CAPSTONE PROJECT

CUSTOMER CHURN

FINAL PROJECT

DATA SCIENCE AND BUSINESS ANALYTICS

By: Sanjam Preet Singh Bhullar

September 2023

Tables of contents

List of figures

List of tables

PROBLEM STATEMENT

An e-commerce company or DTH (you can choose either of these two domains) provider is facing a lot of competition in the current market and it has become a challenge to retain the existing customers in the current situation. Hence, the company wants to develop a model through which they can do churn prediction of the accounts and provide segmented offers to the potential churners. In this company, account churn is a major thing because 1 account can have multiple customers. Hence by losing one account the company might be losing more than one customer.

You have been assigned to develop a churn prediction model for this company and provide business recommendations on the campaign.

Your campaign suggestion should be unique and be very clear on the campaign offer because your recommendation will go through the revenue assurance team. If they find that you are giving a lot of free (or subsidized) stuff thereby making a loss to the company; they are not going to approve your recommendation. Hence be very careful while providing campaign recommendation.

Data dictionary

Variable	Description
Account ID	Account unique identifier
Churn	Account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_LY	How many times all the customers of the account has contacted customer care in last 12months
Payment	Preferred payment mode of the customers in the account
Gender	Gender of the primary customer of the account
Service_Score	Satisfaction score given by customers of the account on service provided by company
Account_User_Count	Number of customers tagged with this account
Account_Segment	Account segmentation on the basis of spend
CC_Agent_Score	Satisfaction score given by customers of the account on customer care service provided by company
Marital_Status	Marital status of the primary customer of the account
Rev_Per_Month	Monthly average revenue generated by account in last 12 months
Complain_LY	Any complaints has been raised by account in last 12 months
Rev_Growth_YoY	Revenue growth percentage of the account (last 12 months vs last 24 to 13 months)
Coupon_Used_LY	How many times customers have used coupons to do the payment in last 12 months
Day_Since_CC_Connect	Number of days since no customers in the account has contacted the customer care
Cashback	Monthly average cashback generated by account in last 12 months
Login_Device	Preferred login device of the customers in the account

Problem Understanding

Defining problem statement

The dataset is about an e-commerce company that has been hit by a high customer churn rate. It is interested in predicting the churn rate and putting the brakes on it.

Put simply, customer churn is the percentage of customers who stopped availing a company's service or buying your business's products. The customer churn rate means how likely your existing customers are not going to make the next purchase from your business/store.

The basic premise is that the customer acquisition cost is always way higher than the customer service cost. So, the company try to keep the existing customers.

To reduce the customer churn rate, the e-commerce company wants to build a model that predicts the churn rate and draw invaluable insights from the historical data. In this manner, the company will be able to target its customers in a better manner and give segment-based offers to retain them.

As for the dataset that we have been provided with, it concerns supervised learning as we have the target column, Churn.

Need of the study/project

Given the growing competition in the market, it becomes imperative for any business to retain existing customers. The reasons for studying churn rate are as follows:

- Losing customers translates into decreasing revenue. Therefore, the focus of any business is to avoid high churn rate.
- Acquiring new customers is more difficult than retaining the existing ones. Besides, acquiring new customers entails cost and thus drives up the expenditure.
- On the contrary, retaining existing customers does not demand high cost. A loyal customer base adds to the esteem of a company that can boast of a loyal customer base. Therefore, examining the churn rate and identifying variables that impact it becomes important.
- In some cases, customers churn not on the basis of their dissatisfaction with the products of the company but on the basis of the poor customer relationship. In such situations, studying the churn also becomes imperative.
- High churn rates give an opportunity to the company to study its operations and that of their competitors. To stay in business, it is important to keep an eye on the rivals – what offers they are doling out and how they are keeping afloat in a competitive market.

Data report

Data sample

AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score
20000	1	4	3.0	6.0	Debit Card	Female	3.0	3	Super	2.0
20001	1	0	1.0	8.0	UPI	Male	3.0	4	Regular Plus	3.0
20002	1	0	1.0	30.0	Debit Card	Male	2.0	4	Regular Plus	3.0
20003	1	0	3.0	15.0	Debit Card	Male	2.0	4	Super	5.0
20004	1	0	1.0	12.0	Credit Card	Male	2.0	3	Regular Plus	5.0

Marital_Status	rev_per_month	Complain_ly	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback	Login_device
Single	9	1.0	11	1	5	159.93	Mobile
Single	7	1.0	15	0	0	120.9	Mobile
Single	6	1.0	14	0	3	NaN	Mobile
Single	8	0.0	23	0	3	134.07	Mobile
Single	3	0.0	11	1	3	129.6	Mobile

Table 1: First five rows of dataset

The dataset has **11,260 rows and19 columns**. In other words, it has 11,260 observations and 19 features.

There are **no duplicates**.

The first column, Account ID, is redundant and will be dropped in the data pre-processing stage.

The target column, Churn, has values 0 and 1. Class 0 signifies the customer has not churned, while Class 1 means that the customer has churned.

Data summary

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Churn	11260.0	NaN	NaN	NaN	0.168384	0.374223	0.0	0.0	0.0	0.0	1.0
Tenure	11042.0	NaN	NaN	NaN	11.025086	12.879782	0.0	2.0	9.0	16.0	99.0
City_Tier	11148.0	NaN	NaN	NaN	1.653929	0.915015	1.0	1.0	1.0	3.0	3.0
CC_Contacted_LY	11158.0	NaN	NaN	NaN	17.867091	8.853269	4.0	11.0	16.0	23.0	132.0
Payment	11151	5	Debit Card	4587	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	11152	2	Male	6704	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Service_Score	11162.0	NaN	NaN	NaN	2.902526	0.725584	0.0	2.0	3.0	3.0	5.0
Account_User_Count	10816.0	NaN	NaN	NaN	3.692862	1.022976	1.0	3.0	4.0	4.0	6.0
Account_Segment	11163	5	Regular Plus	4124	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CC_Agent_Score	11144.0	NaN	NaN	NaN	3.066493	1.379772	1.0	2.0	3.0	4.0	5.0

Marital_Status	11048	3	Married	5860	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Rev_Per_Month	10469.0	NaN	NaN	NaN	6.362594	11.909686	1.0	3.0	5.0	7.0	140.0
Complain_LY	10903.0	NaN	NaN	NaN	0.285334	0.451594	0.0	0.0	0.0	1.0	1.0
Rev_Growth_YoY	11257.0	NaN	NaN	NaN	16.193391	3.757721	4.0	13.0	15.0	19.0	28.0
Coupon_Used_For_Payment	11257.0	NaN	NaN	NaN	1.790619	1.969551	0.0	1.0	1.0	2.0	16.0
Day_Since_CC_Connect	10902.0	NaN	NaN	NaN	4.633187	3.697637	0.0	2.0	3.0	8.0	47.0
Cashback	10787.0	NaN	NaN	NaN	196.23637	178.660514	0.0	147.21	165.25	200.01	1997.0
Login_Device	11028	2	Mobile	7850	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 2: Five-point summary of dataset

It can be observed from the table that some variables have records fewer than 11,260, implying that such **features have null values**.

The difference between the 75th percentile and the maximum value of some of the numerical variables is huge. For example, the 75th percentile of Cashback is 200 whereas its maximum value is 1997. Take the case of Tenure. Its 75th percentile is 16, whereas its maximum value is 99. This means that such **variables have outliers**.

As for the categorical variables, Payment has five unique records with Debit Card occurring most of the times. Account_Segment also has five unique records, with Regular Plus having the maximum records. Login_Device has two records.

Data types

```
RangeIndex: 11260 entries, 0 to 11259
Data columns (total 19 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   AccountID                11260 non-null  int64
 1   Churn                    11260 non-null  int64
 2   Tenure                   11158 non-null  object
 3   City_Tier                11148 non-null  float64
 4   CC_Contacted_LY          11158 non-null  float64
 5   Payment                  11151 non-null  object
 6   Gender                   11152 non-null  object
 7   Service_Score            11162 non-null  float64
 8   Account_user_count       11148 non-null  object
 9   account_segment          11163 non-null  object
 10  CC_Agent_Score           11144 non-null  float64
 11  Marital_Status           11048 non-null  object
 12  rev_per_month            11158 non-null  object
 13  Complain_ly              10903 non-null  float64
 14  rev_growth_yoy           11260 non-null  object
 15  coupon_used_for_payment  11260 non-null  object
 16  Day_Since_CC_connect     10903 non-null  object
 17  cashback                 10789 non-null  object
 18  Login_device             11039 non-null  object
dtypes: float64(5), int64(2), object(12)
memory usage: 1.6+ MB
```

Table 3: Data types

Some of the numerical variables such as Tenure, Account_User_Count, rev_per_month, rev_growth_yoy, coupon_used_for_payment, Days_Since_CC_connect and cashback have object data type. This points to a discrepancy in the dataset. These columns should either be integer or float.

These columns have been typecast as object because these variables have special characters such as $, #, &, *, @ and +. For example, Account_User_Count has @, while Rev_Growth_YoY has $.

The special characters are bad data. In other words, it is missing data. The special characters must be replaced with null values.

Some of the column names need to be cleaned to bring uniformity.

