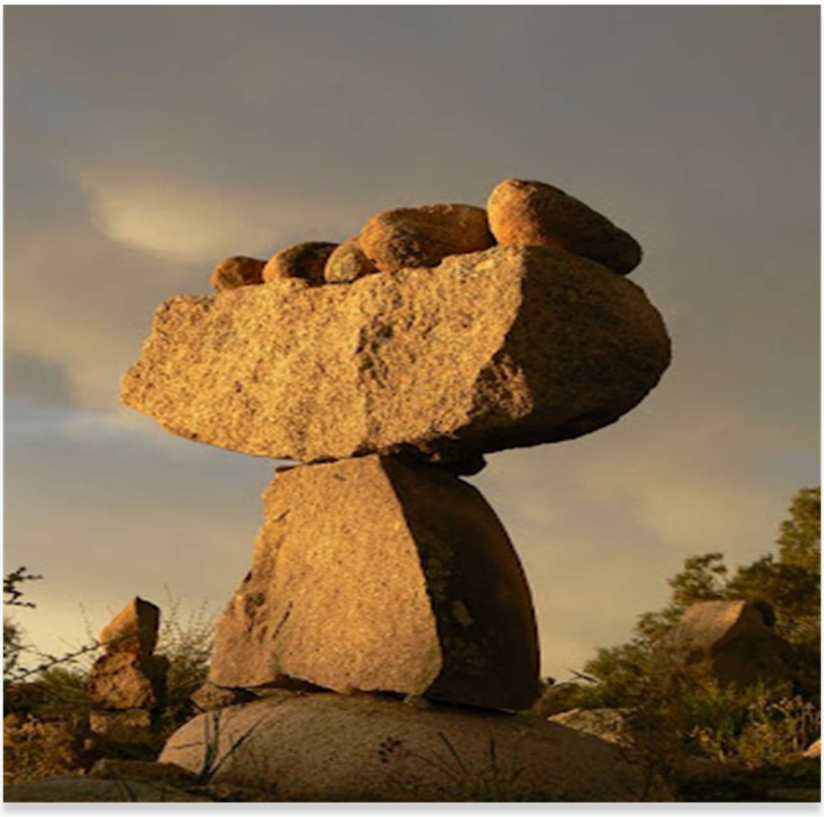

Customer Churning Analysis

Business Report

Final Report

DSBA

Vikash Kumar

Aug'22 Batch

Contents

List of figures and plots

List of Tables

Introduction :

Objective doing the Project

While conducting the churn prediction project, the main goal is to develop a robust and accurate churn prediction model that can identify potential churners among the existing accounts of the DTH provider company.

The Churn rate is around 17% which is high at per Industry standard. Hence, reducing the Churn to 10% will increase the revenue by 7% at least.

The primary objectives and desired achievements of the project are as follows:

Accurate Churn Prediction: The main objective is to build a churn prediction model that can accurately identify accounts at risk of churning. The model should be able to distinguish between customers who are likely to churn and those who are likely to stay, enabling the company to take proactive measures to retain valuable customers.

Customer Segmentation: The project aims to segment customers based on their churn probability and behaviour. By categorizing customers into different groups, the company can tailor personalized offers and incentives for each segment, improving the effectiveness of retention efforts.

Actionable Insights: The analysis should provide actionable insights that the company can use to make informed decisions and develop targeted strategies. The insights should highlight the key factors influencing churn and provide recommendations on how to retain potential churners effectively.

Cost-Effective Campaign Recommendations: The project should generate unique and cost-effective campaign recommendations that align with the company's business objectives. The proposed campaigns should be innovative, tailored to customer needs, and designed to minimize financial losses while maximizing customer retention.

Business Growth and Competitiveness: Ultimately, the goal is to contribute to the company's business growth and competitiveness. By improving customer retention rates, the company can increase revenues, enhance market share, and strengthen its position in the market.

Customer Satisfaction and Loyalty: Improving customer retention not only impacts the bottom line but also enhances customer satisfaction and loyalty. Satisfied and loyal customers are more likely to become brand advocates and refer new customers, leading to organic growth.

Data-Driven Decision Making: The project should promote data-driven decision-making within the company. By leveraging data and analytics, the company can make well-informed decisions that are backed by evidence and insights from the churn prediction model.

Positive Impact on Company Reputation: Reducing customer churn and retaining valuable customers can have a positive impact on the company's reputation. A company known for its customer-centric approach and high customer retention rates can attract more customers and gain a competitive edge in the market.

Overall, the desired achievement of this project is to equip the company with a powerful churn prediction model and actionable recommendations that can be used to improve customer retention, reduce churn, and optimize marketing efforts. By achieving these goals, the company can enhance its position in the market and achieve sustainable business growth in the face of intense competition.

Data Overview

Data Skewness : Skewness can impact the analysis like Data Imbalance, Impact on Model Evaluation , Feature Importance, Overfitting, Class Separability

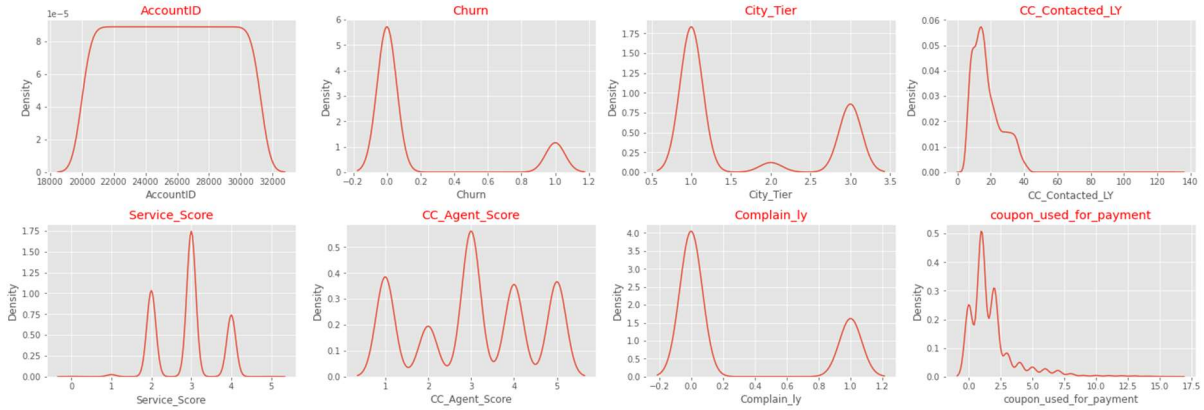

FIGURE 1 SKEWNESS

The positive skewness value suggests a right-skewed distribution of days since the last customer care connection. This indicates that there may be a higher concentration of customers with relatively fewer days since their last customer care connection, with a long tail of customers who have had a longer time since their last connection. Overall, interpreting skewness values helps identify the shape and asymmetry of the distributions, which can provide valuable insights into the underlying patterns and behaviours of the variables.

Five-Fold Statical Summary

	Count	mean	std	min	25%	50%	75%	max
Tenure	11260	10.83	12.83	0	2	8	16	99
CC_Contacted_LY	11260	17.87	8.81	4	11	16	23	132
Account_user_count	11260	3.7	1	1	3	4	4	6
CC_Agent_Score	11260	3.07	1.37	1	2	3	4	5
rev_per_month	11260	6.13	11.52	1	3	4	7	140
rev_growth_yoy	11260	16.19	3.76	4	13	15	19	28
coupon_used_for_paymen	11260	1.79	1.97	0	1	1	2	16
Day_Since_CC_connect	11260	4.58	3.65	0	2	3	7	47
cashback	11260	194.53	175.06	0	147.89	163.17	197.31	1997

TABLE 1DATA DESCRIPTION

Insights: The analysis explores various numerical variables and their relation to churn in a business. Average customer tenure, city tier, customer care contacts, service score, agent score, revenue per month, and complaint rate are examined. Churn rates may vary based on these factors, suggesting potential areas for improvement or targeting. The standard deviations indicate the diversity within each variable, shedding light on the impact of different customer behaviours and experiences on churn.

Univariate Analysis

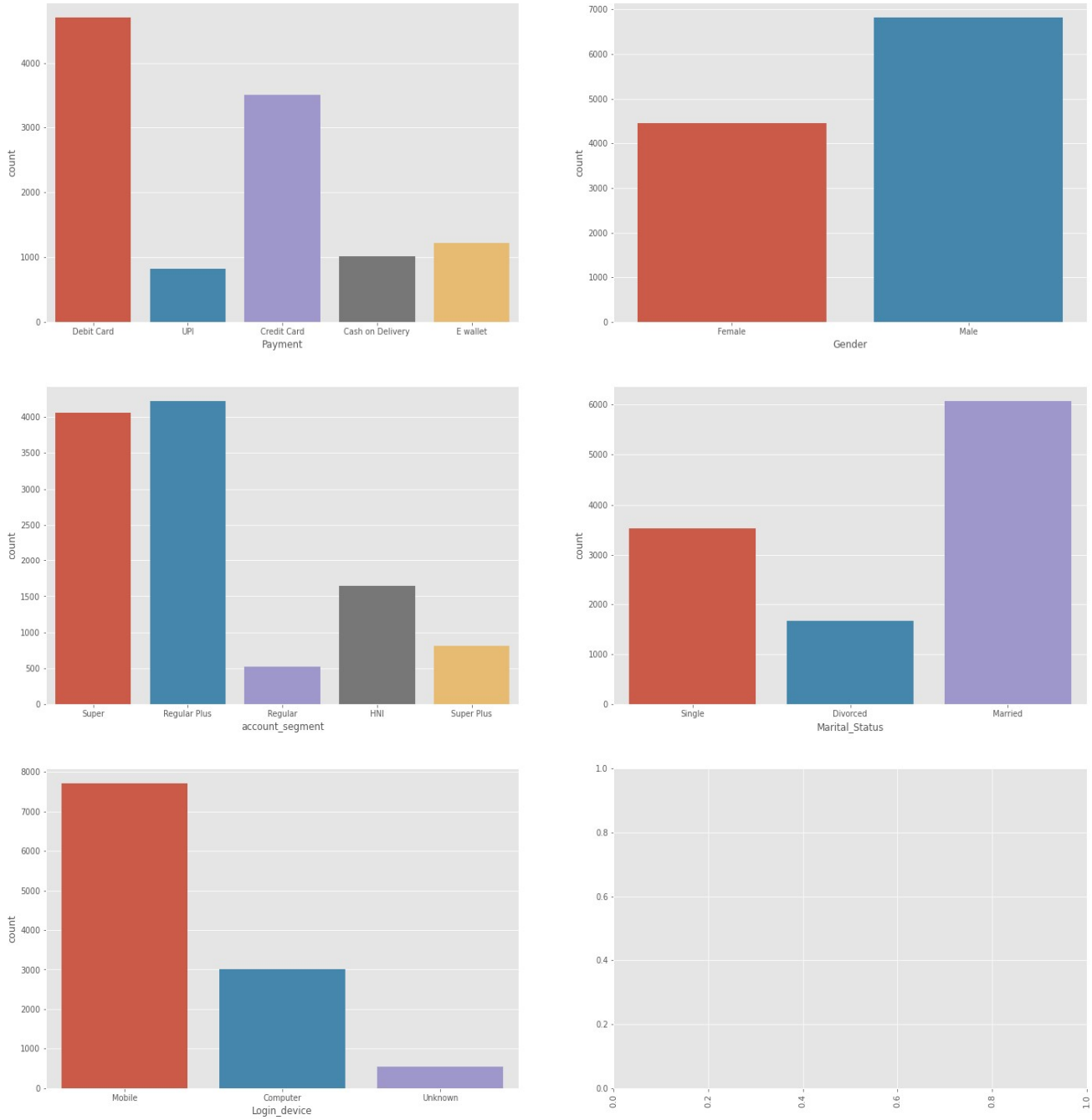

FIGURE 2 UNIVARIATE FOR CATEGORICAL VARIABLES

Categorical data Insights:

Our comprehensive analysis reveals noteworthy insights into payment preferences and customer demographics, shedding light on pivotal aspects for strategic decision-making. Here's a distilled summary of the key findings:

Payment Trends: Maximizing Convenience

Debit Dominance**:** Debit card payments take the lead, registering approximately 4500 transactions, followed closely by credit card payments at around 3500 transactions. This signifies a preference for seamless and convenient debit card transactions.

