average than in smaller cities and towns in lower tiers. The link indicates that address density and type of location (metro vs. smaller cities vs. towns) influence how many addresses clients have across city types.

HYPOTHESIS TESTING - ANOVA

Here , to discover if the satisfaction score is related to the payment method, an Analysis of Variance (ANOVA) test will be performed. It will test the null hypothesis that the mean satisfaction scores are the same across payment forms, implying that there is no association between payment mode and satisfaction scores.

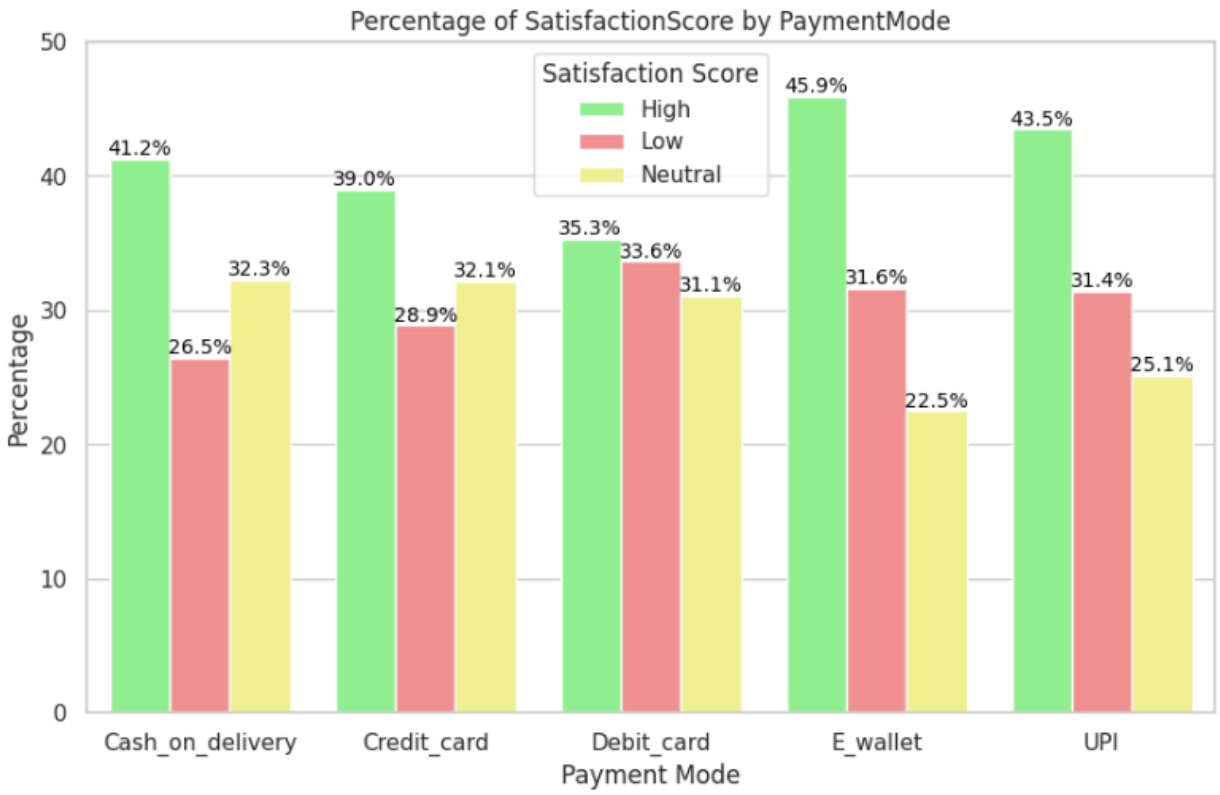
	df	sum_sq	mean_sq	F	PR(>F)
C(PaymentMode)	4.0	29.363281	7.340820	3.861413	0.003892
Residual	5625.0	10693.525529	1.901071	NaN	NaN

The p-value (PR(>F)) for 'PaymentMode' is 0.003892, which is less than 0.05 (a common alpha level for significance). The null hypothesis is rejected. This indicates that there is a statistically significant difference in the SatisfactionScore across different PaymentModes



Relationship between the satisfactory score of customer on service (SatisfactionScore) and the preferred payment method (PaymentMode).

It is preferable to reorganise the satisfaction score into broader categories ("Low," "Neutral," and "High") since it provides a more streamlined way to analysing client feedback and promotes easier communication of satisfaction levels inside the e-commerce business.



Cash On Delivery: While a big portion of consumers who use COD as a payment method are highly satisfied, there is also a sizable minority of customers who are not.

Credit Card: A relatively high percentage of "Low" satisfaction shows areas that may require attention.

Debit Card: indicates a balanced but diverse experience for debit card users.

E-wallet customers exhibit a clear preference for "High" satisfaction.

UPI: Has the highest "High" satisfaction score (43.5%), followed by "Neutral" (31.4%) and the lowest "Low" satisfaction (25.1%) among all payment methods. This suggests a very positive response to UPI.

According to this analysis, clients that utilise UPI and E-wallet as payment options are highly satisfied. To capitalise on this, the company should incorporate these payment alternatives into its client acquisition tactics and prioritise them in promotional activities, thereby increasing adoption rates. Building deeper partnerships with UPI and E-wallet providers may also be advantageous, resulting in more favourable conditions and combined promotional initiatives for client benefits. Furthermore, the organisation should consider evaluating and maybe adjusting its pricing strategies or cost structures for different payment sources. This modification could have a

significant impact on customer satisfaction with credit and debit card transactions, potentially improving their favorability.