After having replaced the special characters with null values and cleaning the column names, let us have another look at the data info.

```
RangeIndex: 11260 entries, 0 to 11259
Data columns (total 19 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   AccountID                11260 non-null  int64
 1   Churn                    11260 non-null  int64
 2   Tenure                   11042 non-null  float64
 3   City_Tier                11148 non-null  float64
 4   CC_Contacted_LY          11158 non-null  float64
 5   Payment                  11151 non-null  object
 6   Gender                   11152 non-null  object
 7   Service_Score            11162 non-null  float64
 8   Account_User_Count       10816 non-null  float64
 9   Account_Segment          11163 non-null  object
 10  CC_Agent_Score           11144 non-null  float64
 11  Marital_Status           11048 non-null  object
 12  Rev_Per_Month            10469 non-null  float64
 13  Complain_LY              10903 non-null  float64
 14  Rev_Growth_YoY           11257 non-null  float64
 15  Coupon_Used_For_Payment  11257 non-null  float64
 16  Day_Since_CC_Connect     10902 non-null  float64
 17  Cashback                 10787 non-null  float64
 18  Login_Device             11028 non-null  object
dtypes: float64(12), int64(2), object(5)
memory usage: 1.6+ MB
```

Table 4: Data types after imputation

If we omit Account ID, the dataset has five object variables and 13 numerical variables.

Of the 13 numerical variables, three – Churn, City_Tier and Complain_LY – are flag columns.

The target column, Churn, is a binary categorical variable. Class 1 (customer has churned) is a class of interest.

Exploratory Data Analysis & business implications

Univariate analysis

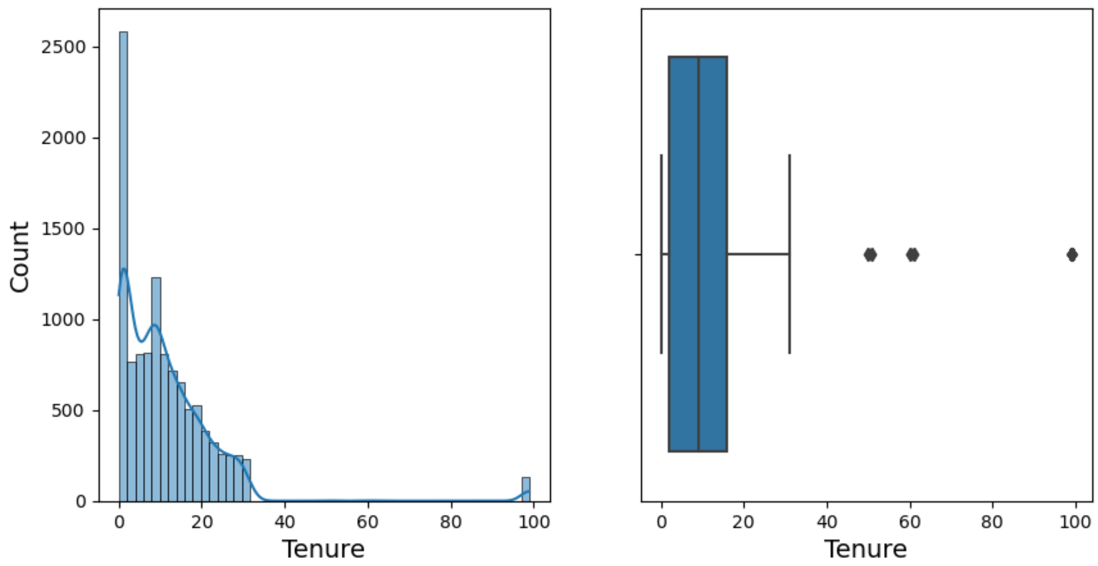


Figure 1: Histogram and boxplot for Tenure

The data is right-skewed and has a few outliers.

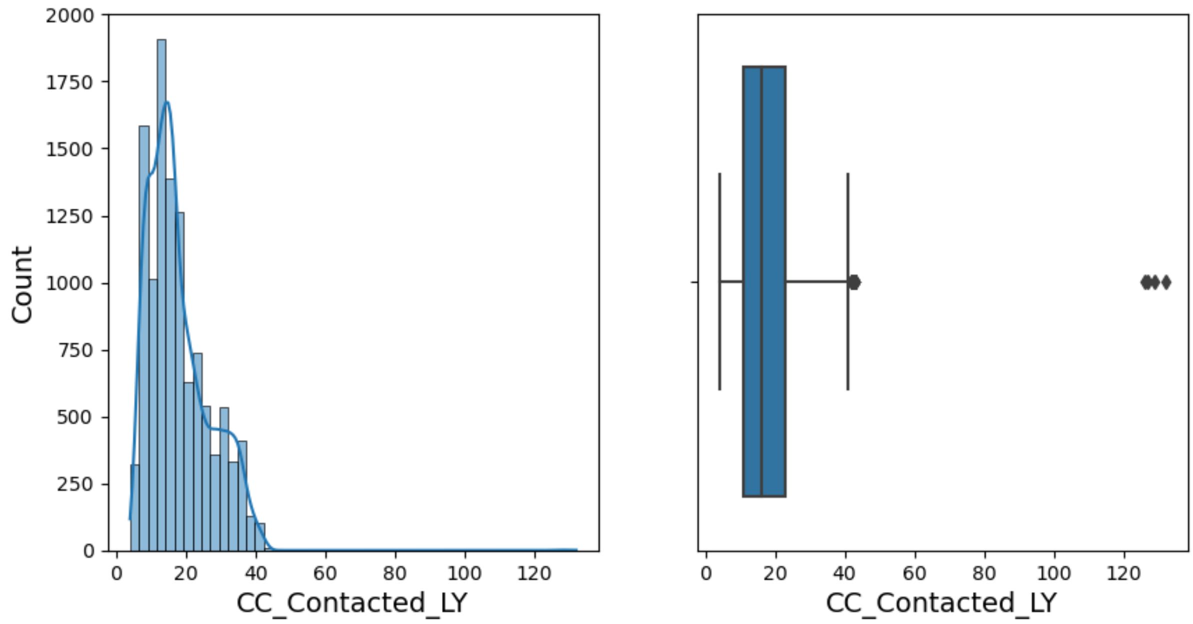


Figure 2: Histogram and boxplot for CC_Contacted_LY

The data is right-skewed and has a few outliers.

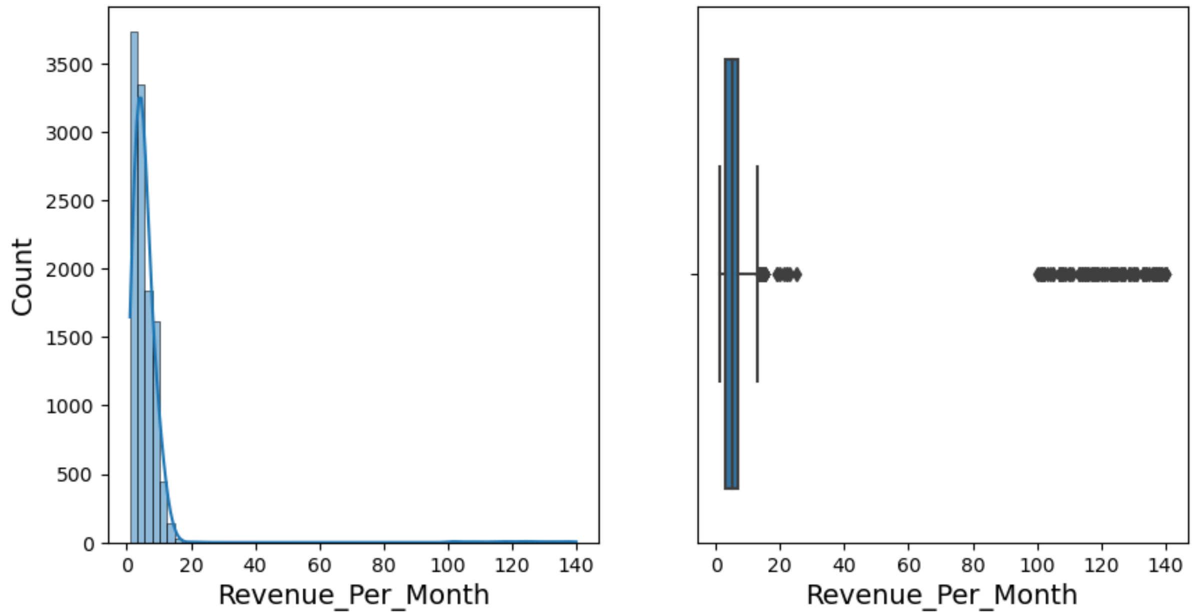


Figure 3: Histogram and boxplot for Revenue Per Month

The data for Revenue Per Month is highly right-skewed. It has a lot of outliers.