Gender Dynamics: The male customer segment holds a prominent presence, surpassing female customers by over 50% in terms of transaction count.

Championing Regular Plus: The Regular Plus customer category emerges as the frontrunner with an impressive count of approximately 4200 customers. Following closely, Super customers contribute significantly with around 4000 individuals. This robust distribution bodes well for business continuity and enhanced contributions.

Marital Status and Market Impact:

Marital Landscape: The married segment stands out as the most substantial, encompassing nearly 6000 customers. This sizeable population offers promising opportunities for targeted engagement strategies and sustainable growth.

Diverse Payment Modes: Coupons Count

Coupon Currency: Coupons emerge as a noteworthy payment mode, with Coupon 1 reigning supreme with a usage count exceeding 4000. This observation aligns coupons alongside the debit and credit card modes, showcasing their prevalence and relevance.

Device Preference: Mobile Mastery

Mobile Dominance: The mobile device emerges as the undisputed leader in login device preference, commanding usage over computers by an astonishing ratio of more than 200 to 1. This emphasizes the significance of optimizing mobile platforms for customer interaction and engagement.

These insights encapsulate pivotal trends, offering a comprehensive view of payment behaviour and customer demographics. Leveraging these findings, businesses can strategically tailor their approaches to capitalize on preferences, enhance customer experiences, and drive sustainable growth.

Bivariate Analysis for Numerical Variables

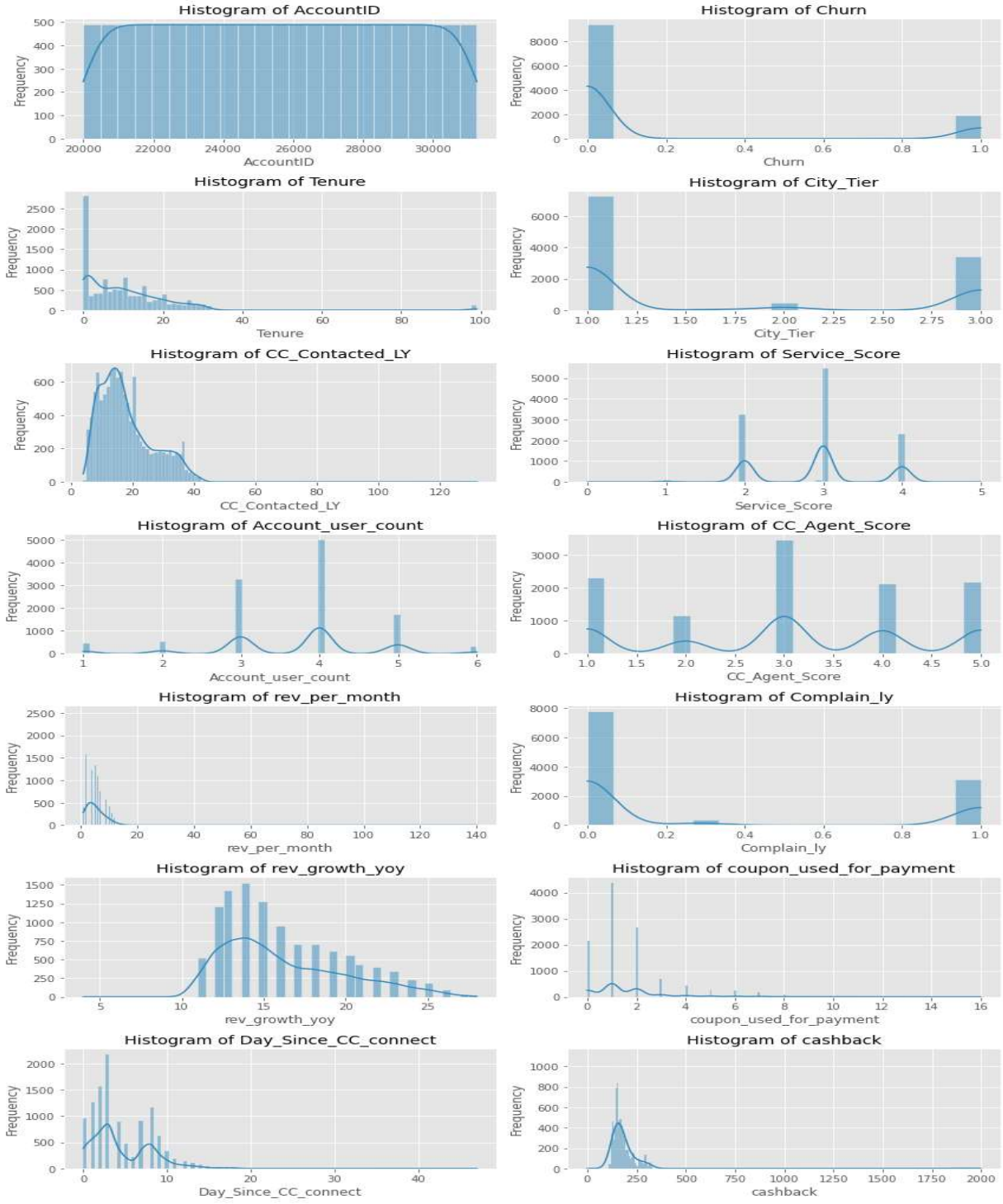

FIGURE 3 BIVARIATE FOR NUMERICAL VARIABLES

Numerical data insights:

CC_Contacted_LY and Day_Since_CC_connect has right-handed data distribution which means the long tail at right side of data in variables.

Rest variables are binary.

Bivariate Analysis

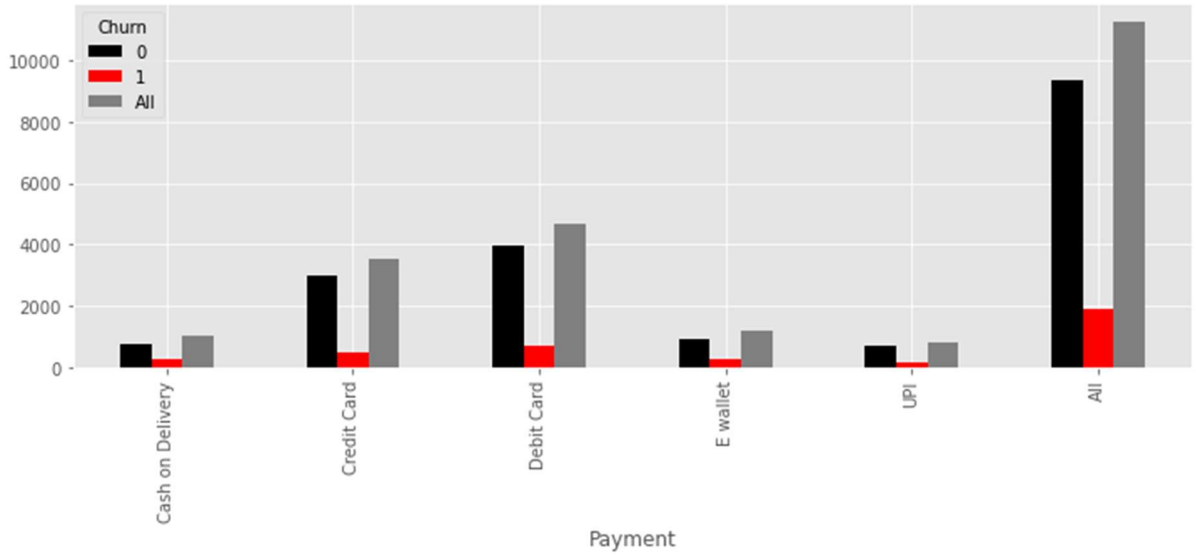

FIGURE 4 CHURN VS PAYMENT MODE

Insight: In case of Debit card transaction there is high Churning of customers followed by Credit Card .The Churn customers from debit card payment are 702 number of customers while Credit card has aprox 498.Seems some technical snag in payment receiving line

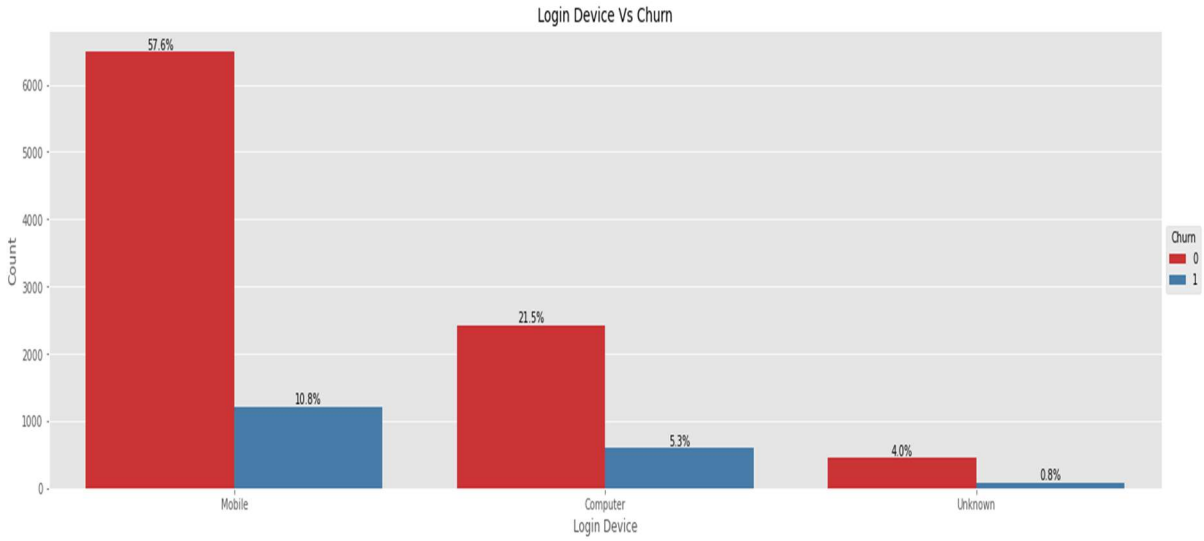

FIGURE 5 CHURN VS LOGIN DEVICE

Insight: Out of total the 58% customers are using mobile as login device and 21% are computer. The Churning rate of customers are higher in case of Mobile device user in respect to computer users as login device.

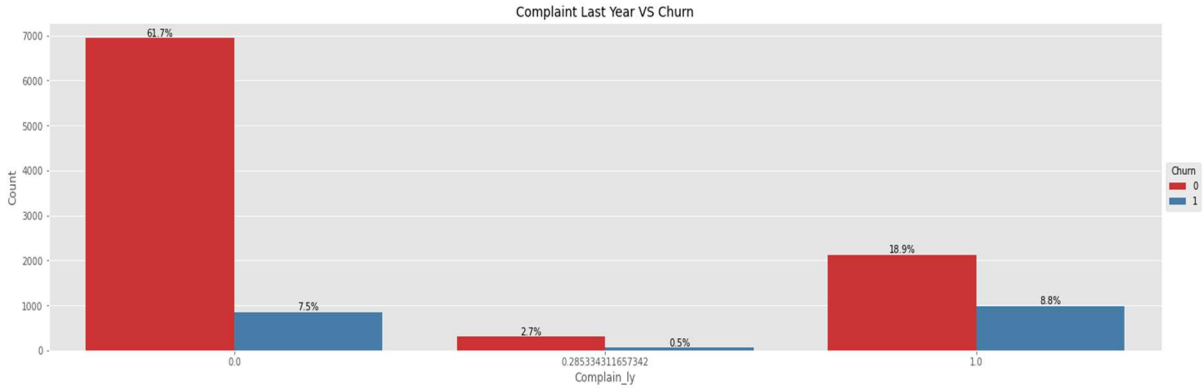

FIGURE 6 CHURN VS COMPLAINT LAST YEAR

Insights: Out of total the 19% customers are made complaint last year and 9% are left the services. This might be due to the Unsatisfactory or Non resolution of concerns in time to customers.