GENERALISED INFERENCE

There isn't a significant difference between males and females in terms of average order. Males are more prone to churn, as we have 63.3% churned males from the app. Perhaps the company should consider increasing the items that attract the attention of males. We will investigate whether there are any other aspects that contribute to males being the most churned customers. Married individuals are the largest customer category in the company; perhaps the company could consider catering to the items that suit both single and married customers, as singles are the most likely to churn from the app. City Tier 2 has the highest tenancy rate, although it does not appear to be a significant factor. City Tier 3 has the highest average order, although it is not a major determinant in customer attrition. People with a lower satisfaction score spend less time on the app than people with a satisfaction score of 5, but I do not think there is any genuine relationship between the satisfaction score and people's time spent on the app.

City Tier 1 has the highest spent hours on the app. City Tier has a negative connection with NumberOfAddress. Higher City Tiers are associated with a lower average NumberOfAddress and a more concentrated distribution. Customers in larger cities (City Tier 1) tend to have more addresses on average than smaller cities and towns in lower tiers. This relationship suggests that address density and type of locality (metro vs smaller cities vs towns) influence how many addresses customers have across city types. There is a weak negative correlation between complaints and the number of days since the last order. Mobile phone users are more likely to churn, which could signal an issue with the app's mobile version.

Since the distance from warehouse to residence is identical across all city tiers, the company built a warehouse in the lowest city tier as well.

Laptops, accessories, and mobile phones are the preferred categories for all city tiers.

Preferred payment method for City Tier '1': Debit Card

Preferred payment option for City Tier '2': UPI.

Preferred payment method for City Tier '3': Electronic wallet.

There is a large commonality in debit card usage among levels.

City Tier '1' has the highest order count (10298 orders).

City Tier '3' has the highest mean order count, which suggests that their count is small but they have a lot of orders. 'Rich Tier'

When the proportion of orders from last year increases, the churn rate decreases, so OrderAmountHikeFromlastYear has a positive effect on the churn rate and we need to focus when customers.

When the percentage of orders last year increases, the churn rate decreases, therefore OrderAmountHikeFromlastYear has a positive influence on the churn rate, and we need to focus when the client has a percentage of 12% to 15%.

Customers who did not make a complaint have a higher DaySinceLastOrder, but it is only one customer, thus it is an outlier. If we remove it, customers with no complaints will have a lower DaySinceLastOrder.

Customers who spend a lot of time on the app have an OrderCount of 2, with a percentage of 67%.

Top two preferred categories for males: Others, Mobile Phone

Top two preferred categories for females: Grocery, Fashion

Churn decreases when more coupons are used.

SatisfactionScore has no effect on OrderCount.

There is no relationship between cashback amount and order count, although there is a positive relationship between cashback.

MACHINE LEARNING MODELS

Machine Learning (ML) is an innovative subset of Artificial Intelligence (AI) that centers around the utilization of data and algorithms to emulate the way that humans learn, gradually improving its accuracy. It enables systems to autonomously learn and improve from their experiences without being explicitly programmed, making it an essential tool in fields where it's impractical or impossible for humans to process vast amounts of data.

In essence, Machine Learning provides the system with the ability to automatically learn and progress through experience. It focuses on the advancement of computer algorithms, improving their performance over time by gaining access to data and learning from it.

In this project, several machine learning models were used to analyze and predict customer churn. These models include:

1. **Logistic Regression (LR)**: Logistic Regression is a statistical model that is used in machine learning for binary classification problems. In the context of customer churn, it can be used to predict two possible outcomes: whether a customer will churn or not. Logistic Regression uses the concept of odds ratios to predict outcomes and is an efficient algorithm when dealing with high dimensionality.
2. **Support Vector Machine (SVM)**: The Support Vector Machine is a powerful, flexible supervised machine learning algorithm used for classification or regression. It uses a technique called the kernel trick to transform input data, then identifies the optimal boundary that maximizes the margin between different classes of data.
3. **Decision Tree (DT)**: The Decision Tree is a non-parametric supervised learning method used for classification and regression. Decision trees learn from data to approximate a target variable with a set of if-then-else decision rules, which make them interpretable and transparent.
4. **Random Forest (RF)**: Random Forest is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes of the individual trees. A Random Forest ensures that the behavior of each individual tree is not too correlated with the behavior of any of the other trees in the model, enhancing the overall result.
5. **XGBoost (XGBM)**: XGBoost, or Extreme Gradient Boosting, is a powerful machine learning algorithm that is a type of gradient boosting. It is renowned for its effectiveness and speed, making it a popular choice for many machine learning competitions.
6. **AdaBoost (Adaptive Boosting)**: AdaBoost is a machine learning meta-algorithm that can be used with many types of learning algorithms to improve their performance. The output of the weak learning algorithms is combined into a weighted sum that represents the final output of the boosted classifier, making it adaptive in the sense that weak learners are tweaked in favor of those instances misclassified by previous classifiers.

Each of these models comes with its own strengths and weaknesses, and their performance can greatly vary depending on the specific nature of the data. By employing a variety of models in our analysis, we are better equipped to ensure that our predictions are both robust and reliable, thereby providing a comprehensive and balanced analytic approach to predicting customer churn.