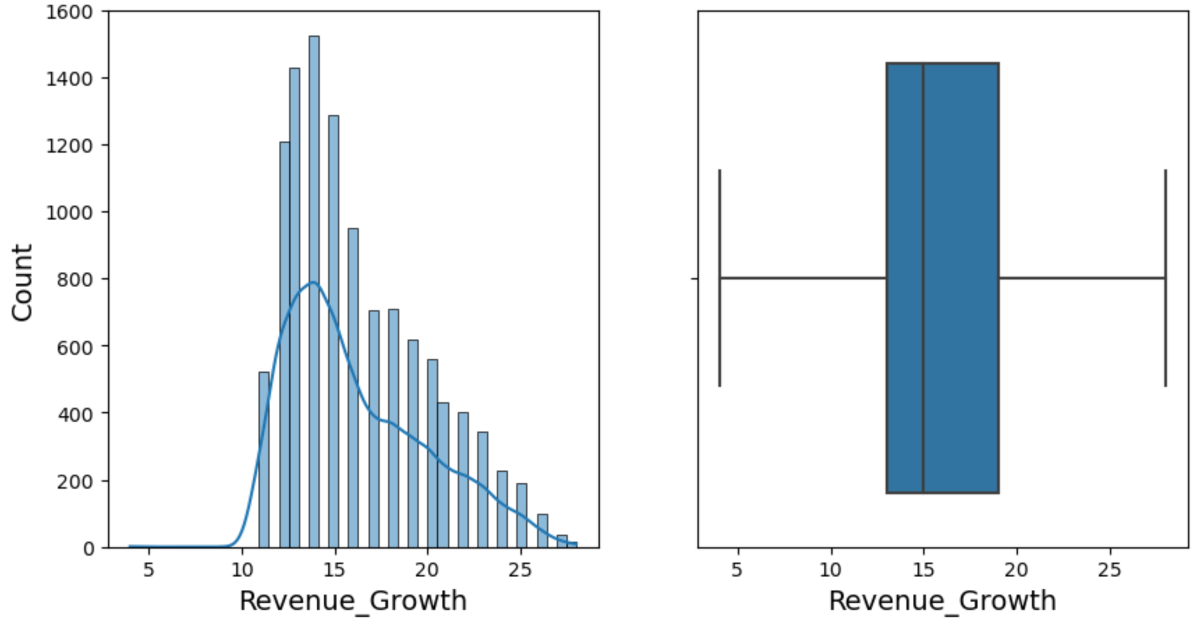


Figure 4: Histogram and boxplot for Revenue Growth

The data for year-on-year revenue growth is less skewed than other variables. The KDE plot gives the impression of a bell-shaped curve. This feature has no outliers.

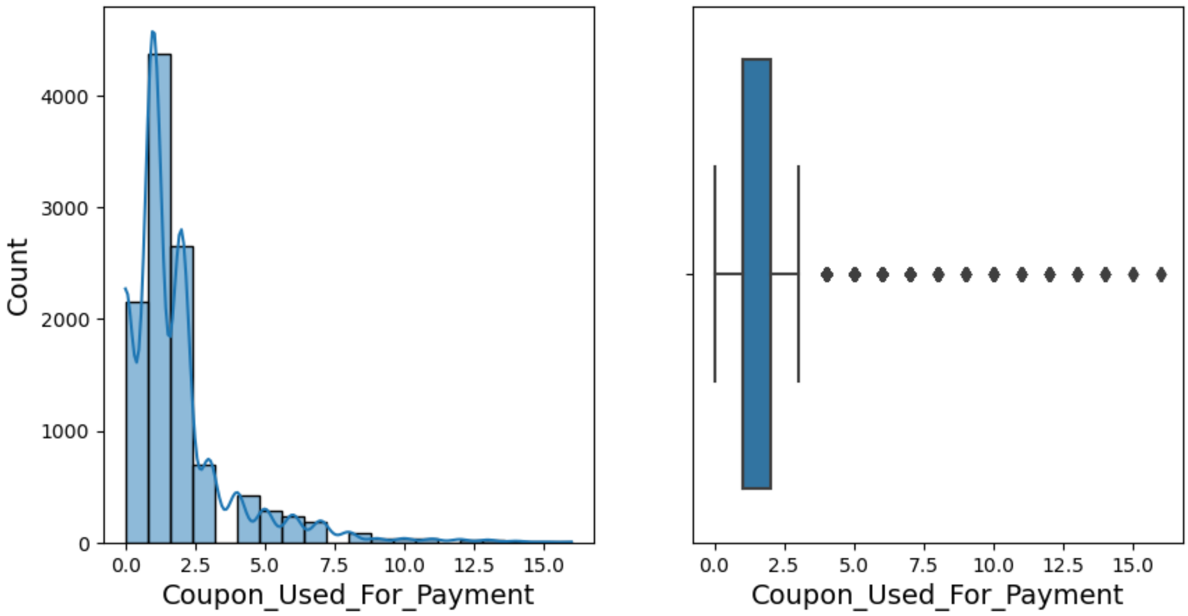


Figure 5: Histogram and boxplot for Coupon used for Payment

The data is right-skewed and has many outliers.

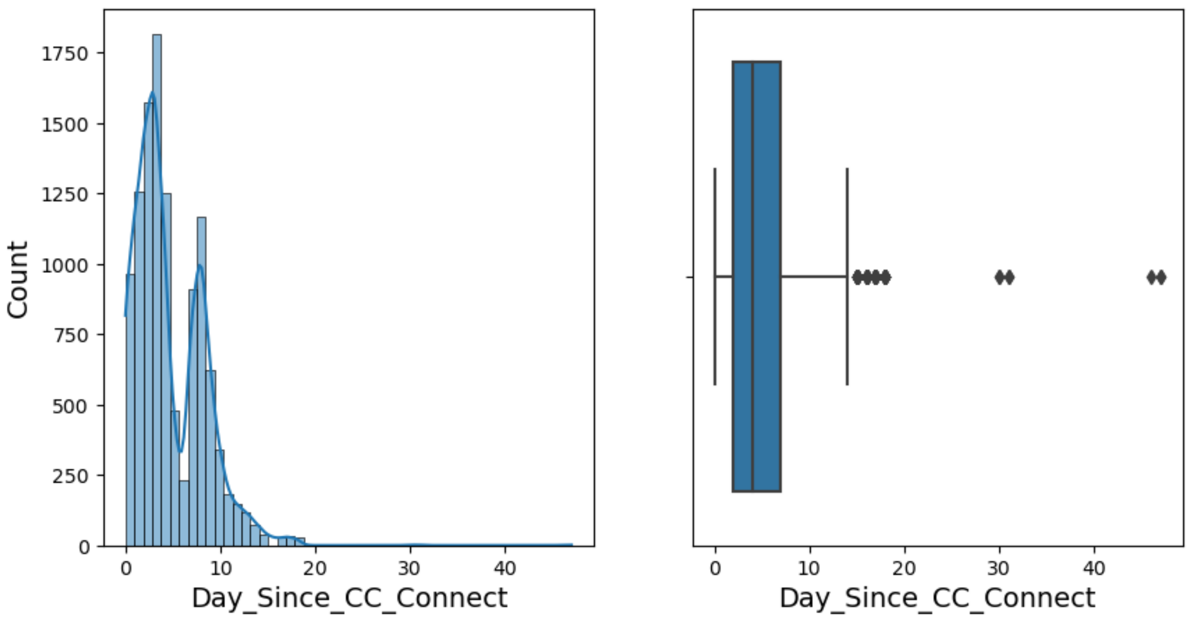


Figure 6: Histogram and boxplot for Day_Since_CC_Connect

Like other variables, this feature is no different. It is right-skewed and has a few outliers.

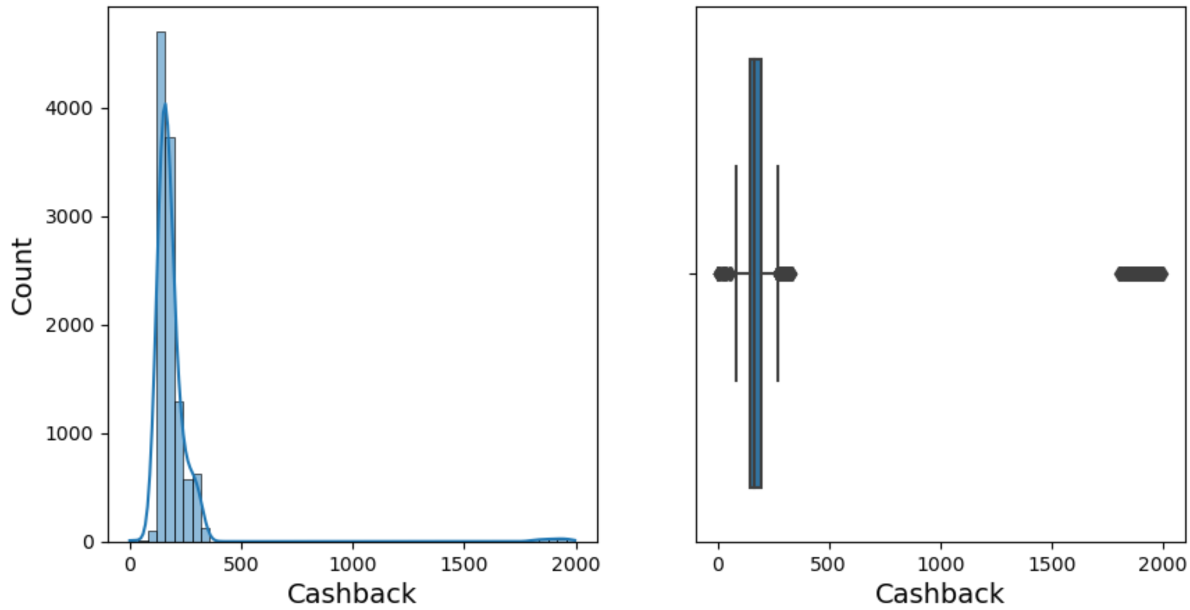


Figure 7: Histogram and boxplot for Cashback

The data for Cashback is also right-skewed. It has many outliers. In other words, the records in this feature have extremely high values.