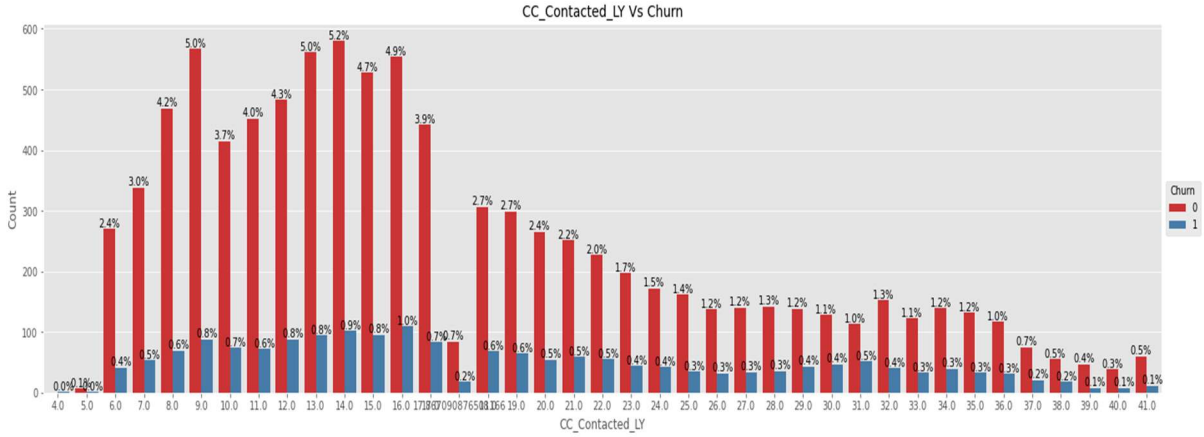

FIGURE 7 CHURN VS CC_CONTACTED_LY

Insight: Those customers who contacted the Customer care last year are more registered the more churn of customers. Straight meaning of it is the Customer care are having casual approach and not able to resolve the customer complaints or quarries.

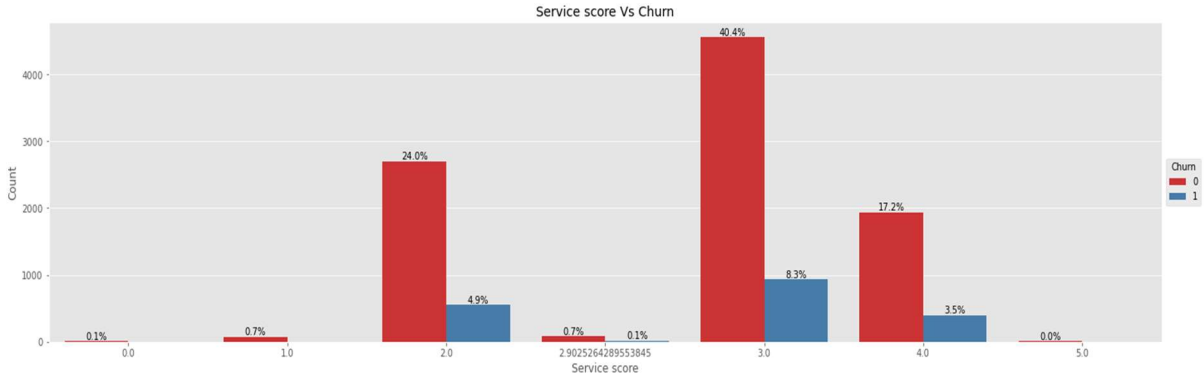

FIGURE 8 CHURN VS SERVICE SCORE

Insights: The customers rating the average (3)to the service score are churned more. That means they are not satisfied with customer care resolutions. While the rating service score below average(2) and above average has almost same number of customer churning. In final inference almost 40% of customers are availing customer service and 8% are churned. Seems some gap at end of customer service executives or agents.

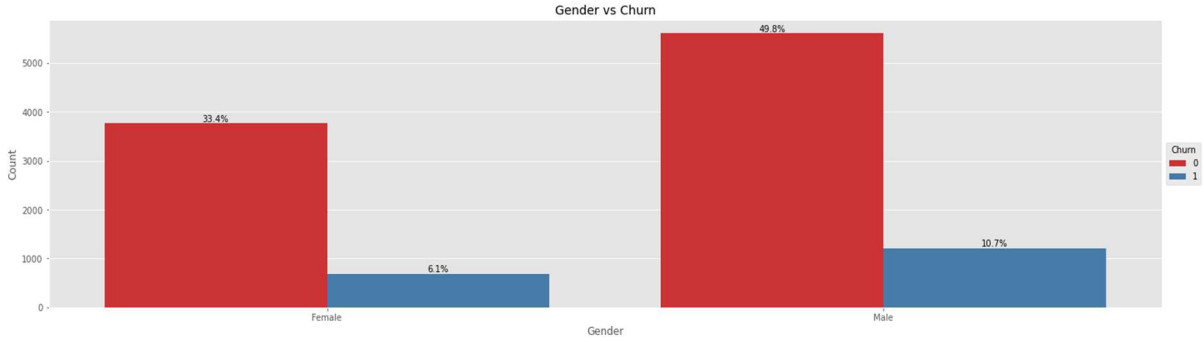

FIGURE 9 CHURN VS GENDER

Insight: Males customers are more than 60% and their churning is almost 11% while female churning is 6 percent only. Seems there is gap in content part of Male customers.

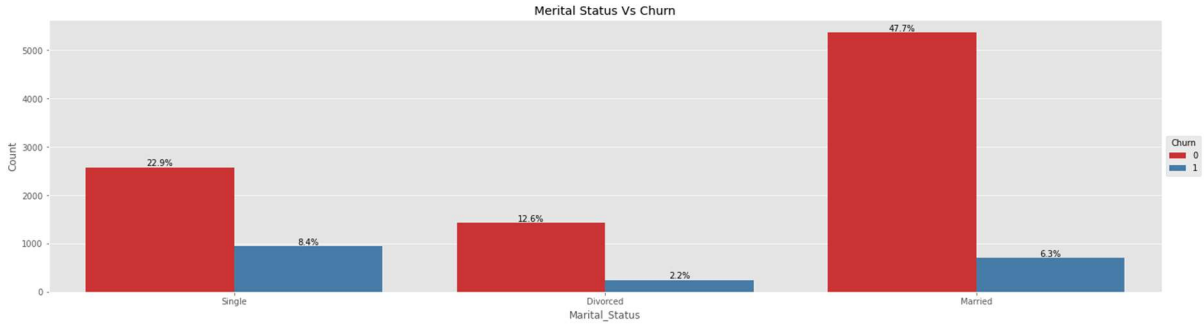

FIGURE 10 CHURN VS MARITAL STATUS

Insight: The customer with Single status has 8% churned and its more than married and forced single customers.

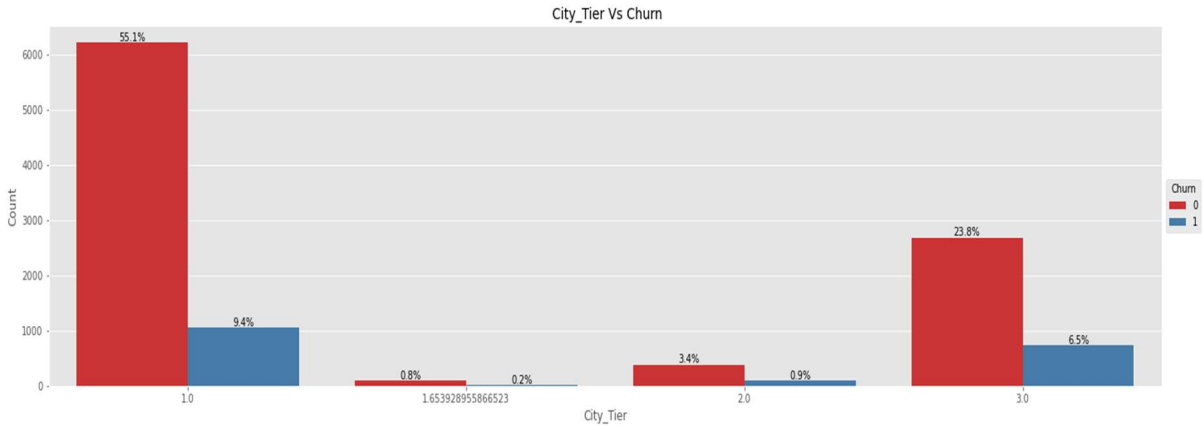

FIGURE 11 CHURN VS CITY TIER

Insight: The company has large base in city tier 1.Which has more than 55% of customers and churning rate of customers are highest to 9.4%. While class 3 tier has almost 30% of stake and churning is around6.5%.Seems the awareness of schemes and benefits are more in city tier 1 customers over other class.

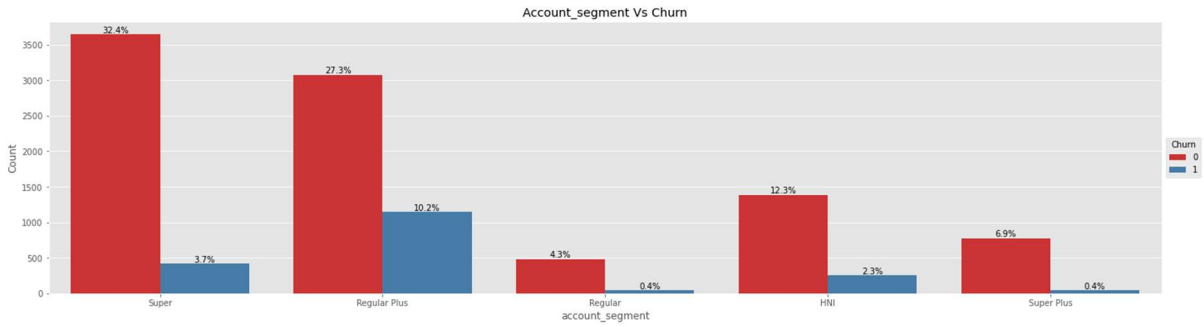

FIGURE 12 CHURN VS CUSTOMER SEGMENT

Insight: Company has good base of Super plus customer almost 37%, followed by Regular which are 35%Aprox.From Both the premium segments the erosion is highest for Regular segment 10%.Which means these customers feeling some missing on their level of satisfaction to stay.