	Train_Accuarcy	Test_Accuarcy
Logistic Regression	0.768270	0.773204
Support Vector Machine	0.904278	0.876893
Decision Tree	1.000000	0.937087
Random Forest	1.000000	0.966990
XGBClassifier	1.000000	0.957282
AdaBoostClassifier	0.873314	0.815146

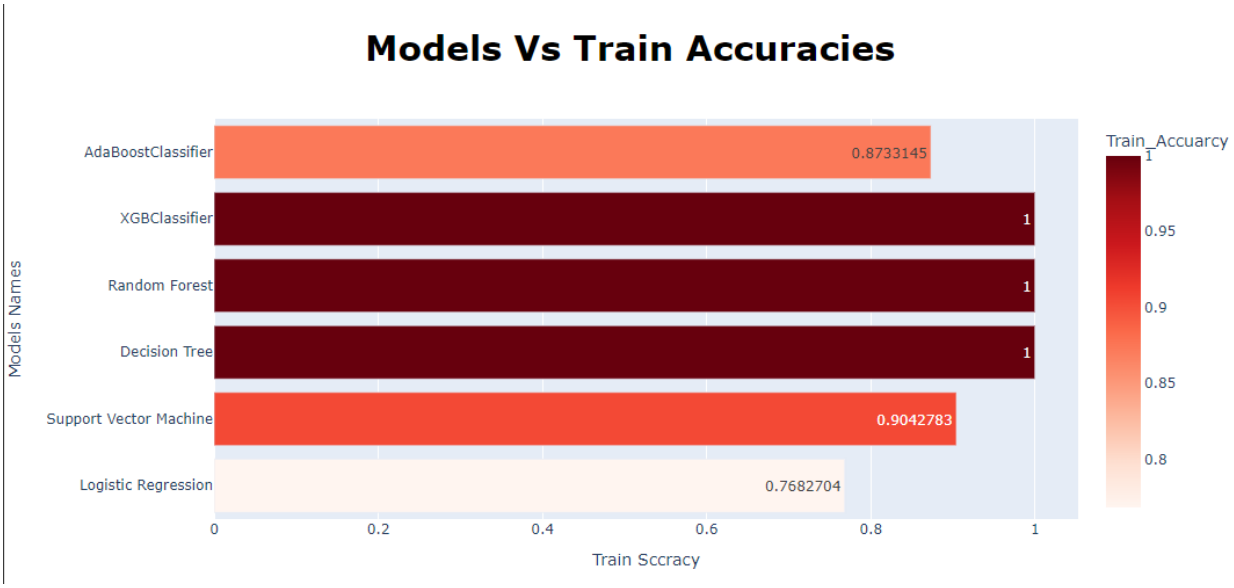
```
Using model: Logistic Regression
Trainning Score: 0.7682703512568669
Test Score: 0.7732038834951457
Acc Train: 0.7682703512568669
Acc Test: 0.7732038834951457

************************************************************

Using model: Support Vector Machine
Trainning Score: 0.9042783419344098
Test Score: 0.8768932038834951
Acc Train: 0.9042783419344098
Acc Test: 0.8768932038834951

************************************************************

Using model: Decision Tree
Trainning Score: 1.0
Test Score: 0.9370873786407767
Acc Train: 1.0
Acc Test: 0.9370873786407767

************************************************************

Using model: Random Forest
Trainning Score: 1.0
Test Score: 0.9669902912621359
Acc Train: 1.0
Acc Test: 0.9669902912621359

************************************************************

Using model: XGBClassifier
Trainning Score: 1.0
Test Score: 0.9572815533980582
Acc Train: 1.0
Acc Test: 0.9572815533980582

************************************************************

Using model: AdaBoostClassifier
Trainning Score: 0.8733144664558016
Test Score: 0.8151456310679611
Acc Train: 0.8733144664558016
Acc Test: 0.8151456310679611
```

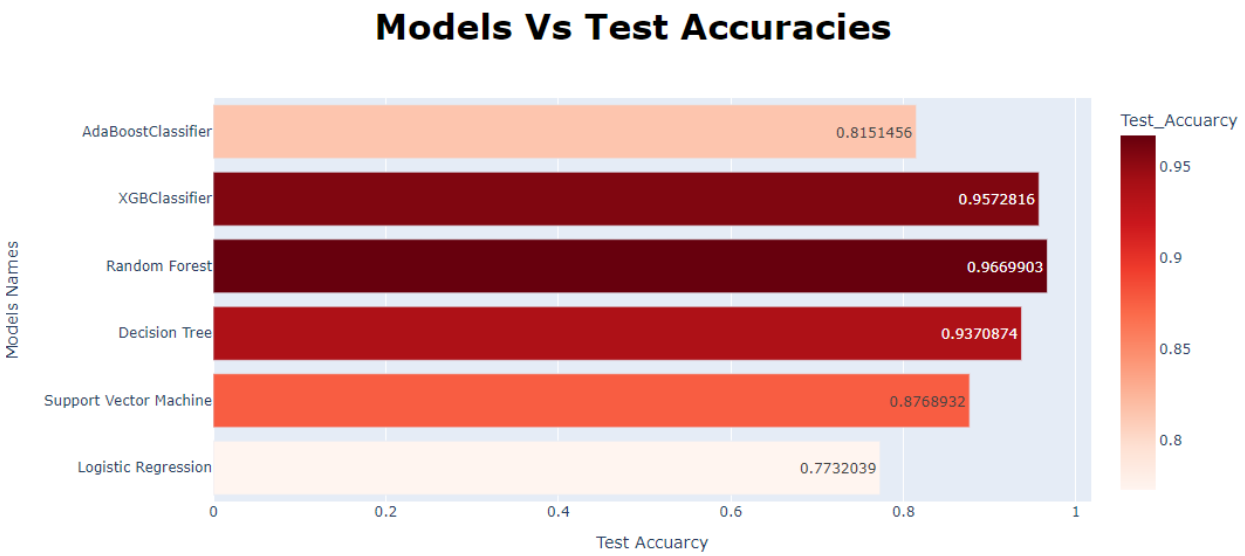
INFERENCE:

1. **Decision Tree, Random Forest, and XGB Classifier**: These tree-based models reached a perfect training accuracy of 1.0, suggesting they fitted the training data exceptionally well. However, their test accuracies were lower, with Decision Tree having the best test accuracy at 0.937087, followed by XGBClassifier at 0.957282 and Random Forest at 0.968990. This suggests that while these models were able to learn the training data thoroughly, they may have overfitted to some extent, resulting in a minor decline in performance on the unknown test data.
2. **Support Vector Machine (SVM)**: The SVM model has a fairly high training accuracy of 0.904278 and a test accuracy of 0.876893. This shows that the SVM model generalised well to the test data, with no notable overfitting or underfitting concerns.