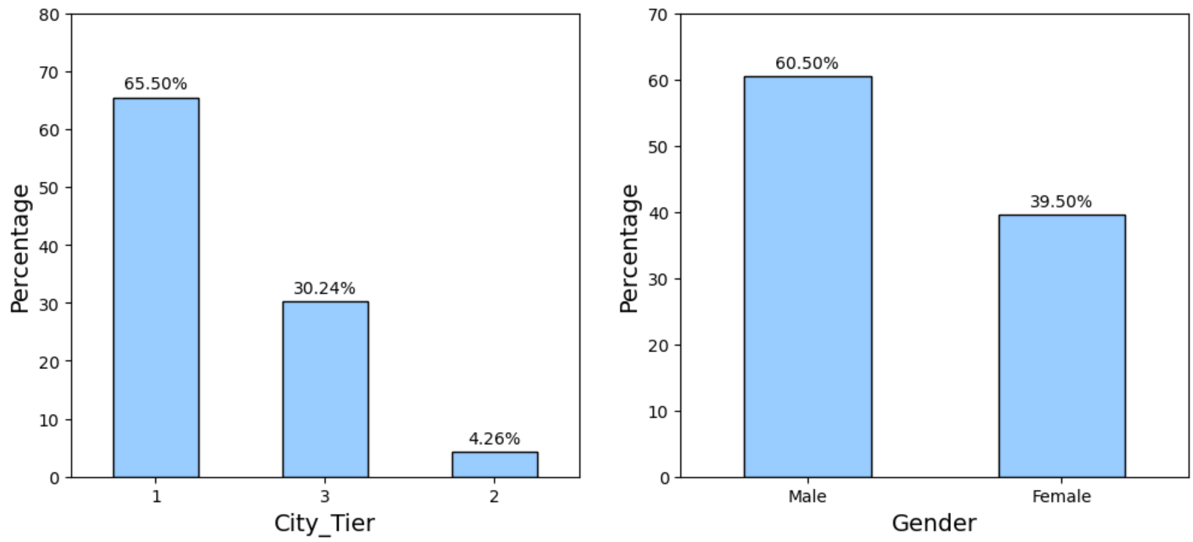


Figure 8: Bar charts for City Tier and Gender

Tier 1 cities have the highest records in the dataset followed by Tier 3 and Tier 2.

The number of men are more than that of women.

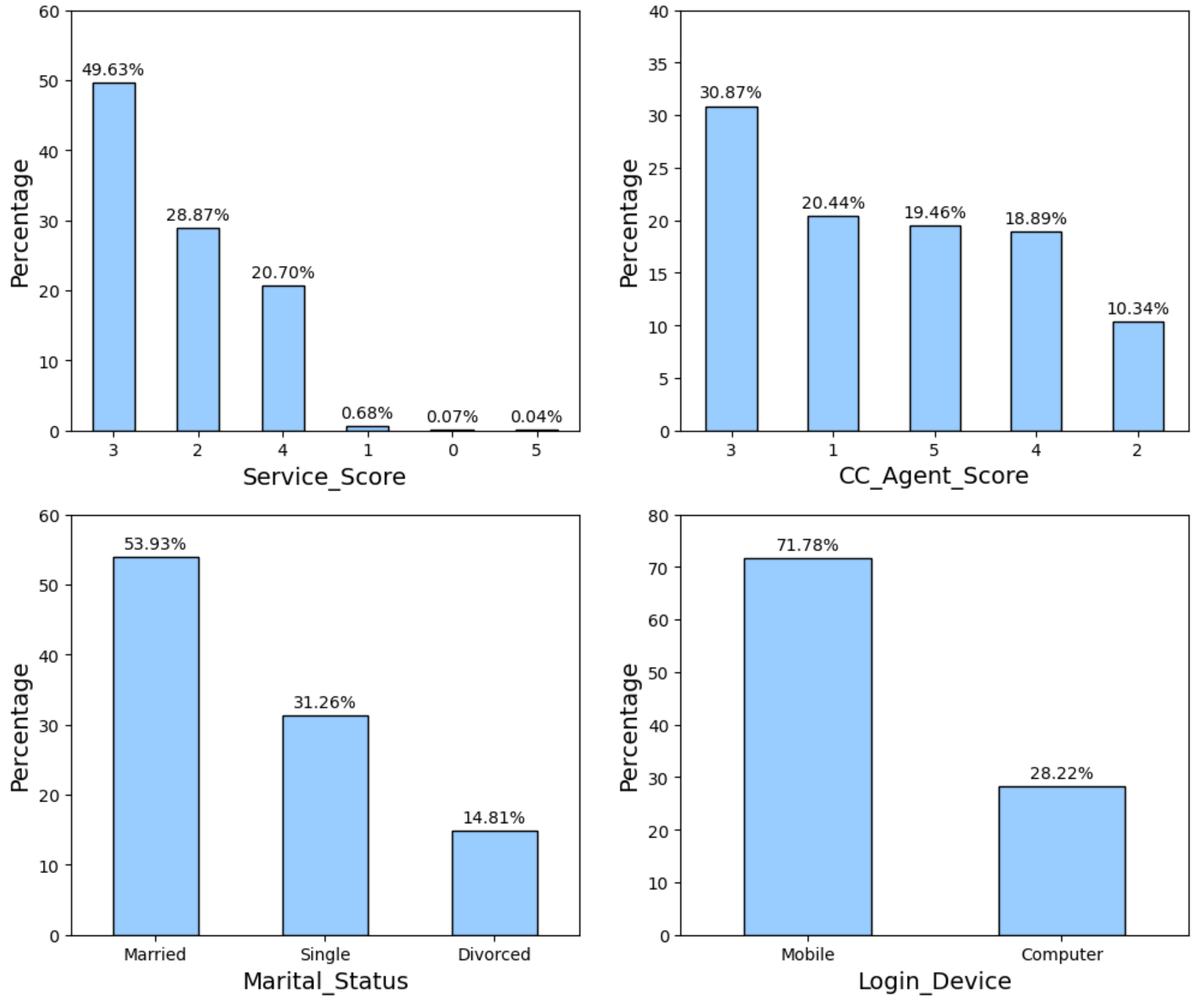


Figure 9: Bar charts for categorical variables

Most of the customers have given a service score of 3 to the e-commerce company followed by a score of 2. Only a few customers have given a rating of 0, 1 and 5 to the company.

The majority of the customers have given a score of 3 to the customer care agent followed by a score of 1. Less than 20 per cent of user have given high ratings of 4 and 5

As for the Marital Status, married persons dominate the dataset followed by single persons. Divorced persons have the least number.

When it comes to Login Device, most of the customers prefer mobile over computer.

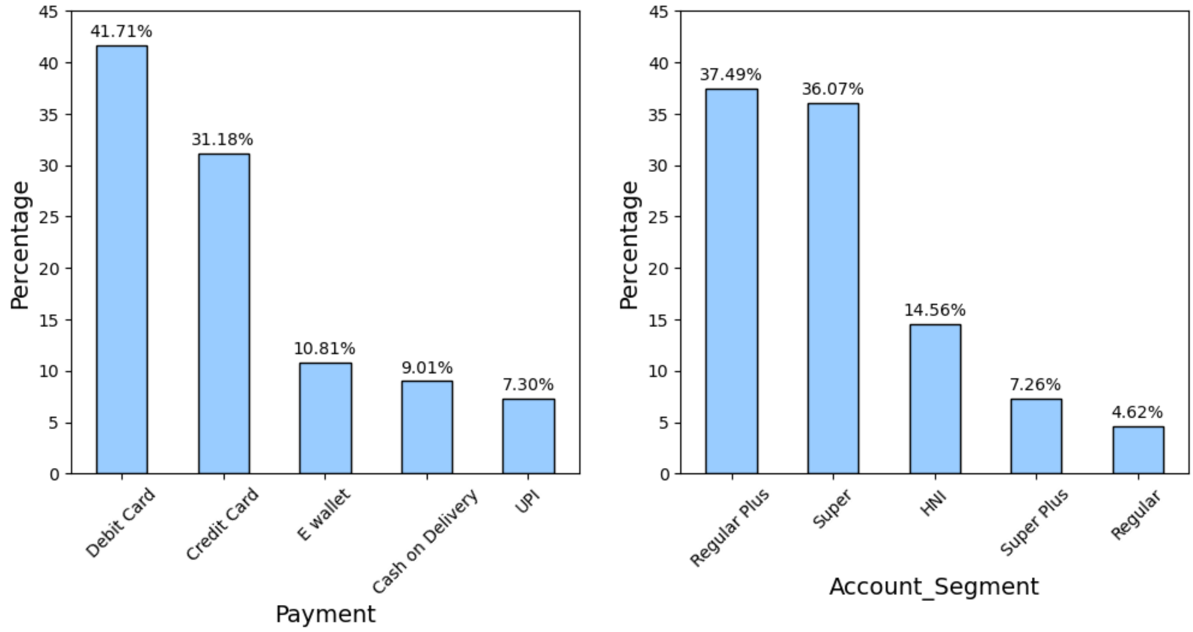


Figure 10: Bar charts for Payment and Account Segment

Debit card is the most preferred mode of payment followed by credit card and e-wallet. UPI is the least preferred mode of payment.

The majority of the customers prefer Regular Plus account followed by Super. The difference between the two account segments, however, is negligible. Regular account is the least preferred category.