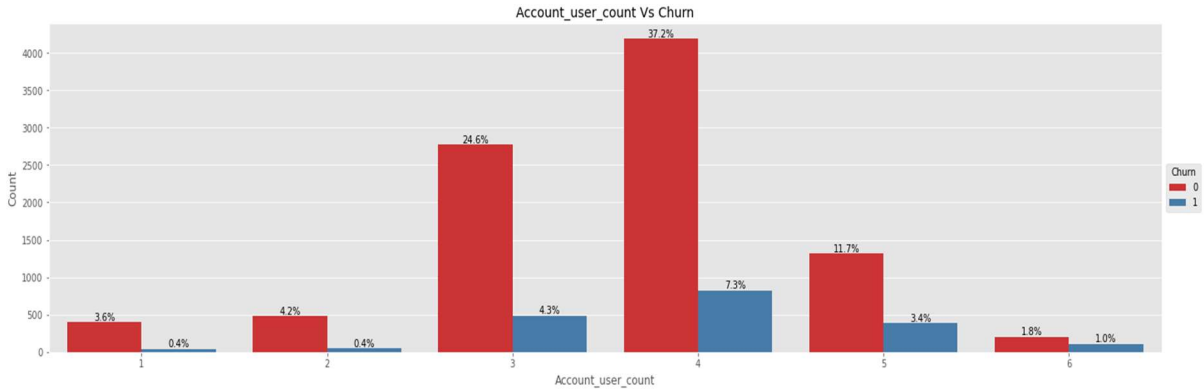

FIGURE 13 CHURN VS ACCOUNT_USER_COUNT

Insight : User account 4 has maximum number of customers aprox 4200 at tagged here and churning is highest in this category nearly 7%, followed by category 3 which is 4% and Account segment 5 with 3.4%.

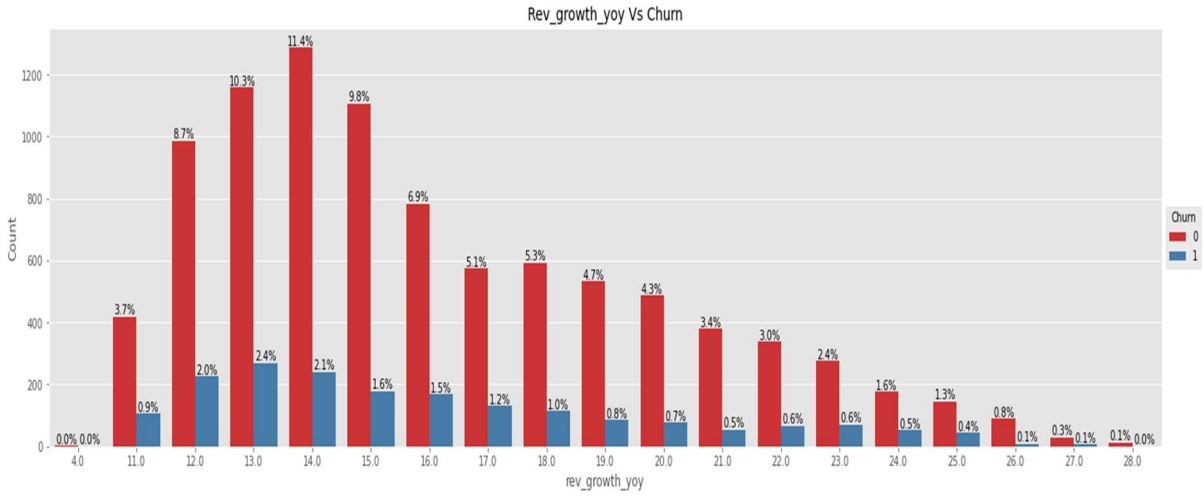

FIGURE 14 CHURN VS REVENUE GROWTH

Insight: Highest Churn 10% among the customers who are giving more revenue to the company

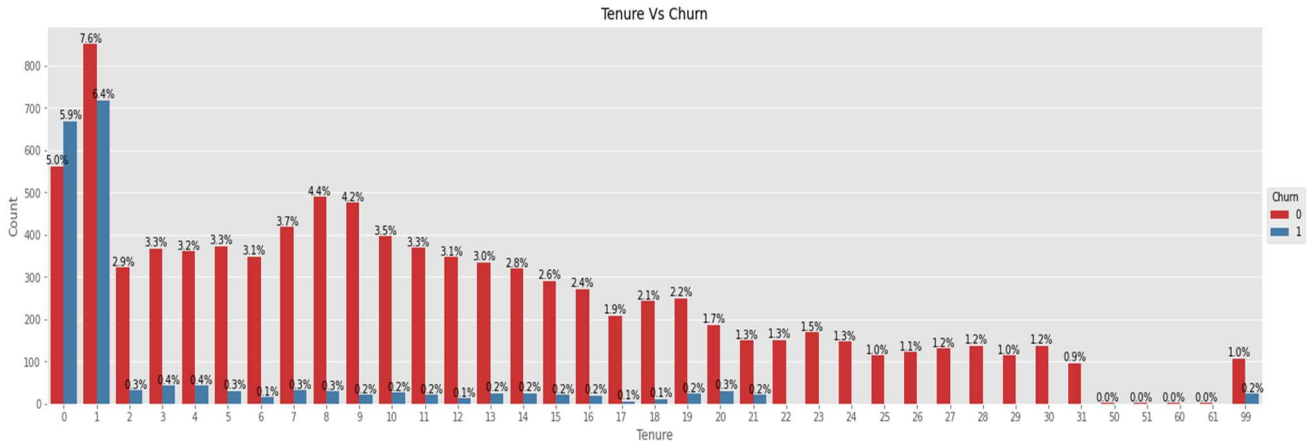

FIGURE 15 CHURN VS TENURE

Insights: The Customer in period of 0 to 1 has highest 11% of churning, seems the Incoming customers are not being care properly. While stability can be after the 2 and more.

Multi Variate Analysis

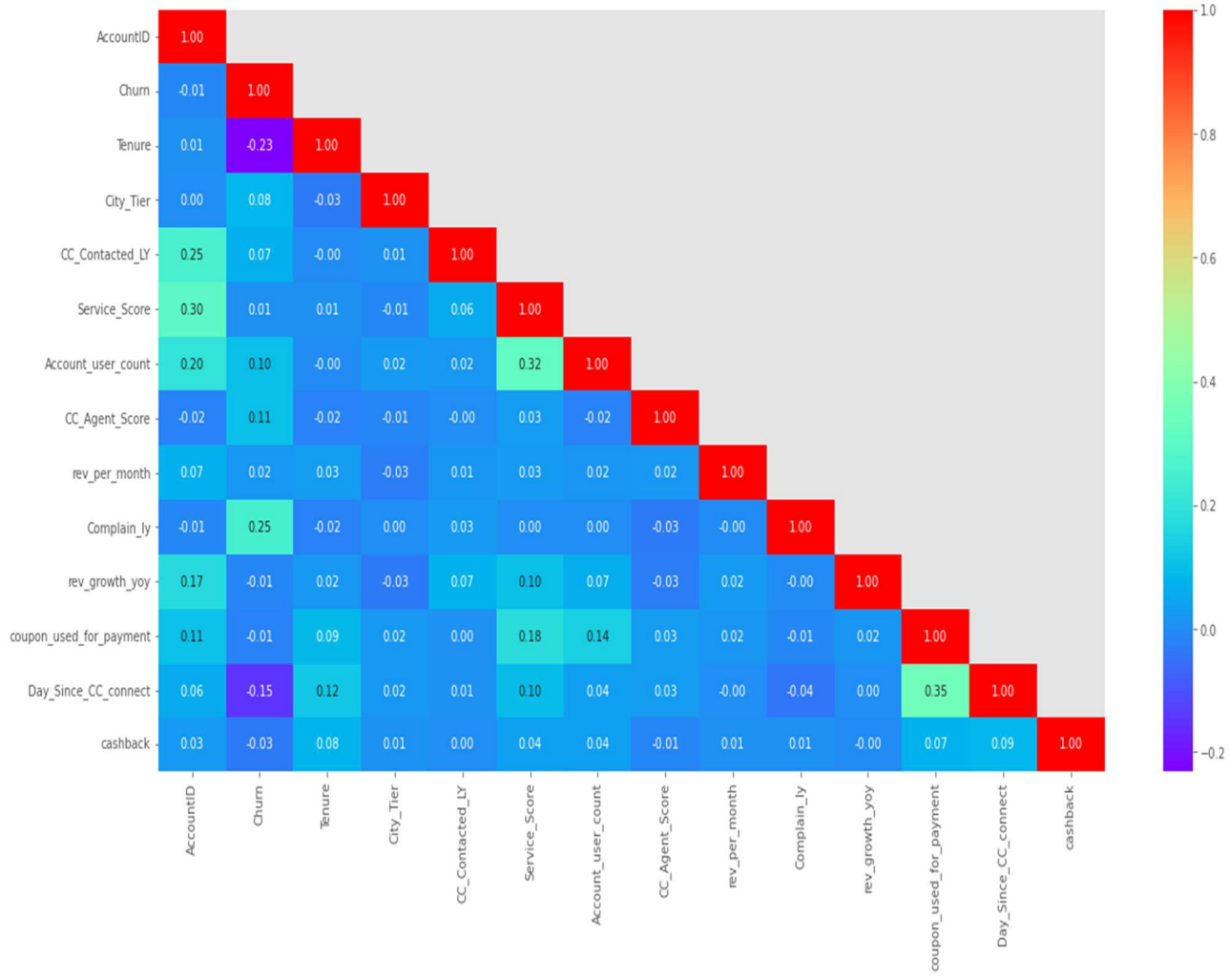

FIGURE 16 MULTIVARIATE ANALYSIS

Insights:

Multivariate Analysis for Churn Prediction: Key Feature Insights

In the pursuit of building a robust churn prediction model, a multivariate analysis of the most important features relative to the Churn target is pivotal. This analysis delves into the interplay of multiple variables and their collective impact on predicting customer churn. Here are the significant feature insights extracted from the provided data:

Strongest Predictors of Churn:

Tenure: Customer tenure exhibits a significant negative correlation with churn. Longer-tenured customers are less likely to churn, highlighting the importance of nurturing long-term relationships.

Positive Impact on Churn:

Complain_ly: Customers with a higher complaint rate are more likely to churn. Addressing and resolving customer complaints could potentially contribute to retention efforts.

CC_Agent_Score: A higher customer care agent score corresponds to a slight increase in churn probability. This suggests that improving agent interactions might positively influence retention strategies.

Account_user_count: As the number of users associated with an account increase, there's a slight rise in churn likelihood. Tailoring strategies to engage multiple users within an account could aid in retention.

Neutral or Mild Impact on Churn:

City_Tier: The city tier's influence on churn is moderate. Analysing city-specific trends could uncover insights into regional customer behaviour and guide targeted retention approaches.

CC_Contacted_LY: Customer interactions with customer care in the last year exhibit a mild connection to churn. Carefully examining the nature of these interactions could provide insights into customer sentiments and expectations.

rev_per_month: Revenue per month has a relatively subtle effect on churn. Strategizing around revenue enhancement might contribute to retention efforts.

Negative Impact on Churn:

Service_Score: A higher service score is associated with a slight decrease in churn. Focusing on maintaining high service quality can potentially contribute to customer loyalty.

Cashback: Customers utilizing cashback tend to exhibit lower churn rates. Employing cashback incentives strategically might aid in fostering customer loyalty.

Day_Since_CC_connect: longer periods since connecting with customer care correspond to lower churn rates. Regular customer engagement and proactive outreach may play a role in customer retention.

Data Cleaning and Pre-processing

FIGURE 17 DATA OVERVIEW

Missing Values:

There is actual missing values are found in a dataset by using dedicated Python code .Apart from Null Values there are large number of JUNK values are detected in many variables.

We May go for deletion of row having low Junk values but it may lead to deletion of some valuable information from other variables. So decided to convert these Junk values as Null values and impute accordingly.

Missing Values Analysis			
Variables	**Actual**	**Junk Values**	**Total**
cashback	471	2	473
Day_Since_CC_connect	357	1	358
Complain_ly	357	0	357
Login_device	221	0	221
Marital_Status	212	0	212
CC_Agent_Score	116	0	116
City_Tier	112	0	112
Account_user_count	112	332	444
Payment	109	0	109
Gender	108	0	108
Tenure	102	116	218
CC_Contacted_LY	102	0	102
rev_per_month	102	689	791
Service_Score	98	0	98
account_segment	97	0	97
coupon_used_for_payment	0	3	3
rev_growth_yoy	0	3	3

TABLE 2 MISSING VALUES

Reason of Junk /Null Data Set: There are number of reasons but we can identify probable cause as Data Entry Errors, Inapplicable or Not Applicable, Inapplicable or Not Applicable, improper data transfer from one software to another, Human Error.