3. **Logistic Regression**: The Logistic Regression model achieved the lowest training accuracy of 0.768270 and test accuracy of 0.773204. This could imply that the Logistic Regression model was underfitting the training data to some level, although performing relatively well on the test data.
4. **AdaBoostClassifier**: The AdaBoostClassifier had a training accuracy of 0.873314 and a test accuracy of 0.815146. While it did not have the same training accuracy as tree-based models or SVM, it outperformed Logistic Regression on test data.
5. **Overfitting and Generalisation**: The discrepancy between training and test accuracies may indicate overfitting or underfitting. Models with a big difference between training and test accuracies (e.g., Decision Tree, Random Forest, and XGBClassifier) may be overfitting to the training data, whereas models with a smaller difference (e.g., SVM and Logistic Regression) are more likely to generalise to new, unseen data.

In terms of test accuracies, the Random Forest model ranks highest at 0.968990, followed by XGBClassifier at 0.957282, Decision Tree at 0.937087, Support Vector Machine at 0.876893, AdaBoostClassifier at 0.815146, and Logistic Regression at 0.773204.

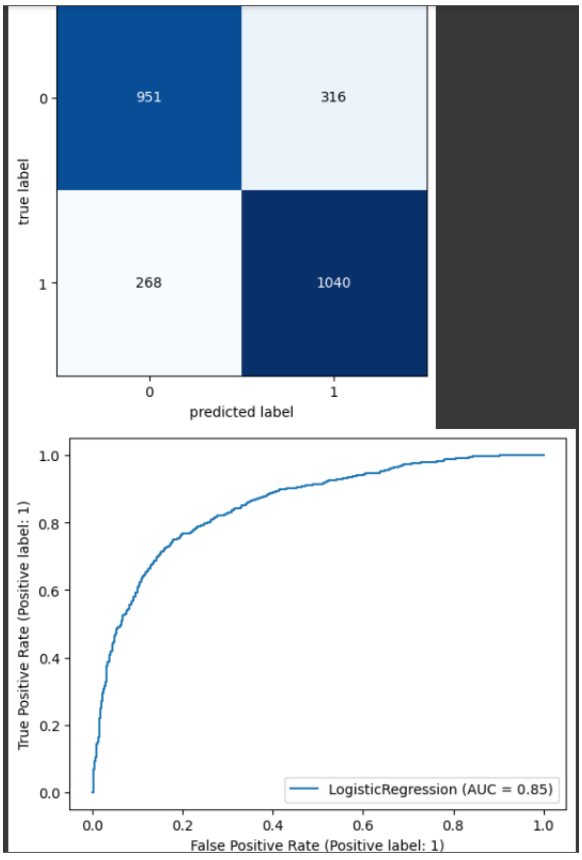




CONFUSION MATRIX DISPLAY OF EACH MODEL

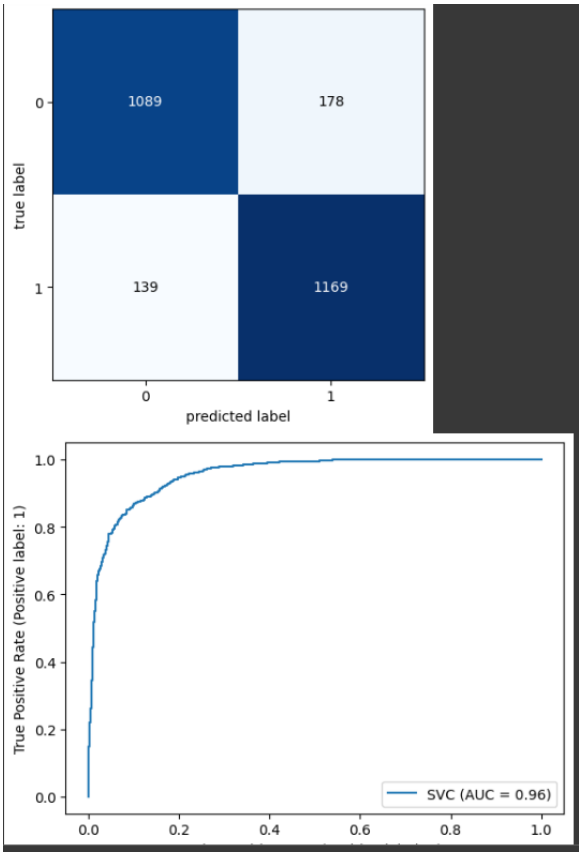
Logistic Regression

```
Accuracy = 0.7732038834951457
ROC Area under Curve = 0.7728494915630604
              precision    recall  f1-score   support

           0    0.78015   0.75059   0.76508      1267
           1    0.76696   0.79511   0.78078      1308

    accuracy                        0.77320      2575
   macro avg    0.77355   0.77285   0.77293      2575
weighted avg    0.77345   0.77320   0.77306      2575
```