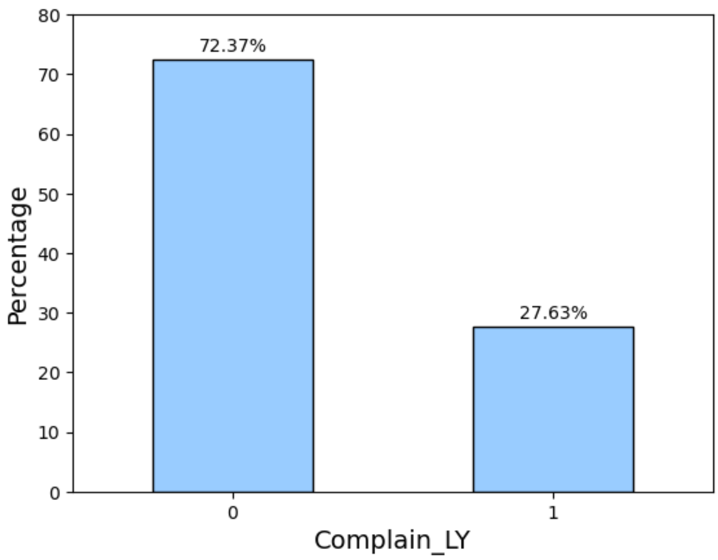


Figure 11: Bar chart for Complain_Last_Year

The majority of the customers have not complained (label 0) about the e-commerce company's service.

Bivariate analysis

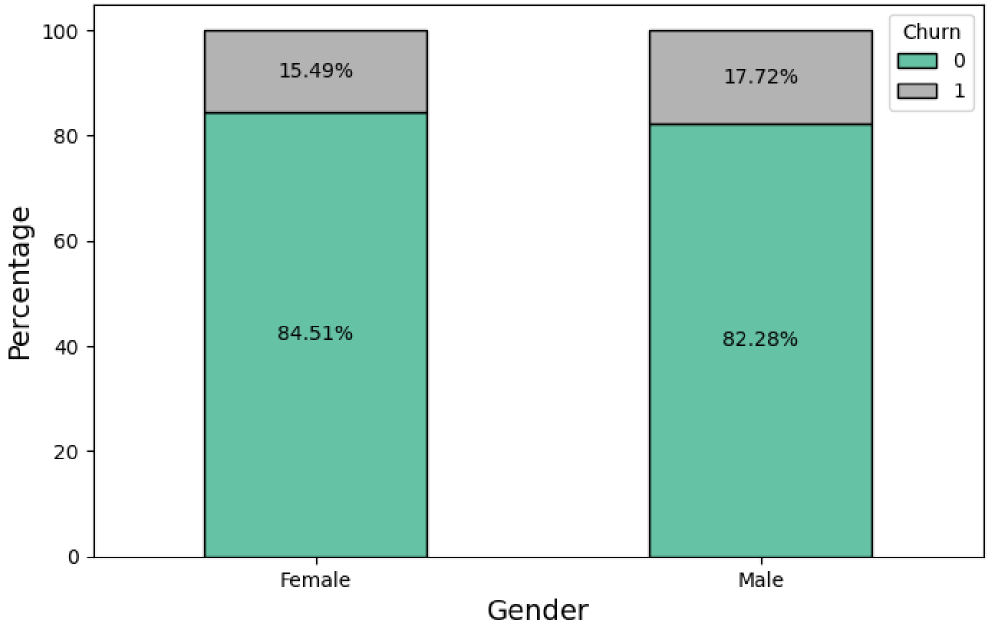


Figure 12: Churn rate by Gender

Men have a slightly more tendency to churn (label 1) than women. That being said, Gender does not have a significant predictive power.

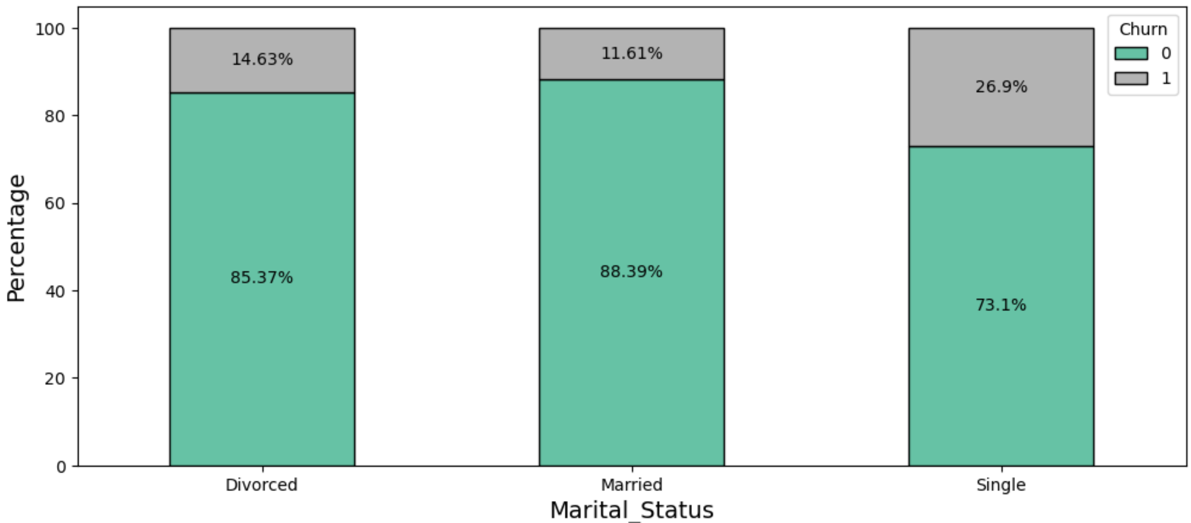


Figure 13: Churn rate by Marital Status

Single persons are more likely to churn than married and divorced persons. Married persons have the least tendency to churn.

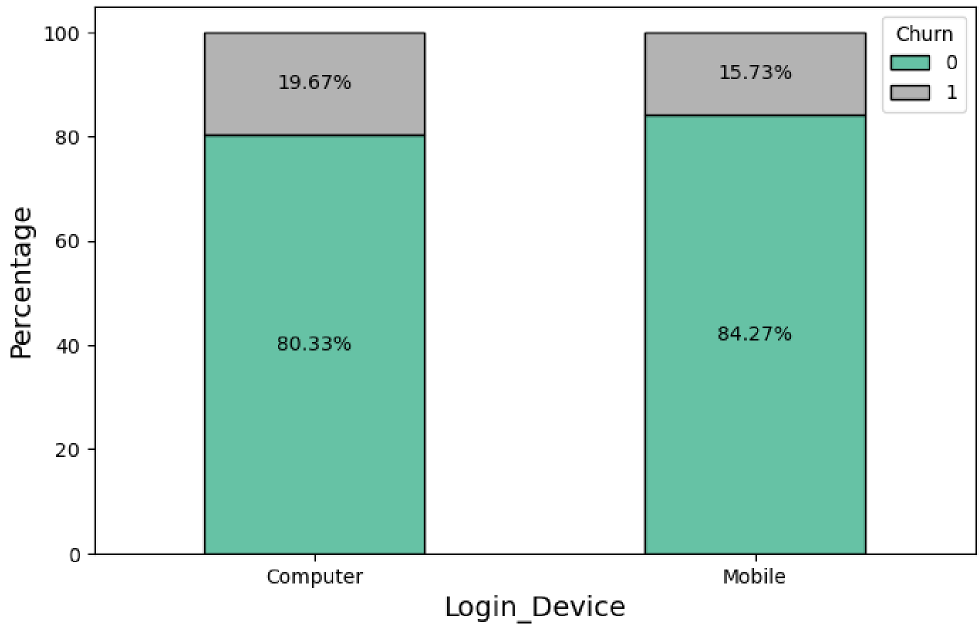


Figure 14: Churn rate by Login Device

Computer users have a higher probability to churn than mobile users.

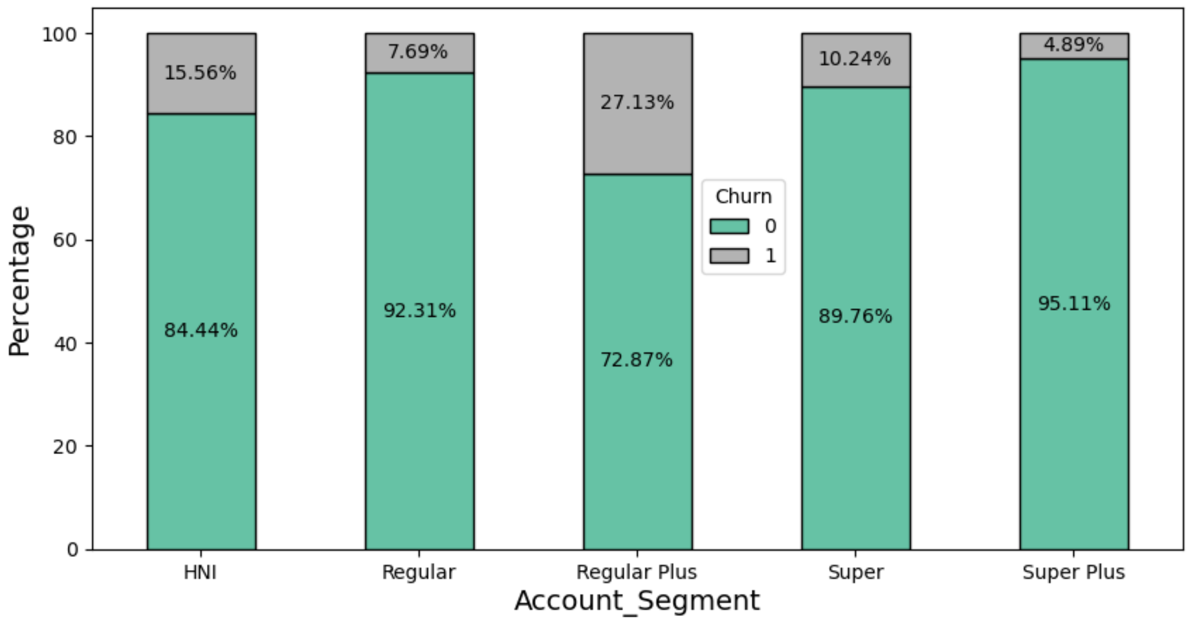


Figure 15: Churn rate by Account Segment

Regular Plus customers are more likely to churn followed by HNI users. Super Plus customers are least likely to churn. In other words, they are more likely to remain loyal to the company.