Imputation of Missing Values:

- The **Numerical** Variables are treated with **Mean** values because for right-skewed numerical variables, the mean is pulled towards the higher values by the long tail of the distribution. This makes the mean a suitable choice for imputation because it represents the central tendency of the data.
- For **categorical** variables, the **mode** (most frequent category) is often used for imputation. This is because it represents the category that occurs most frequently and is a reasonable estimate for the missing value. Using the mode helps maintain the distribution of the variable and ensures that the imputed values align with the observed pattern in the data.

Outliers

Outlier treatment is required for several important reasons, each contributing to the overall quality, reliability, and interpretability of data analysis and modelling. Treatment of outlier are essential Preserving Data Quality, Enhancing Statistical Significance, Improving Model Performance, Avoiding Misleading Conclusions, Enhancing Visualization, Meeting Assumptions of Statistical Tests, Reducing Noise, Preventing Errors in Decision-Making.

Outlier Removal: Enhancing Data Quality and Reliability

In the pursuit of refining data integrity and robust analysis, the identification and treatment of outliers hold paramount importance. Here's a comprehensive breakdown of the meticulous process we employed to cleanse our dataset of outliers:

Strategic Outlier Identification: To ensure data fidelity, we harnessed the potent Interquartile Range (IQR) method. This technique meticulously calculates the lower and upper thresholds for isolating outliers, thus bolstering the integrity of our insights.

Quartile Quartet: Anchored in statistical rigor, we extracted the first and third quartiles (Q1 and Q3) from our data. This pivotal step, executed through the quantile() method with carefully selected arguments [0.25, 0.75], equips us with the essential bedrock for outlier detection.

Boundaries of Exception: Armed with the quartile bedrock, we sculpted the upper range via an elegant calculation: Q3 augmented by a 1.5-fold amplification of the IQR. This discerning approach bestowed us with a solid boundary, beyond which outliers dared to tread.

Box Plot Wisdom: Harnessing the visual acumen of a box plot, we embarked on an insightful voyage to identify potential outliers. This graphical representation unveiled dots beyond the lofty upper threshold, hinting at the presence of these statistical deviants.

Dots of Distinction: Those daring dots, embellishing the realm beyond the upper quantile limit, ceremoniously unveiled the presence of outliers within the variable. They marked the inflection points of distinction, steering our gaze towards potential data anomalies.

Liberating Limits: Armed with upper and lower limits, we embarked on an audacious quest to liberate our dataset from the clutches of outliers. This proactive stride towards data refinement invigorates the credibility of our analyses.

A Pictorial Odyssey: As true champions of transparency, we encapsulated the tale of outlier treatment in a pictorial rendition. This visual narrative masterfully contrasts variables before and after their transformative outlier treatment, encapsulating the profound impact of our endeavours.

Through this meticulous process, we unfailingly safeguarded the sanctity of our dataset, enriching its reliability and fortifying the underpinnings of our analyses. In our unwavering pursuit of data excellence, outlier treatment emerges as a steadfast sentinel, upholding the integrity of our insights and laying the groundwork for resolute decision-making.

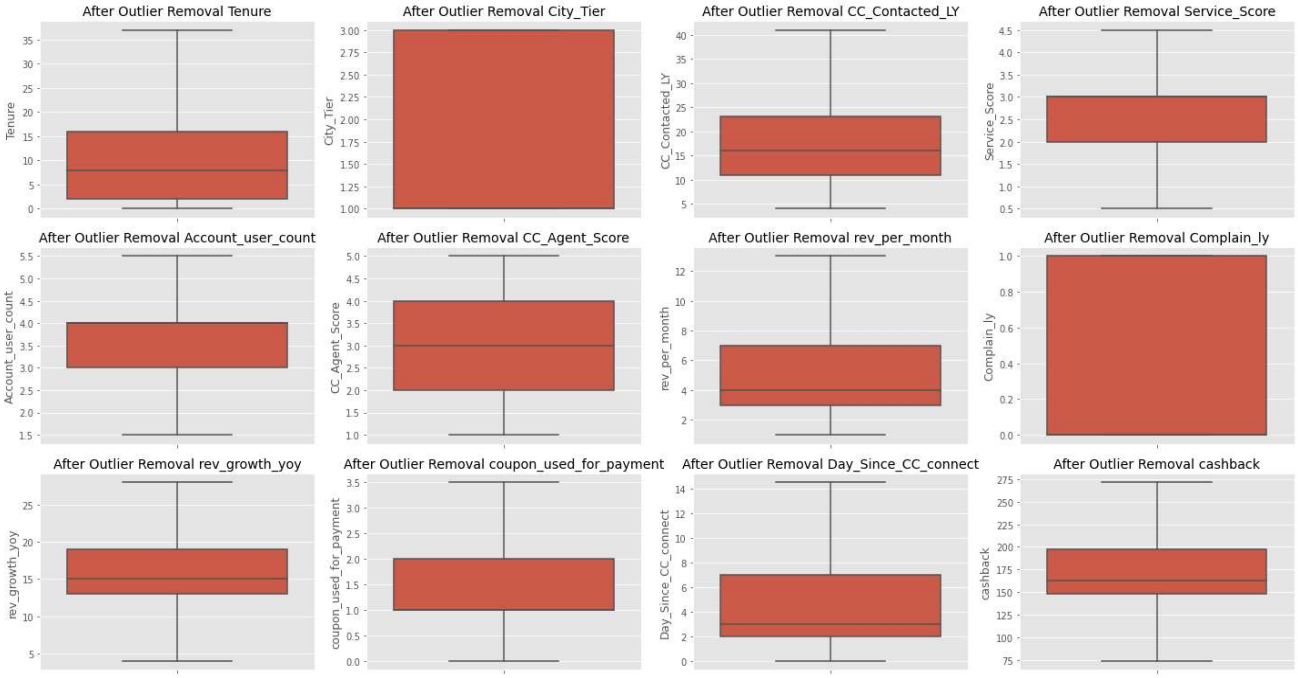

FIGURE 18 OUTLIER TREATED VARIABLES

Variable Addition or Removal:

There is No such requirement are arrived to create new variable or drop any variables. The Number o variables are less and NO case of Multicollinearity detected to do so.

Transformation of Variables

Scaling of data involves transforming variables so that they have a similar scale, making them comparable and ensuring that no variable dominates the analysis simply due to its larger magnitude. Here the variables like “Cashback, “City Tier” and “Complain Ly ”are at different magnitude Needs transformation to bring down on same scale.

We choose the Standard Scaling method for the standardisation and transformation of variables.

- Using Standard scaling because want to transform the data to have zero mean and unit variance.
- It is useful because the distribution of data is approximately Gaussian or want to compare features on the same scale.

Data Description after Varaible Transformation								
Variables	count	mean	std	min	25%	50%	75%	max
Tenure	11260	0	1	-0.84	-0.69	-0.22	0.4	6.87
City_Tier	11148	0	1	-0.71	-0.71	-0.71	1.47	1.47
CC_Contacted_LY	11158	0	1	-1.57	-0.78	-0.21	0.58	12.89
Service_Score	11162	0	1	-4	-1.24	0.13	0.13	2.89
Account_user_count	11260	0	1	-2.69	-0.7	0.29	0.29	2.29
account_segment	11260	0	1	-1.64	-0.74	0.15	0.15	1.94
CC_Agent_Score	11144	0	1	-1.5	-0.77	-0.05	0.68	1.4
rev_per_month	11260	0	1	-0.45	-0.27	-0.18	0.08	11.63
Complain_ly	10903	0	1	-0.63	-0.63	-0.63	1.58	1.58
rev_growth_yoy	11260	0	1	-3.25	-0.85	-0.32	0.75	3.14
coupon_used_for_payment	11257	0	1	-0.91	-0.4	-0.4	0.11	7.21
Day_Since_CC_connect	11260	0	1	-1.26	-0.71	-0.43	0.66	11.62
cashback	11260	0	1	-1.11	-0.27	-0.18	0.02	10.3
Gender_Male	11260	0	1	-1.24	-1.24	0.81	0.81	0.81
Marital_Status_Married	11260	0	1	-1.08	-1.08	0.92	0.92	0.92
Marital_Status_Single	11260	0	1	-0.67	-0.67	-0.67	1.48	1.48
Login_device_Mobile	11260	0	1	-1.47	-1.47	0.68	0.68	0.68
Login_device_Unknown	11260	0	1	-0.22	-0.22	-0.22	-0.22	4.46
Payment_Credit Card	11260	0	1	-0.67	-0.67	-0.67	1.49	1.49
Payment_Debit Card	11260	0	1	-0.85	-0.85	-0.85	1.18	1.18
Payment_E wallet	11260	0	1	-0.35	-0.35	-0.35	-0.35	2.87
Payment_UPI	11260	0	1	-0.28	-0.28	-0.28	-0.28	3.56
Churn	11260	0.17	0.37	0	0	0	0	1

TABLE 3 : TRANSFORMED DATA DESCRIPTION

Model Building Approach

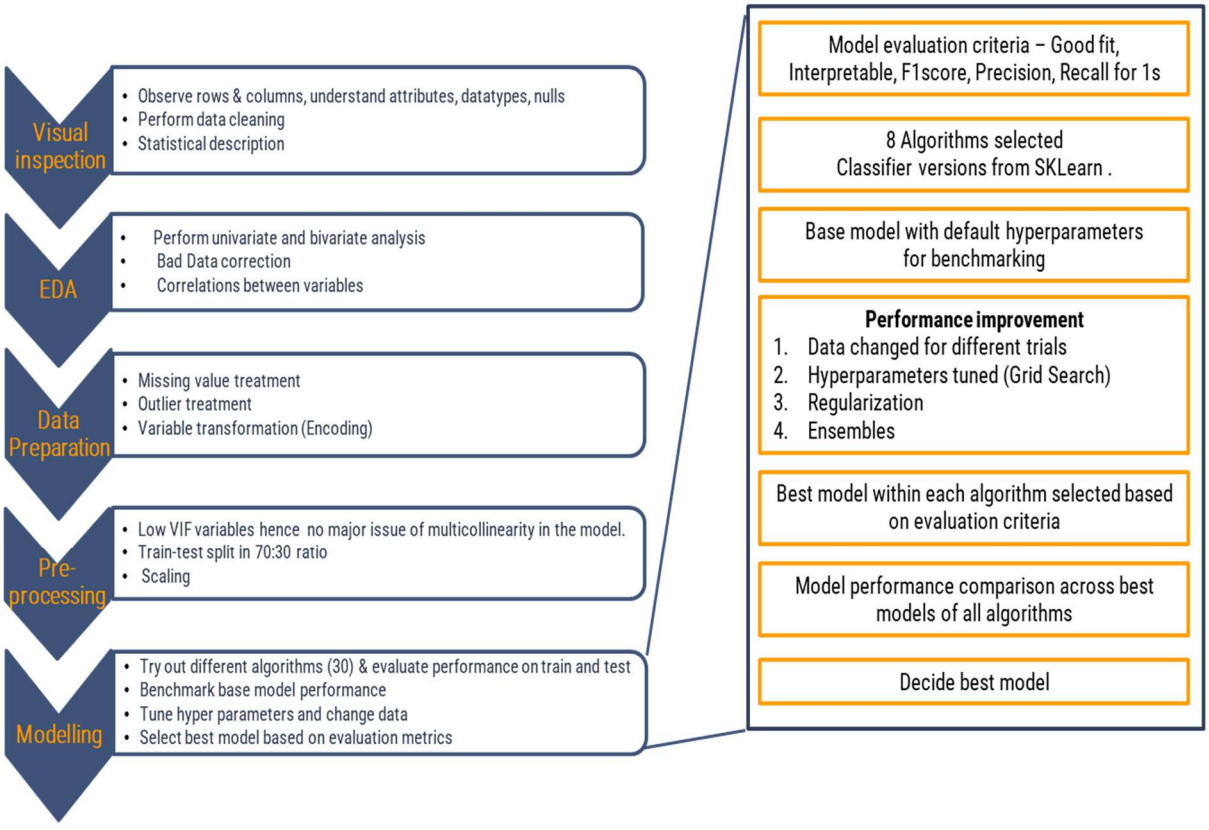

FIGURE 19 MODEL BUILDING APPROACH

Feature Selection

Multicollinearity: Certain features display a degree of multicollinearity, which means they are correlated with each other. For instance, "Cashback" and "Marital_Status_Married" exhibit negative correlations, possibly indicating that married customers might avail cashback offers more frequently. But there are such higher percentages of Multi collinearity.

```
Find most important features relative to Churn-target
Churn                        1.00
Complain_ly                  0.25
CC_Agent_Score               0.11
Account_user_count           0.10
City_Tier                    0.08
CC_Contacted_LY              0.08
rev_per_month                0.04
Service_Score                0.01
AccountID                   -0.01
rev_growth_yoy              -0.01
coupon_used_for_payment     -0.02
cashback                    -0.15
Day_Since_CC_connect        -0.15
Tenure                      -0.33
Name: Churn, dtype: float64
```