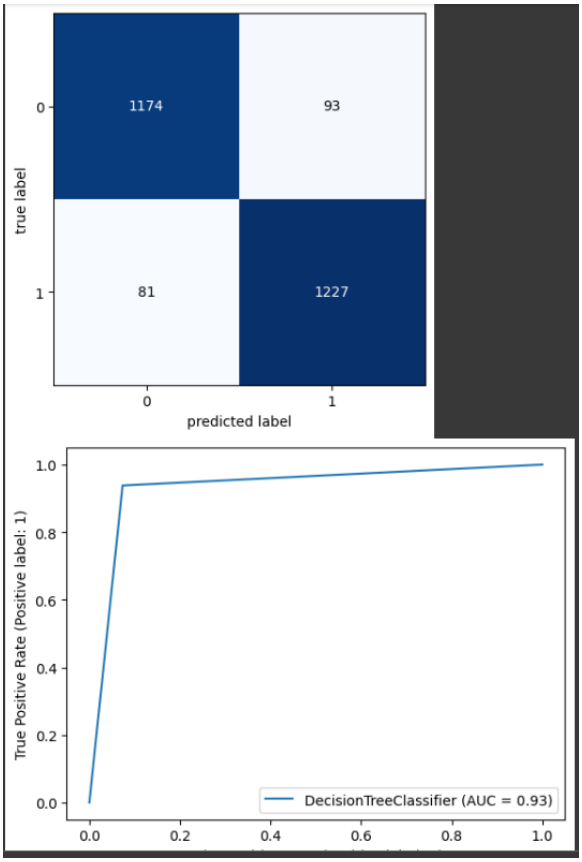
Support Vector Machine

```
Accuracy = 0.8768932038834951
ROC Area under Curve = 0.8766207709704593
              precision    recall  f1-score   support

           0    0.88681   0.85951   0.87295      1267
           1    0.86785   0.89373   0.88060      1308

    accuracy                        0.87689      2575
   macro avg    0.87733   0.87662   0.87677      2575
weighted avg    0.87718   0.87689   0.87684      2575
```



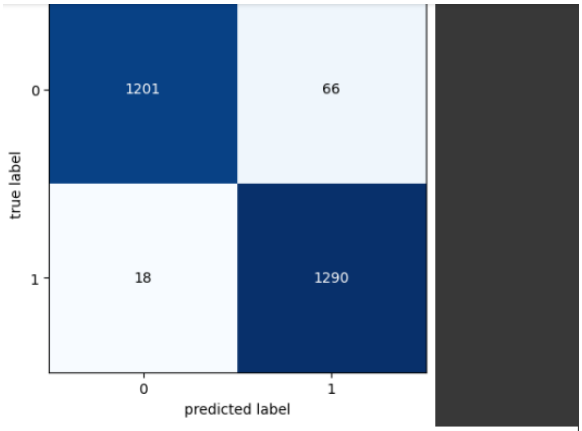
Decision Tree

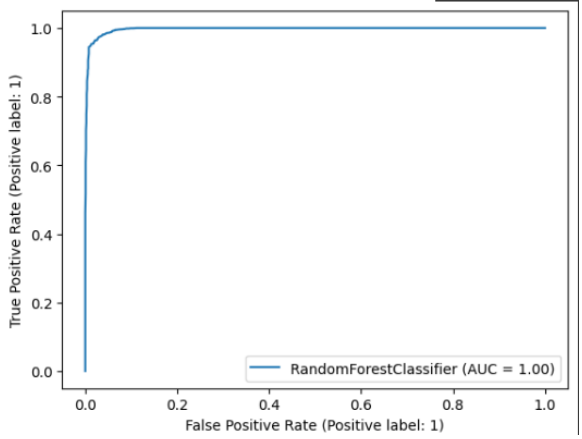
```
Accuracy = 0.9324271844660195
ROC Area under Curve = 0.9323358290551256
              precision    recall  f1-score   support

           0    0.93546   0.92660   0.93101      1267
           1    0.92955   0.93807   0.93379      1308

    accuracy                        0.93243      2575
   macro avg    0.93250   0.93234   0.93240      2575
weighted avg    0.93245   0.93243   0.93242      2575
```



Random Forest

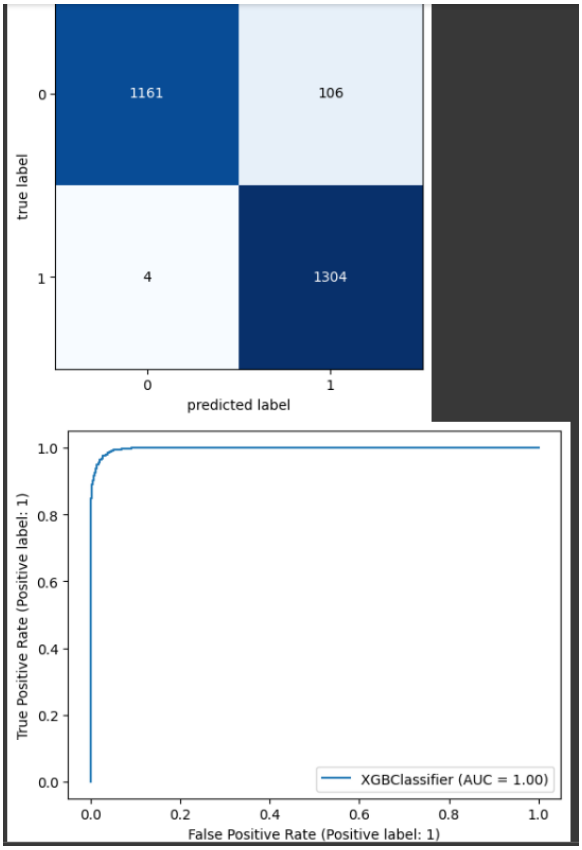
```
Accuracy = 0.967378640776699
ROC Area under Curve = 0.9670734886280531
              precision    recall  f1-score   support

           0    0.98523   0.94791   0.96621      1267
           1    0.95133   0.98624   0.96847      1308

    accuracy                        0.96738      2575
   macro avg    0.96828   0.96707   0.96734      2575
weighted avg    0.96801   0.96738   0.96736      2575
```