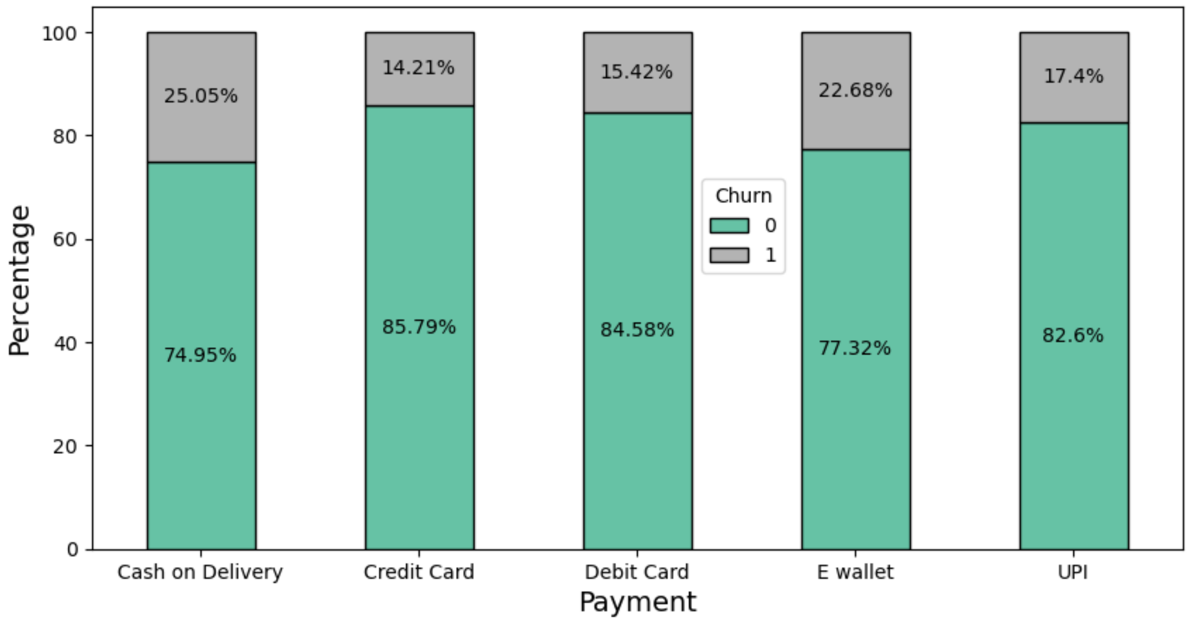


Figure 16: Churn rate by Payment

Customers preferring cash on delivery as a mode of payment has the highest probability to churn followed by those using e-wallet. Credit card users churn the least.

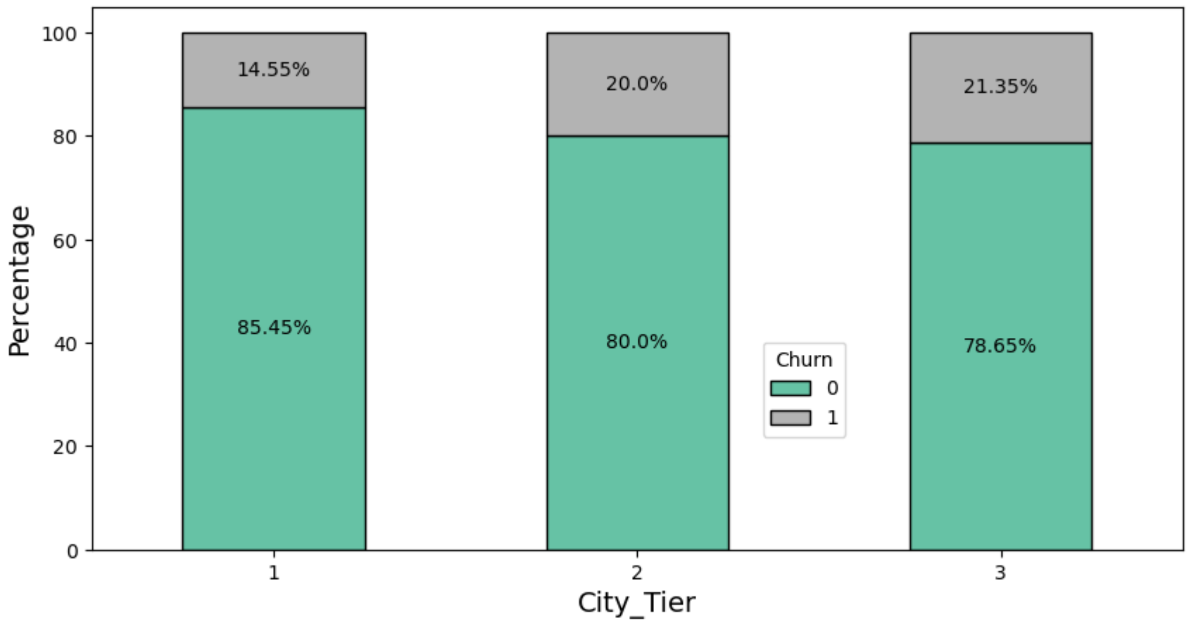


Figure 17: Churn rate by City Tier

Churn rate is the highest in Tier 3 cities. Tier 2 cities also have a high churn rate.

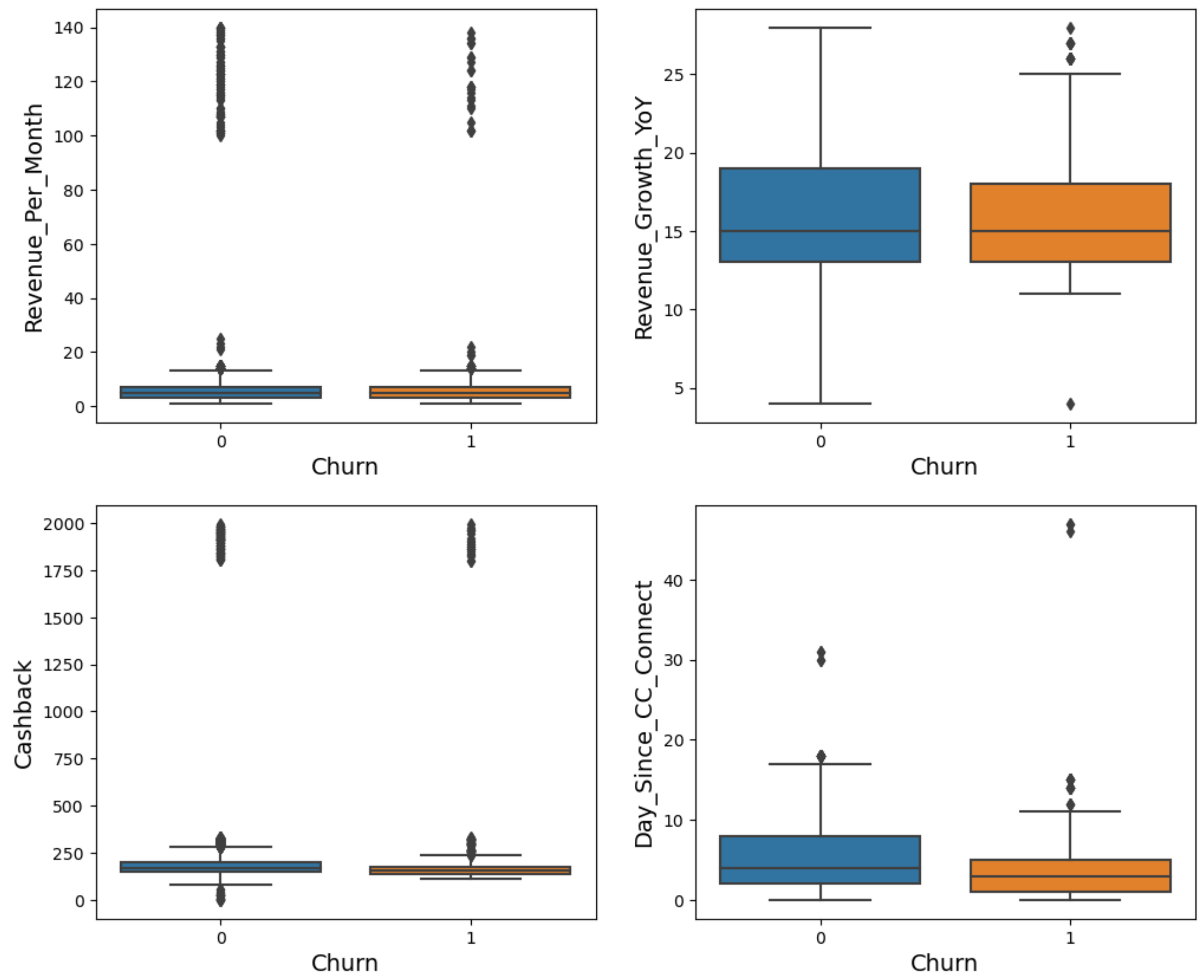


Figure 18: Boxplots for Churn vs other variables

Revenue Per Month has no impact on the churn rate. There are a few outliers.