TABLE 4 MULTICOLLINEARITY

VIF : The VIF values for all variables are relatively low, indicating that there is no major issue of multicollinearity in the model. This means that the independent variables can be considered independent of each other, and their coefficient estimates are reliable for interpretation

	variables	VIF
20	Payment_Debit Card	3.33
19	Payment_Credit Card	3.12
21	Payment_E wallet	2.31
16	Marital_Status_Single	2.22
15	Marital_Status_Married	2.20
22	Payment_UPI	1.69
13	cashback	1.46
2	City_Tier	1.39
1	Tenure	1.38
0	AccountID	1.38
12	Day_Since_CC_connect	1.28
11	coupon_used_for_payment	1.28
23	Churn	1.27
4	Service_Score	1.26
8	rev_per_month	1.22
5	Account_user_count	1.18
18	Login_device_Unknown	1.14
17	Login_device_Mobile	1.14
3	CC_Contacted_LY	1.09
9	Complain_ly	1.07
10	rev_growth_yoy	1.04
6	account_segment	1.04
7	CC_Agent_Score	1.03
14	Gender_Male	1.01

TABLE 5 VIF

VIF values, we can infer the following:

Payment_Debit Card, Payment_Credit Card, and Payment_E wallet have VIF values around 3, indicating low to moderate multicollinearity. There may be some correlation among these payment-related variables, but it is not severe enough to significantly impact the regression model's performance.

Marital_Status_Married and Marital_Status_Single have VIF values around 2. These variables might have some correlation, but it is not strong enough to cause major multicollinearity issues.

Clus_kmeans has a VIF value of 2, suggesting a relatively low level of multicollinearity. This variable might have some correlation with other predictors, but it is not highly collinear.

Payment_UPI and Churn have VIF values around 1.7 and 1.6, respectively, indicating minimal multicollinearity with other variables in the model.

The majority of the remaining variables have VIF values around or below 1.5, indicating no significant multicollinearity among them.

Train Test Split: For the given business problem, 'Churn' is the target variable since the problem is to come up with a model to predict whether a particular customer will Churn or not. That is **Classification Problem.**

X – Independent variable (Removing 'Churn' variable)

y – Dependent/ Target variable (Having only 'Churn' variable)

1.The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

2.The train_test_split() method is used to split our data into train and test sets. Samples from the original training dataset are split into the two subsets using random selection. This is to ensure that the train and test datasets are representative of the original dataset.

```
Number of rows and columns of the training set for the independent variables: (7882, 23)
Number of rows and columns of the training set for the dependent variable: (7882,)
Number of rows and columns of the test set for the independent variables: (3378, 23)
Number of rows and columns of the test set for the dependent variable: (3378,)
```

TABLE 6 SHAPE OF TRAIN AND TEST DATA

SMOTE : : In this case, since the target variable is imbalanced with a ratio of 82.2 to 16.8, applying SMOTE could be a suitable approach to balance the classes and enhance the model's performance.

Clustering : Cluster analysis can be a compelling data-mining means for any organization that has to recognise discrete groups of customers, transactions, or other kinds of behaviours and things.

We have applied K-means clustering on the dataset and created 3 clusters. Each cluster is represented by the cluster number (Clus_kmeans) and shows the average values of different features within each cluster. Clusters 0, 1, and 2 have different characteristic profiles based on the features' average values within each cluster. The values within each cluster are standardized (normalized), which means that the average values have been scaled to have a mean of 0 and a standard deviation of 1. Therefore, the positive and negative values represent the deviation from the overall mean for each feature

Modelling

The given case is having variable "Churn" as target variable which is either "0" or "1" so we apply all the models of Classification to get the high accuracy in modelling as outcome.

We Tried multiple Classification models to get the highest Accuracy and other validation score to find the best fit model to predict the most accurate on Churning.

Here the top performing model details are depicted as Model comparison on different scores apart from Accuracy.

TOP performing Models Summary							
Model	**Hyperparameter**	**DataSet**	**Accuracy Score**	**Precision Score**	**Recall**	**AUC score**	**F1 Score**
RandomForestClassifierBag	Base Model of Bagging	Test	**0.99**	**0.97**	**0.99**	**0.99**	**0.99**
		Train	**1**	**1**	**1**	**1**	**1**
RandomForestClassifier	Base Model	Test	**0.99**	**0.97**	**0.99**	**0.99**	**0.99**
		Train	**1**	**1**	**1**	**1**	**1**
DecisionTreeClassifierGS	GridsSearch CV , Best model for F1 score	Test	**0.99**	**0.97**	**0.97**	**0.98**	**0.99**
		Train	**1**	**1**	**1**	**1**	**1**
DecisionTreeClassifier	Base Model	Test	**0.99**	**0.97**	**0.97**	**0.98**	**0.99**
		Train	**1**	**1**	**1**	**1**	**1**
RandomForestClassifierRS	Resampling Model, Best model for F1 score	Test	**0.96**	**0.93**	**1**	**0.96**	**0.96**
		Train	**1**	**1**	**1**	**1**	**1**
RandomForestClassifierRS	Resampling Model, Best model for F1 score	Test	**0.96**	**0.92**	**1**	**0.96**	**0.96**
		Train	**1**	**1**	**1**	**1**	**1**
DecisionTreeClassifierRS	Resampling Model, Best model for F1 score	Test	**0.96**	**0.92**	**0.99**	**0.96**	**0.96**
		Train	**1**	**1**	**1**	**1**	**1**
GaussianNBRS	Resampling Model, Best model for F1 score	Test	**0.96**	**0.93**	**0.99**	**0.96**	**0.96**
		Train	**1**	**1**	**1**	**1**	**1**

TABLE 7 MODEL PERFORMANCE COMPARISON

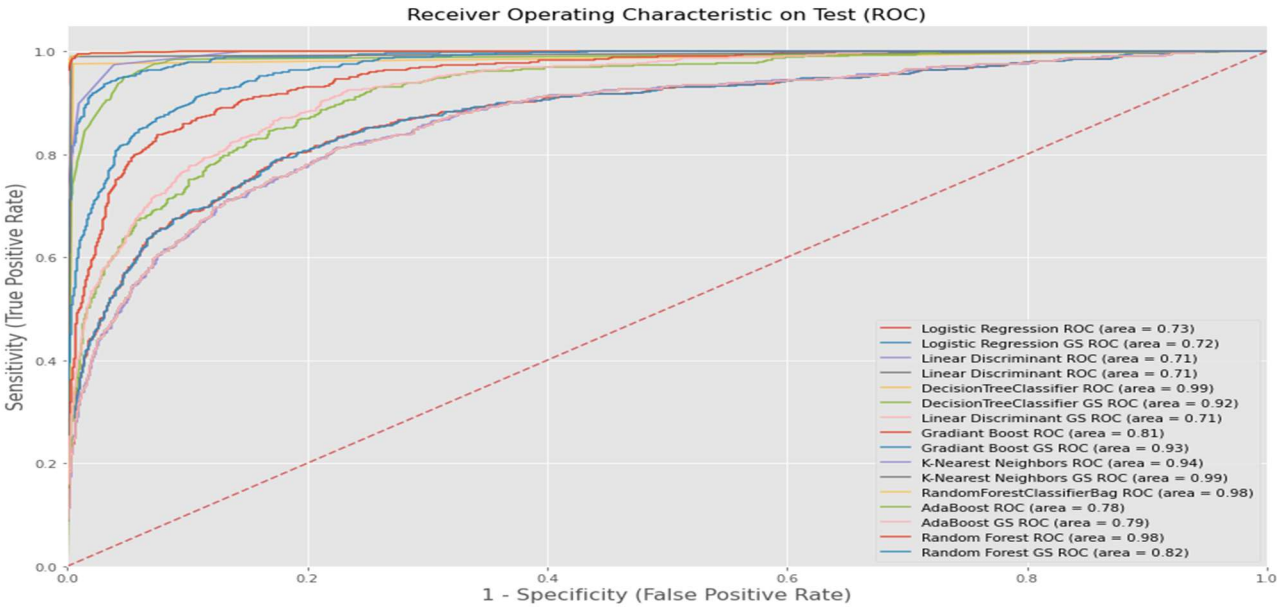

FIGURE 20 ROC FOR MODELS PERFORMED

Best Model Selection: Overall, the **Random Forest Classifier** stand on the top models based on their consistently high scores across multiple evaluation metrics. They demonstrate strong performance in terms of accuracy, precision, recall, AUC score, and F1 score, making them the best choices for classification tasks on the given test dataset.

Model	Hyperparameter	DataSet	Accuracy Score	Precision Score	Recall	AUC score	F1 Score
RandomForestClassifierBag	Base Model of Bagging	Test	**0.99**	**0.97**	**0.99**	**0.99**	**0.99**
		Train	**1**	**1**	**1**	**1**	**1**

TABLE 8 BEST MODEL

The model shown the high accuracy and performs well in classifying both classes (0 and 1). The model has high precision for both classes, indicating a low false positive rate. The recall for class 0 is perfect (1.00), indicating that the model correctly identifies all instances of class 0. The recall for class 1 is also high (0.97), indicating a good ability to identify positive instances. The F1-scores for both classes are high, indicating a good balance between precision and recall. The number of false negatives (15) is relatively low, suggesting that the model is effective in identifying positive instances correctly. Class 0 (negative instances) has a slightly larger support compared to class 1 (positive instances). Overall, the model shows strong performance in classifying both positive and negative instances. The high precision, recall, and F1-scores indicate that the model is reliable in distinguishing between the two classes

Advantages of Random Forest Classifier

1.Random Forests are known for their high accuracy and robustness. They tend to perform well across various datasets and handle a wide range of input features.
2.Random Forests can handle a large number of input features and can handle both numerical and categorical variables.
3.They are less prone to overfitting as they reduce variance by averaging predictions from multiple trees.
4.Random Forests provide feature importance rankings, which can help identify the most influential features in the mode