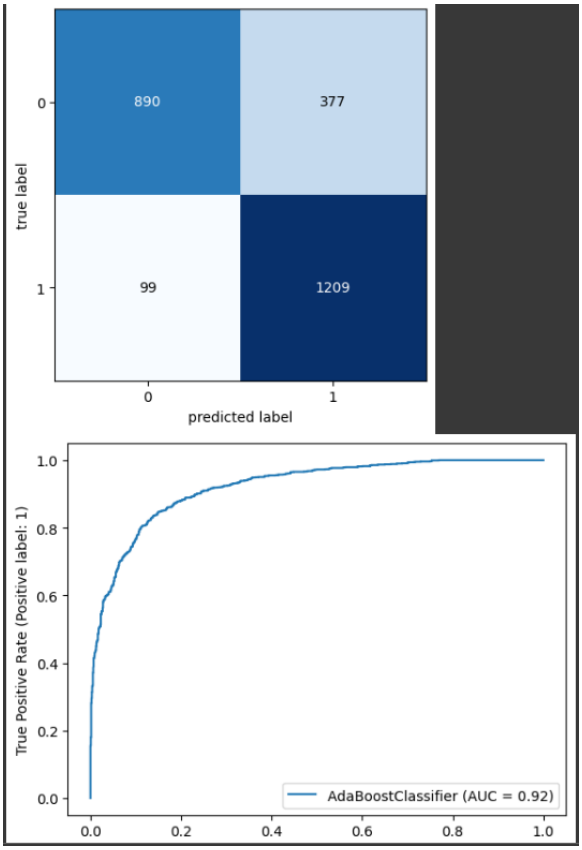
XGBoost

```
Accuracy = 0.9572815533980582
ROC Area under Curve = 0.9566398509325166
              precision    recall  f1-score   support

           0    0.99657   0.91634   0.95477      1267
           1    0.92482   0.99694   0.95953      1308

    accuracy                        0.95728      2575
   macro avg    0.96069   0.95664   0.95715      2575
weighted avg    0.96012   0.95728   0.95719      2575
```



AdaBoost

```
Accuracy = 0.8151456310679611
ROC Area under Curve = 0.8133793255758384
              precision    recall  f1-score   support

           0    0.89990   0.70245   0.78901      1267
           1    0.76230   0.92431   0.83552      1308

    accuracy                        0.81515      2575
   macro avg    0.83110   0.81338   0.81226      2575
weighted avg    0.83000   0.81515   0.81263      2575
```



From the above predictive model analysis , XGBoost and AdaBoost showcases the best of accuracies . XgBoost with an accuracy of 95.7% accuracy and AdaBoost an accuracy of 82%.

DEEP LEARNING MODEL

Deep Learning Model for Binary Classification using Embeddings

We are developing a deep learning model for binary classification using embeddings and dense layers. The model is built with the Keras package, a high-level neural network API in Python.

Pre-Processing

Before training the deep learning model, the input data (X_encoded) is preprocessed using the following steps:

Each feature column in X_encoded is encoded into numerical values using scikit-learn's LabelEncoder. This step is important since the Embedding layer requires integer input.

The encoded feature columns are concatenated into a two-dimensional numpy array, with each row representing a sample and each column representing a feature.

The input_dim for the Embedding layer is set to one more than the maximum value in X_encoded.This ensures that all input values are within the valid range for the Embedding layer.

Model architecture

The deep learning model is defined using Keras' Sequential API. The model architecture is composed of the following layers:

Masking Layer: This layer is used to mask (or ignore) the input data's padding values, which are set to zero. This is significant because the Embedding layer will accept padding values as acceptable input, which can lead to unwanted behaviour.

The Embedding Layer converts each input (integer) value to a dense vector representation. The input_dim parameter is set to the maximum value in X_encoded plus one, whereas the output_dim parameter specifies the dimensionality of the embedding vectors. The input_length option controls the length of the input sequences.

Dense Layers: After the Embedding layer, there are three dense (fully connected) layers with ReLU activation functions. The first dense layer has 128 units, the second has 64 units, and the final layer has a single unit with a sigmoid activation function for binary classification.