Year-on-year Revenue Growth also has no impact on the churn rate. However, the spread of those who have not churned (label 0) is more than those who have.

Cashback, too, does not have a major impact on the churn rate.

Day_Since_CC_Connect has an impact on the churn rate. Median Day_Since_CC_Connect for label 0 is more than that for label 1.

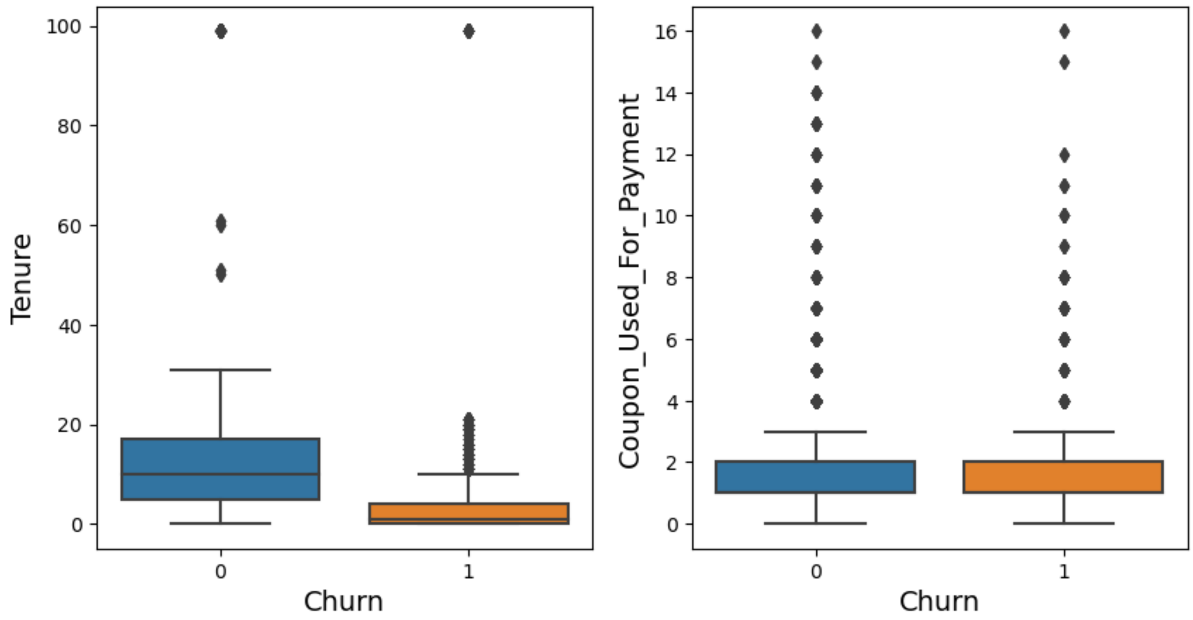


Figure 19: Boxplots for Tenure & Coupon Used vs Churn

Tenure has an impact on the churn rate. The median tenure for those who have churned (label 1) is less than those who have not churned.
Coupon_Used_For_Payment has no impact on the churn rate.

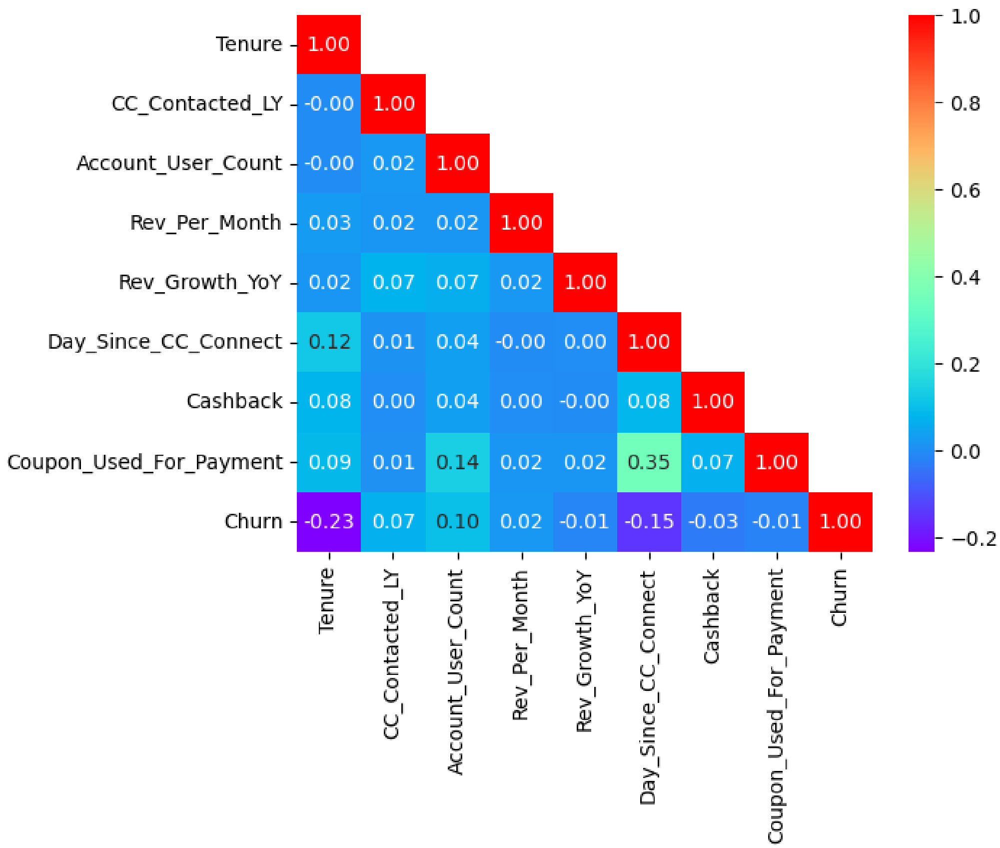


Figure 20: Heatmap for continuous features

The thumb rule is that any value over 0.7 or under -0.7 has a strong linear dependence. Anything between -0.4 and 0.4 indicates absence of linear dependence. The sign indicates the type of correlation – + for positive correlation and - for negative correlation.

It can be seen from the heatmap (Figure 20) that correlation among variables is weak. Some of the variables are not at all correlated with each other.

Multivariate analysis

Pair plots can show the relation between two continuous variables. With hue as Churn, we can show the relation among three features.