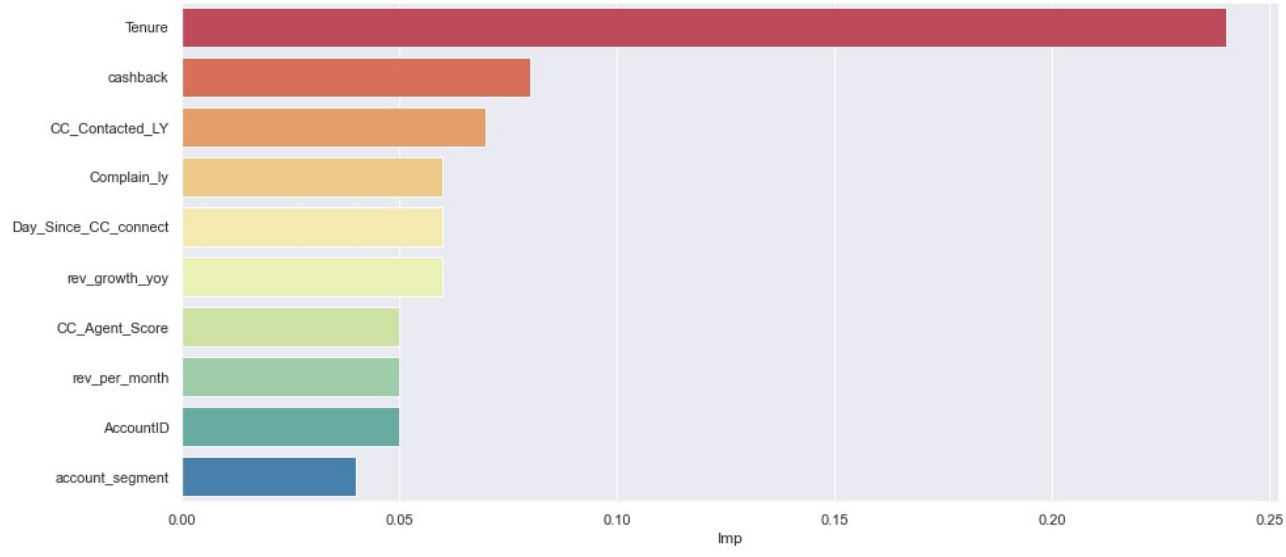

FIGURE 21FEATURE IMPORTANCE RFB

Insights of Feature Importance

To interpret the feature importance in terms of churning, we can analyse the percentage importance of each feature insights on how they relate to the likelihood of churn. Here's an interpretation based on the provided percentage importance values:

Tenure: This feature has the highest importance with 24.00%. It suggests that customers with longer tenure are less likely to churn. Longer tenure could indicate customer loyalty, satisfaction, or a higher level of engagement.

Cashback: With 8.00% importance, this feature suggests that offering cashback incentives may contribute to reducing churn. Cashback programs can incentivize customers to stay and engage with the company's offerings.

CC_Contacted_LY: The 7.00% importance of this feature implies that frequent customer care interactions in the past year may influence churn. Higher contact frequency might indicate dissatisfaction or issues that need resolution.

Complain_ly: With 6.00% importance, this feature suggests that customers who lodged complaints in the last year might be at a higher risk of churn. Addressing complaints effectively and promptly could help mitigate churn.

Day_Since_CC_connect: This feature, also with 6.00% importance, indicates that more recent customer care interactions may impact churn. Regular touchpoints with customers can help maintain engagement and satisfaction.

Rev_growth_yoy: The 6.00% importance of this feature suggests that positive year-on-year revenue growth may contribute to reducing churn. It implies that a thriving business and positive financial performance can attract and retain customers.

CC_Agent_Score: With 5.00% importance, this feature implies that customer care agent performance and satisfaction scores may influence churn. Well-trained and efficient agents can enhance the customer experience and reduce churn.

Rev_per_month: The 5.00% importance of this feature indicates that customers with higher monthly revenues are less likely to churn. It suggests that customers who spend more are more invested in the company's offerings.

These interpretations provide insights into the relationship between each feature and the likelihood of churn. They can guide business decisions and strategies focused on customer retention, such as improving customer care interactions, optimizing cashback programs, addressing complaints, and fostering customer loyalty through personalized engagement and revenue growth.

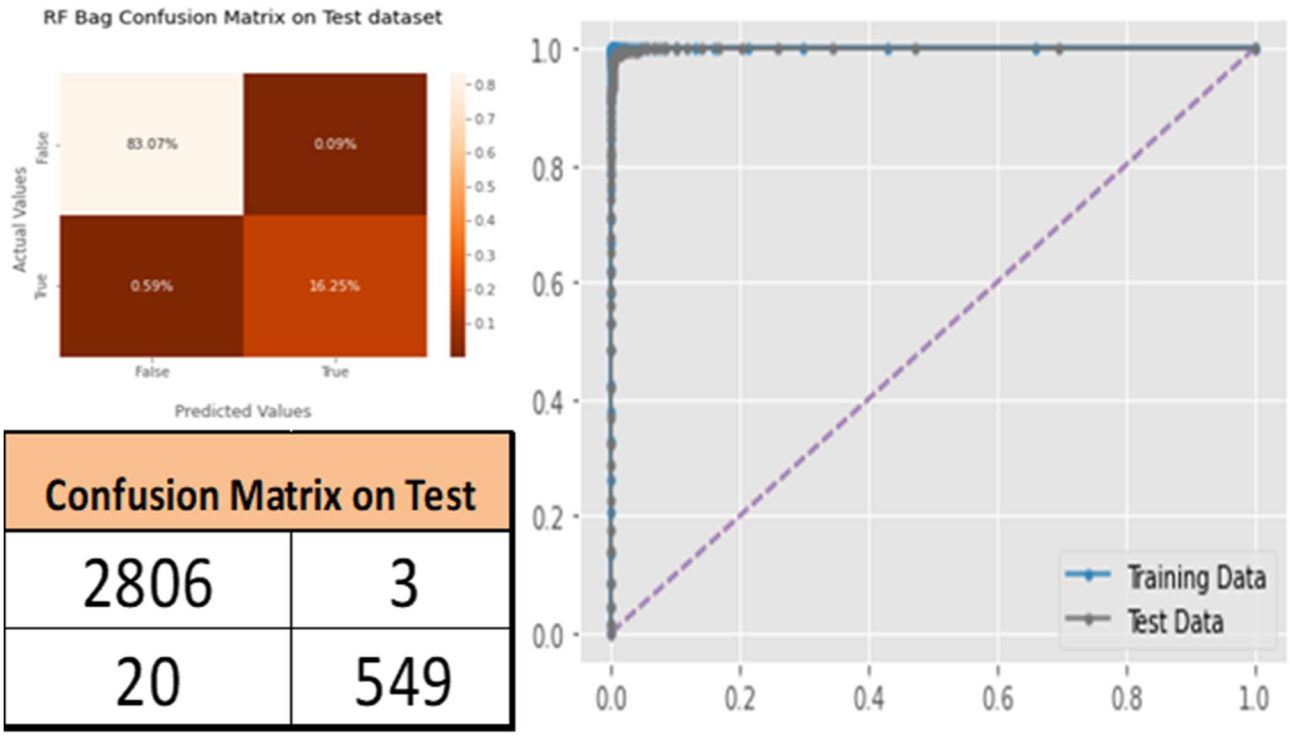

FIGURE 22 : AUC AND CONFUSION MATRIX OF RF

Insights of Model

On the basis of Performing the bagging on base of random forest classifier the following insights has been observed

True Positive (TP): 2806 This represents the number of instances that are actually positive (class 1) and are correctly predicted as positive by the model.

False Positive (FP): 3 This indicates the number of instances that are actually negative (class 0) but are incorrectly predicted as positive by the model.

False Negative (FN): 21 This represents the number of instances that are actually positive (class 1) but are incorrectly predicted as negative by the model.

True Negative (TN): 548 This indicates the number of instances that are actually negative (class 0) and are correctly predicted as negative by the model.

Classification Report: Precision, Recall, F1-score, Support, and Accuracy metrics are provided for each class (0 and 1), as well as macro and weighted averages.

Precision: Precision measures the model's ability to correctly identify positive instances out of the total instances predicted as positive. For class 0, the precision is 0.99, indicating that 99% of the instances predicted as class 0 are correct. For class 1, the precision is also 0.99, meaning that 99% of the instances predicted as class 1 are correct.

Recall: Recall (also known as sensitivity or true positive rate) measures the model's ability to correctly identify positive instances out of the actual positive instances. For class 0, the recall is 1.00, indicating that the model correctly identifies all instances of class 0. For class 1, the recall is 0.96, meaning that the model identifies 96% of the instances of class 1.

F1-score: The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall. For class 0, the F1-score is 1.00, indicating excellent performance. For class 1, the F1-score is 0.99, indicating very good performance.

Support: Support represents the number of instances in each class. For class 0, the support is 2809, and for class 1, the support is 569.

Accuracy: The overall accuracy of the model is 0.99, meaning that it correctly predicts 99% of the instances.

Effort to improve model performance.

Improving model performance is a crucial aspect of building effective predictive models.

Here are several strategies we have taken to enhance the performance of models:

Data Preprocessing: Handle missing values: Impute missing values using Mean and Mode methods .

Outlier treatment: Treated outliers that could negatively impact model performance by transforming them.

Handling Imbalanced Data: Addressed class imbalance: By techniques like oversampling, under sampling, or Synthetic Minority Over-sampling Technique (SMOTE) to balance class distribution.

Feature Scaling: Scale numerical features: Standardize the numerical features to ensure they have a similar scale.

Model Selection: Choose appropriate algorithms: Selected models that are well-suited for problem. Consider techniques like Decision trees, Random forests ,Ensemble models, LR, KNN, etc.

Hyperparameter Tuning: Fine-tune hyperparameters of chosen algorithms using techniques like grid Search CV .

Ensemble Methods: Combine models: Utilized ensemble techniques like bagging, boosting, which can lead to improved performance.

Cross-Validation: Use k-fold cross-validation: Validate model's performance using k-fold (5-Fold)cross-validation to ensure robustness and reliability

Model Validation

Model validation is a critical step in the process of building predictive models to ensure that the model performs well on unseen data and generalizes effectively. Validation helps assess the model's performance, identify potential issues like overfitting, and make informed decisions about model deployment

We use the Some Common Validation techniques like:

Train-Test Split: Divide the dataset into two parts: a training set and a testing (or validation) set. Train the model on the training set and evaluate its performance on the testing set. This helps assess how well the model generalizes to new, unseen data.

Metrics: For our we used common metrics include accuracy, precision, recall, F1-score, ROC curve, and AUC.

Model	Hyperparameter	DataSet	Accuracy Score	Precision Score	Recall	AUC score	F1 Score
RandomForestClassifierBag	Base Model of Bagging	Test	**0.99**	**0.97**	**0.99**	**0.99**	**0.99**
		Train	**1**	**1**	**1**	**1**	**1**

TABLE 9 : METRIC OF MODEL VALIDATIONS

Our main focus was on RECALL score, better the score better the prediction of target variable 1.

k-fold cross-validation: It involves splitting the dataset into k subsets (or folds) of approximately equal size. The model is then trained on k-1 folds and validated on the remaining fold. This process is repeated k5 times, each time using different fold as the validation set and the remaining folds as the training set.

[0.9923031379514505, 0.9937833037300178, 0.9937833037300178, 0.9928952042628775, 0.9940793368857312]

Mean of testing accuracy over 5 folds = 0.99 with std = 0.00 Table: K Fold Validation

TABLE 10 K FOLD VALIDATION OF RF

Grid Search and Hyperparameter Tuning: Hyperparameter tuning helps find the best configuration for model and can improve its performance.

RandomForestClassifier

```
RandomForestClassifier(max_depth=8, max_features=7, min_samples_leaf=2
,
                       min_samples_split=30, oob_score=True, random_state=0)
```

FIGURE 23 RF GRIDSEARCHCV

Ensemble Methods: Ensemble methods like bagging, boosting, and stacking can help improve models' performance by combining the predictions of multiple models.