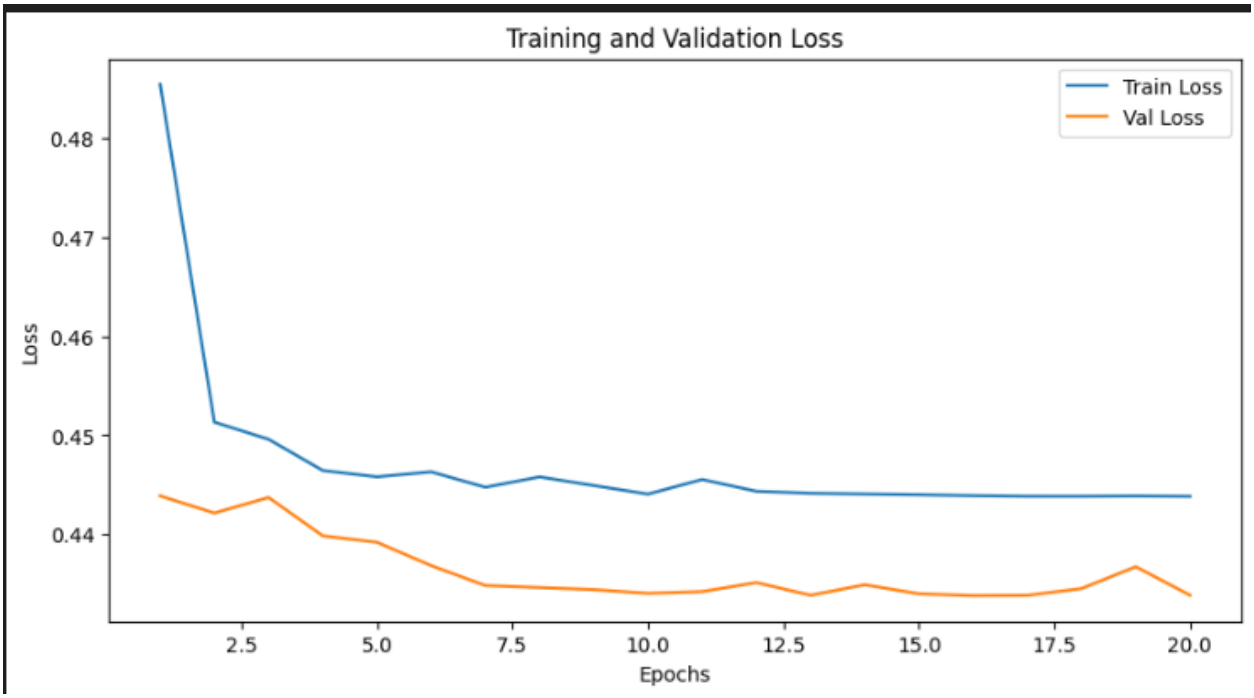
Model Compilation and Training

The model is built using the Adam optimizer and the binary_crossentropy loss function, which are appropriate for binary classification applications. The accuracy metric is also used to track the model's performance while training.

The fit approach is utilised to train the deep learning model. X_encoded and y represent the input data and target labels, respectively. The epochs parameter provides the number of iterations performed across the entire dataset, whereas the batch_size parameter specifies the amount of samples propagated through the network at once. The validation_split option divides a portion of the training data for validation, allowing you to check the model's performance on previously unseen data and avoid overfitting.

Result

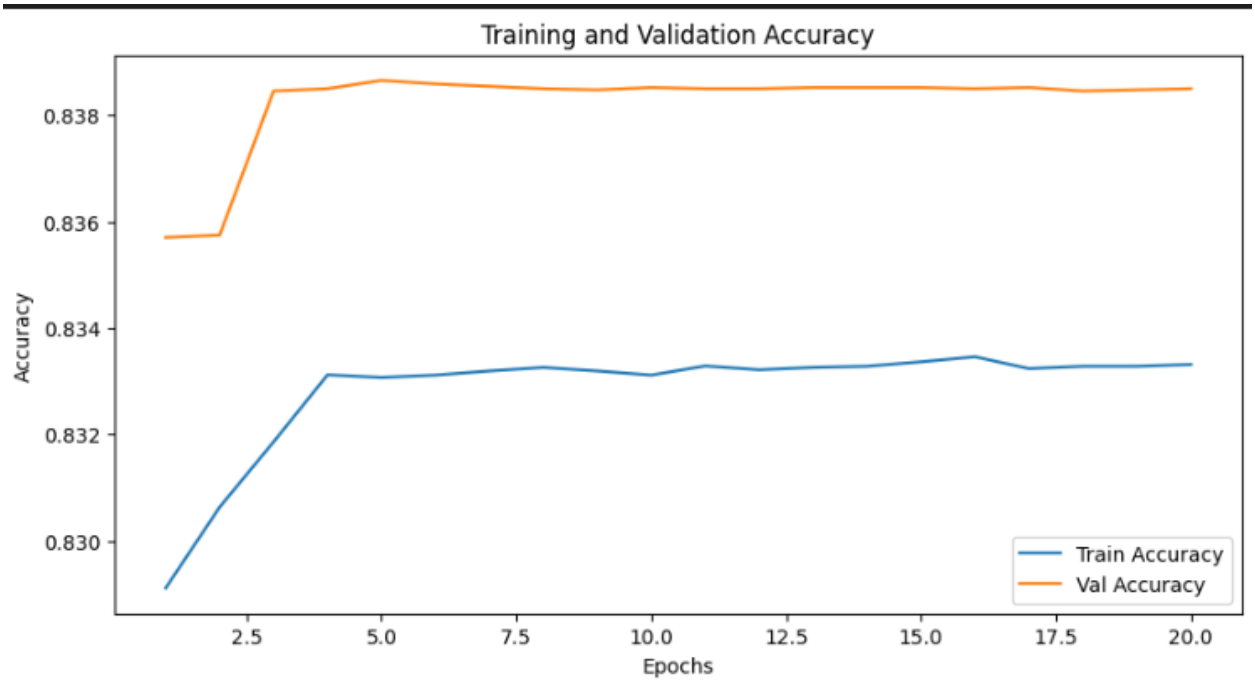
```
Epoch 20/20
Train Loss: 0.4438, Train Accuracy: 0.8333
Val Loss: 0.4338, Val Accuracy: 0.8385
```



1. Initial Loss Values: The training loss is rather large, approximately 0.48, whereas the validation loss is lower, around 0.46. This could mean that the model initially overfits the training data to some degree.
2. Loss Curve Behavior: The initial epochs show a rapid drop in training loss, demonstrating effective model learning from the data. The validation loss diminishes initially, although at a slower rate than the training loss. After a certain point (about epoch 5-6), the validation loss flattens and even slightly increases, whilst the training loss continuously decreases.
3. Overfitting Indication: The divergence of the training and validation loss curves after a particular number of epochs indicates that the model is beginning to overfit the training data. While the training loss continues to reduce, the validation loss does not improve and may even worsen, indicating that the model does not generalize well to new data.
4. Potential Early stopping: To avoid more overfitting, an early stopping strategy could be used. By monitoring the validation loss, the training process can be terminated when it begins to increase or stops dropping after a specified number of epochs. This could aid in developing a model that generalizes better to new data.
5. Hyperparameter Tuning: The overfitting behavior observed in the loss curves may signal that the model architecture or hyperparameters (e.g., learning rate, regularization approaches) require adjustment. Dropout, L1/L2 regularization, and altering model complexity can all help to reduce overfitting and increase generalization.

6. Convergence: While the loss curves have not entirely converged, the validation loss appears to be rather constant until about epoch 15, implying that the model has reached a fair level of convergence. Due to the overfitting issue, more training may not appreciably improve the model's performance on previously unknown data.

Overall, the loss curve behavior indicates that the model initially learns efficiently from the training data, but it begins to overfit after a certain number of epochs. To increase the model's generalization performance on previously unknown data, appropriate strategies such as early stopping, regularization, and hyperparameter tuning should be considered.

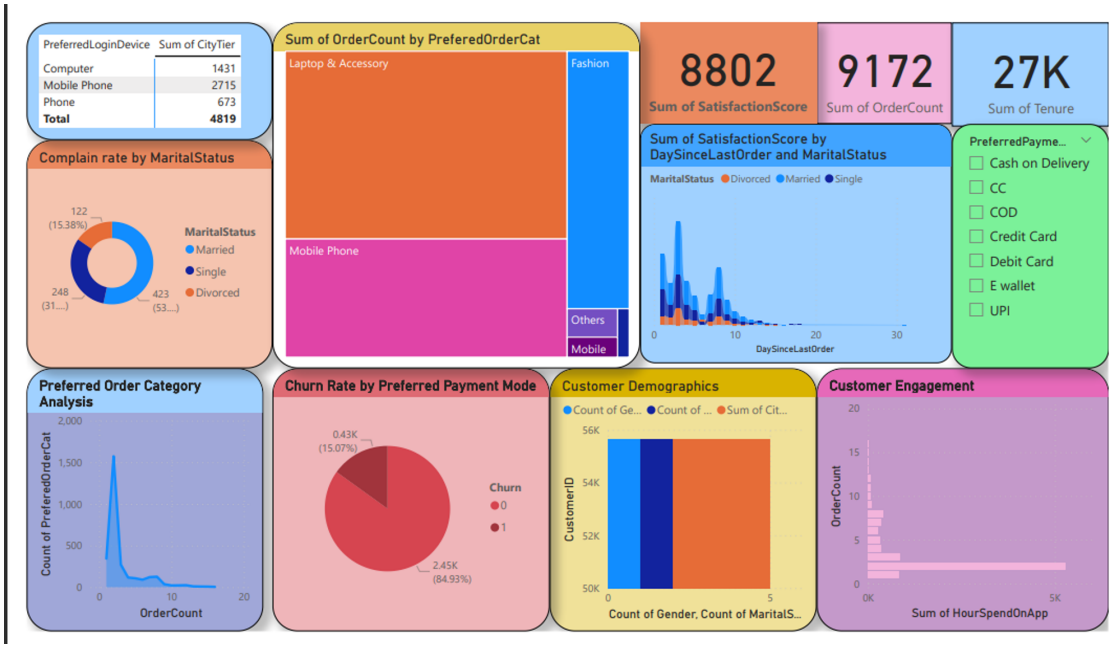


1. Initial Accuracy: Both the training and validation accuracy begin at approximately 0.83, showing that the model has a decent initial accuracy on both the training and validation data.
2. Accuracy Improvement: Training accuracy rapidly increases in early epochs, reaching about 0.838 by epoch 5. The validation accuracy improves initially, although at a slower rate than the training accuracy.
3. Differences in Training and Validation Accuracy: Following the initial improvement, a considerable gap develops between the training and validation accuracy curves. The training accuracy continues to rise, approaching its peak, while the validation accuracy plateaus and even slightly declines.

4. Overfitting Indication: The difference in training and validation accuracy curves indicates that the model is overfitting the training data. While it achieves excellent accuracy on the training set, it does not perform as well on the unseen validation data.
5. Potential Early halting: As with loss curve analysis, an early halting strategy could be used to prevent future overfitting. The training process could be interrupted if the validation accuracy stops improving or begins to decrease after a specified number of epochs.
6. Hyperparameter Tuning and Regularisation: The overfitting behaviour noticed in the accuracy curves suggests that the model's hyperparameters or architecture may require adjustment. Techniques like dropout, L1/L2 regularisation, and model complexity adjustment may assist reduce overfitting and increase generalisation to new data.
7. Maximum Validation Accuracy: The validation accuracy appears to peak at 0.834-0.835, which may be considered the highest possible accuracy for this model on unseen data, given the existing hyperparameters and architecture.

Overall, the accuracy curves reveal that, while the model performs well on training data, it struggles to generalise to new data due to overfitting. To increase the model's generalisation performance and close the accuracy gap between training and validation, appropriate strategies such as early stopping, regularisation, and hyperparameter tuning should be considered.

POWERBI DASHBOARD



CONCLUSION

In conclusion, our comprehensive analysis of churn in the e-commerce domain, combined with the use of a diverse range of machine learning and deep learning models such as Logistic Regression, Support Vector Machines, Random Forest, AdaBoost, and XGBoost, provided valuable insights into customer behaviour and retention strategies.

Through extensive churn analysis, we identified significant factors impacting client attrition, allowing us to build tailored retention tactics. Our data show that tenure, preferred payment mode, happiness score, and order frequency all have a significant impact on churn, emphasising the necessity of personalised customer engagement and service enhancements.

Furthermore, by combining several machine learning and deep learning models, we were able to create robust predictive models that could properly detect probable churners. We achieved greater predictive performance by combining ensemble methodologies and hybrid modelling approaches, which improved the effectiveness of our churn prediction efforts.

Overall, our findings highlight the relevance of data-driven decision-making in e-commerce enterprises, as well as the utility of employing advanced analytics approaches to optimise client retention tactics and achieve long-term business growth. As we continue to enhance and evolve our predictive models, we anticipate increasing customer pleasure, loyalty, and, ultimately, corporate profitability.