Model Comparison Test dataset					
Models	Accuracy Score	Precision Score	Recall	AUC score	F1 Score
RandomForestClassifier Bag	0.99	0.97	0.99	0.99	0.99
GradientBoostingClassifier GS	0.92	0.64	0.83	0.88	0.92
GradientBoostingClassifier	0.92	0.64	0.83	0.88	0.92
AdaBoostClassifier GS	0.9	0.61	0.74	0.83	0.9
AdaBoostClassifier	0.89	0.59	0.72	0.82	0.89

TABLE 11ENSEMBLE FOR MODEL VALIDATION

Overall, our focus was to perform models and its tuning to get better and better Metrics include accuracy, precision, recall, F1-score, ROC curve, and AUC. While for Cross validation the k-fold cross validation methods have been performed with K=5.While other are also taken as reference for models' validations.

Final interpretation & Variable Based Suggestion

Tenure: The tenure of the customer is a key factor in predicting churn, with a high dependency of 30% to 51% weight across different models. Customers are more prone to churn during their early days of 0 to 5.

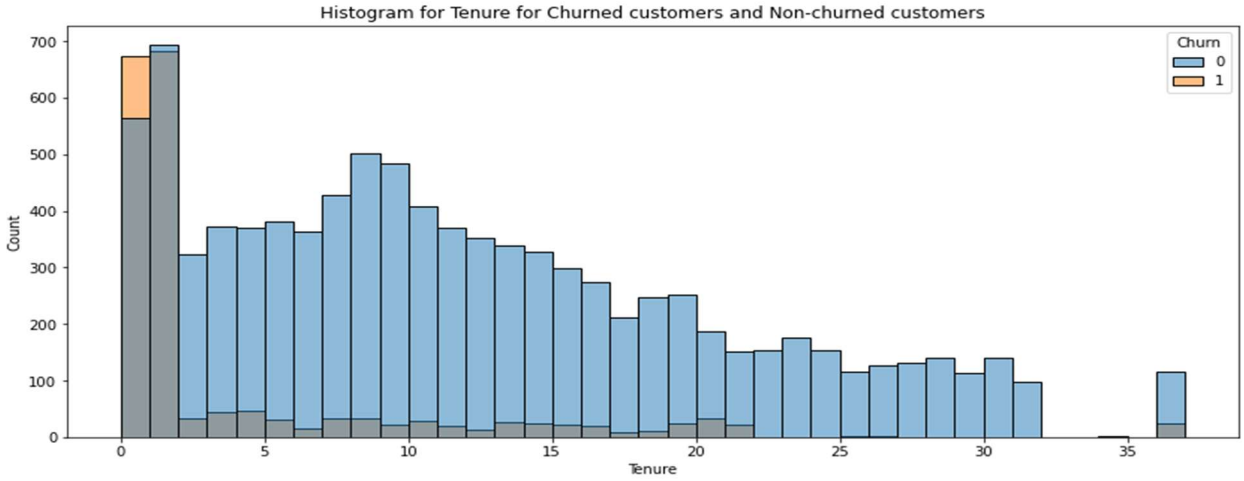

FIGURE 24 CHURN

To address this, the company should formulate policies and strategies to engage new accounts within this age group and provide them with incentives or personalized offers to increase their retention rate.

Customer Segment: Among the customer segments, the company has a significant base of Super customers, followed by Regular Plus customers, with approximately 3,100 customers in this segment. However, around 30% of these customers churn out of total churn of this variable

Indicating a need for targeted efforts to retain them. Implementing **promotional offers or exclusive benefits** for this segment can help in improving customer loyalty and reducing churn

City Tier: In terms of geographical distribution, the company has a large customer base in Tier 1 cities, with over 6,200 customers. However, around 1,100 customers from this segment churn.

To address this, the company should launch specific campaigns or initiatives in Tier 1 cities to retain accounts and prevent churn. Additionally, while Tier 3 and Tier 2 cities have fewer customers, they exhibit a similar churn trend to Tier 1, suggesting the need for retention strategies in these cities as well.

Account_user_count : Among the user account categories, User account 4 has the highest number of customers, approximately 4,200, with a churn rate of nearly 800. This is followed by categories 3 and 5. Interestingly, category 5 has the highest churn rate in terms of percentage, with approximately 25% of the tagged customers churning.

To address this, the company should focus on providing enhanced services and personalized attention to customers in these account categories to reduce churn.

Cashback: Specifically, customers within the cashback range of 125 to 150 and 150 to 175 are more prone to churn. This insight suggests that the Company should pay close attention to these cashback bands and implement strategies to reduce churn within these ranges. The company can consider offering additional incentives or exclusive cashback offers to customers falling within this range.

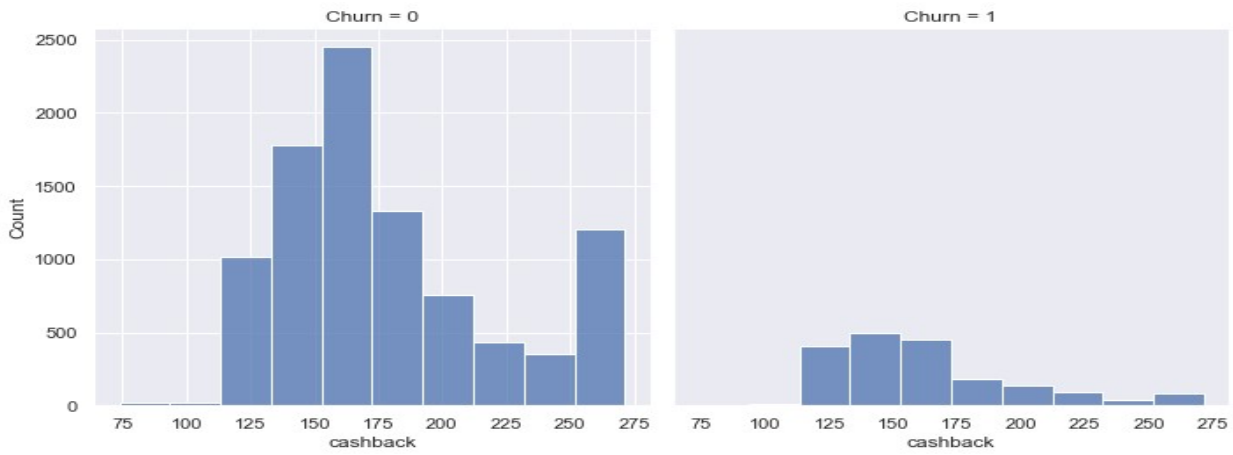

FIGURE 25 CASHBACK

Login Device: In terms of login device usage, approximately 6% of customers using either mobile or other devices churn. However, in terms of absolute numbers, customers using mobile devices for login have the highest turnover, with around 1,200 out of 6,200 customers churning.

The company should pay attention to providing a seamless and optimized experience for customers using mobile devices to reduce churn in this segment.it is most prevailing device in today's scenario.

Payment: It has come out in analysis threat the Debit card and Credit card payment having some issue probably technical one.

To address this issue, Possible reasons could include technical difficulties during the payment process, limited acceptance of credit cards, or customer preferences for other payment methods.

CC_Agent_Score :Customers who rated the company between 2.5 to 3.5 which can be treated as not good in today's scenarios were churned. It indicates that satisfaction levels is low and reflecting in candid rating scores.

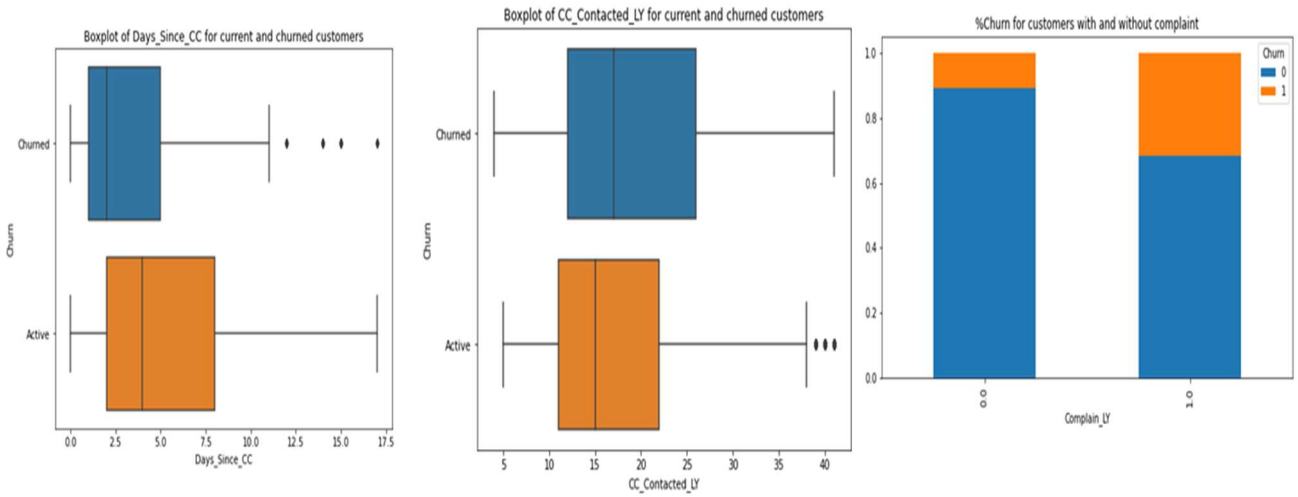

FIGURE 26 CUSTOMER CARE SERVICE

The company should focus on analysing the feedback type and reasons of complaint and its resolution address these concern account should be treated empathically and implement measures to address their concerns and improve overall customer satisfaction index.

Recommendation

Focus on Retention: The analysis highlights several variables that are crucial for improving customer retention. Tenure plays a significant role as per EDA and Models, with customers being more prone to churn during their early days of 0 to 3. The company should formulate policies and strategies to engage new accounts within this age group and provide personalized offers to increase retention. Additionally, focusing on customer segments with higher churn rates, such as Super customers, and implementing targeted efforts to retain them can significantly improve overall retention.

Personalized Campaigns: The findings indicate that different customer segments have varying churn rates. By leveraging this information, the company can design personalized campaigns tailored to each segment's needs and preferences. For example, offering promotional offers or exclusive benefits to Super customers or customers falling within specific cashback bands can enhance their loyalty and reduce churn. Personalized campaigns based on factors like city tier, account user count, or customer preferences can also contribute to higher engagement and retention.

Customer Engagement Program: To increase customer engagement, the company should focus on initiatives that enhance the overall customer experience. This can include providing seamless login experiences across different devices, optimizing mobile app functionality, and ensuring smooth payment processes for both debit and credit cards. Engaging customers through loyalty programs, rewards, and personalized communication can also strengthen their connection with the company and reduce churn.

Feedback and Issue Resolution: Customer feedback is crucial for identifying areas of improvement and resolving issues. By analysing customer ratings, particularly those falling in the 2.5 to 3.5 range, the company can gain insights into customer satisfaction levels and areas requiring attention. Implementing measures to address customer concerns and improve issue resolution can significantly impact customer retention. Regularly collecting feedback, conducting surveys, and proactively addressing customer grievances can demonstrate the company's commitment to customer satisfaction.

In summary, to improve retention, the company should focus on **personalized campaigns**, **customer engagement programs** on the basis of segmentation, **feedback analysis, and issue resolution**. By leveraging variables such as tenure, customer segments, cashback, city tier, account user count, login device usage, and payment methods, the company can tailor its strategies to enhance customer loyalty, drive engagement, and reduce churn we have certain constraints to tailor such campaigns.

Overall, by focusing on **engaging new accounts,** targeting specific customer segments, launching campaigns in key cities, providing enhanced services to high-churn account categories, optimizing the mobile experience, and addressing the concerns of customers with varying rating scores, the company can effectively reduce churn and improve customer retention

----------End of the Report----------