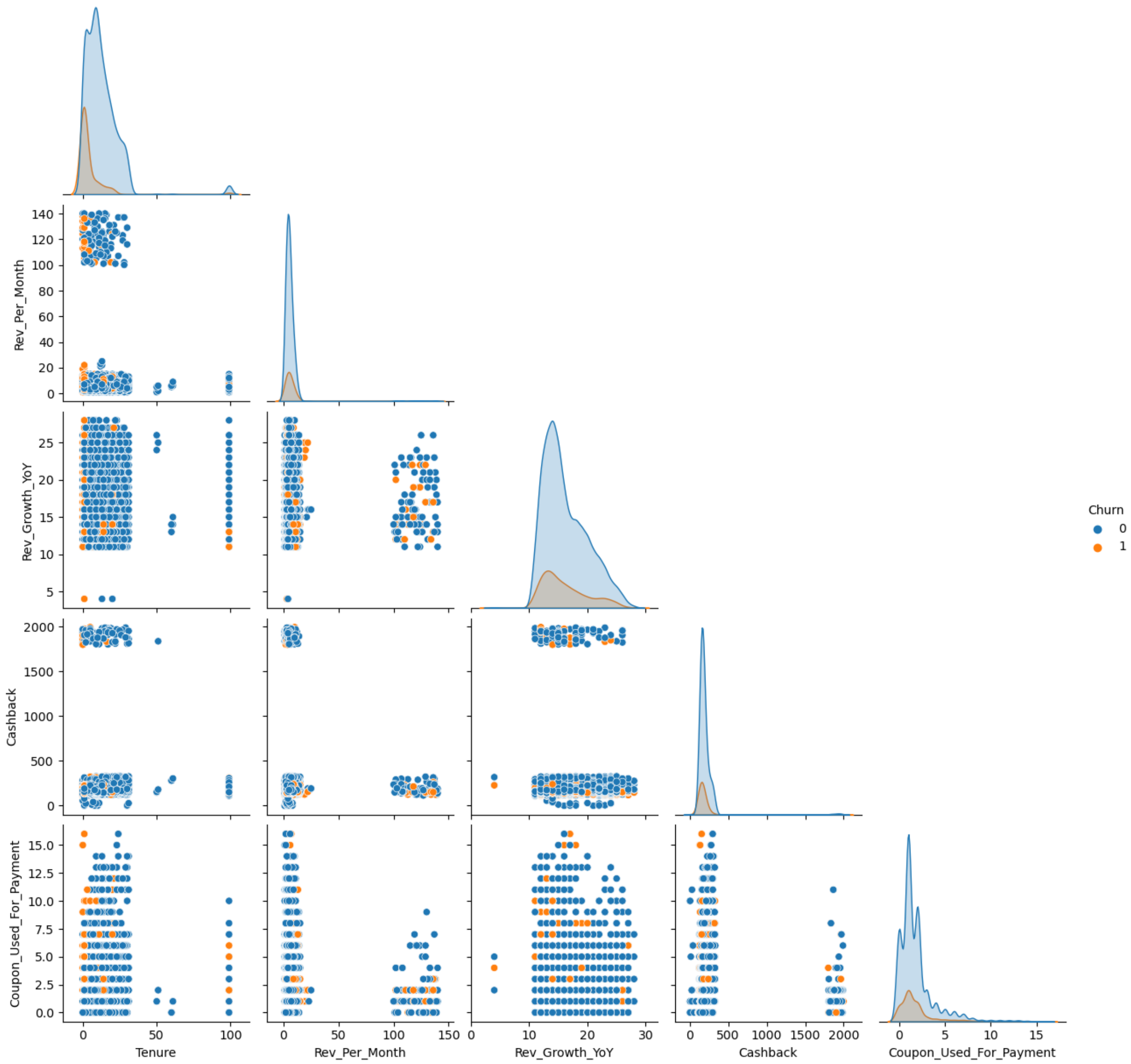


Figure 21: Pair plots for continuous features

Too much scatter on the pair plot indicates there is no correlation between features. It can be observed from the pair plots that no two variables are correlated.

Business implications from EDA

The following business implications can be drawn from the exploratory data analysis.

- The e-commerce company's customer base is highly concentrated in Tier 1 cities. Its presence in Tier 2 cities is the least.
- The majority of the users are men.
- Debit card is the most preferred mode of payment among customers. Since debit card is mostly used by salaried persons, it can be construed that the company's customer base is dominated by salaried persons.
- Online payment methods such as e-wallets and UPI are not preferred by many.
- Regular Plus and Super are the most popular account segments. Regular is the least preferred account segment.
- Most of the customers have given a rating of 3 to the service provided by the company. This can be seen as an 'average' rating, which implies that there is a scope for improvement on the service front.
- Users' experience with customer care agents seems to be 'average' as most of them have a given a rating of 3. In other words, the customer care team needs to improve and answer complaints promptly.
- The company's customer base has more married persons than single and divorced persons.
- The majority of the customers have not complained in the past one year. However, it cannot be construed that they are satisfied with the company's service. Sometimes, customers churn silently.
- Churn rate is associated with
    - Male users
    - Single persons
    - Computer users
    - Regular Plus users
    - Tier 2 and 3 cities
    - Customers preferring cash on delivery
- Tenure impacts the churn rate. New customers leave the company early, which is a cause for concern. The company is failing to capitalise on new customers.

Data cleaning & pre-processing

Null values

AccountID	0
Churn	0
Tenure	218
City_Tier	112
CC_Contacted_LY	102
Payment	109
Gender	108
Service_Score	98
Account_User_Count	444
Account_Segment	97
CC_Agent_Score	116
Marital_Status	212
Rev_Per_Month	791
Complain_LY	357
Rev_Growth_YoY	3
Coupon_Used_For_Payment	3
Day_Since_CC_Connect	358
Cashback	473
Login_Device	232

Table 5: Null values

AccountID	0.000000
Churn	0.000000
Tenure	1.936057
City_Tier	0.994671
CC_Contacted_LY	0.905861
Payment	0.968028
Gender	0.959147
Service_Score	0.870337
Account_User_Count	3.943162
Account_Segment	0.861456
CC_Agent_Score	1.030195
Marital_Status	1.882771
Rev_Per_Month	7.024867
Complain_LY	3.170515
Rev_Growth_YoY	0.026643
Coupon_Used_For_Payment	0.026643
Day_Since_CC_Connect	3.179396
Cashback	4.200710
Login_Device	2.060391

Table 6: Percentage of null values

Table 5 shows missing values after having replaced special characters with the null values.

Except for Account ID and Churn, all variables have null values. Rev_Per_Month has the highest number of null values (7 per cent) followed by Cashback (4.2 per cent).

Null values comprise 1.8 per cent of the data points. It seems to be a small number and, hence, one of the approaches is to drop the null values. However, doing so might result in losing some important data points.

Therefore, null values will be imputed with the help of different techniques.

Missing value treatment

The imputation technique for null values or missing values depends on the type of the variable. Generally, null values in numerical variables are imputed either with mean or median, while missing values in categorical variables are imputed with mode. There are other techniques such as K-nearest neighbours (KNN) imputation and Regression imputation as well.

Since all continuous features are skewed, missing values in all continuous features, but one, have been imputed with median. Variance and the distribution of the variables before and after imputation were checked and it was found that imputation has not altered the original variable. Therefore, **median imputation seems to be**

best choice for the following features: Tenure, CC_Contacted_LY, Account_User_Count, Rev_Per_Month, Complain_LY, Cashback, Rev_Growth_YoY and Coupon_Used_For_Payment.

Meanwhile, null values in Day_Since_CC_Connect have been imputed with mean because median altered the original variable to some extent.

As for categorical variables, null values have been imputed with mode, the observation that appears the most often in a variable.

Outlier check

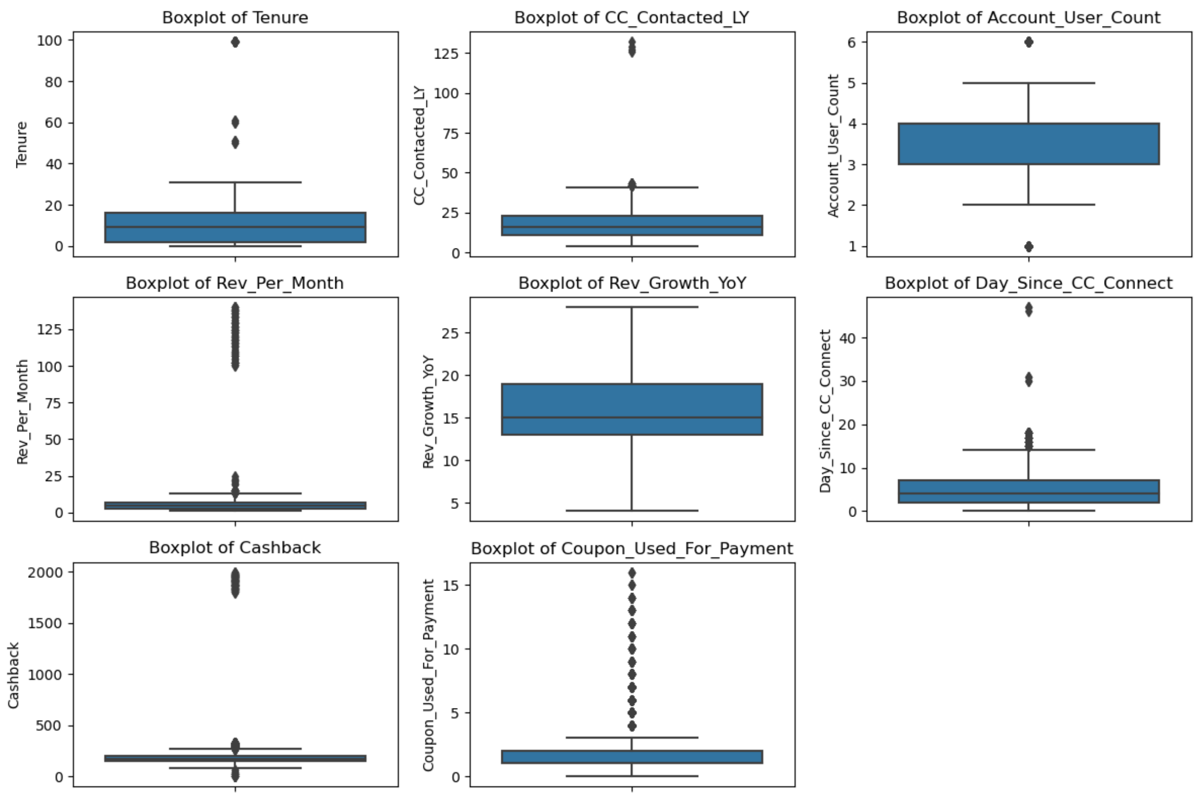


Figure 22: Boxplots for numerical variables

Barring year-on-year revenue growth, all variables have outliers.

For parametric models such as Linear Regression and Logistic Regression, outliers have a negative impact on the performance of the model. Therefore, it is important to treat the outliers.

Different techniques to treat the outliers are as follows:

- Drop the outliers
- Cap the outliers
- Transform the variables

Before dropping and capping the outliers, we should explore the option of transforming the variables. The objective is to include the maximum number of data points in the model-building process.

One of the common transformation techniques is **log transformation,** which will reduce the variance in the variables. We will apply it on the following continuous variables: Tenure, CC_Contacted_LY, Rev_Per_Month, Day_Since_CC_Connect, Cashback and Coupon_Used_For_Payment.

Log transformation

After the log transformation, the number of outliers will comparatively reduce as can be seen from the boxplots.

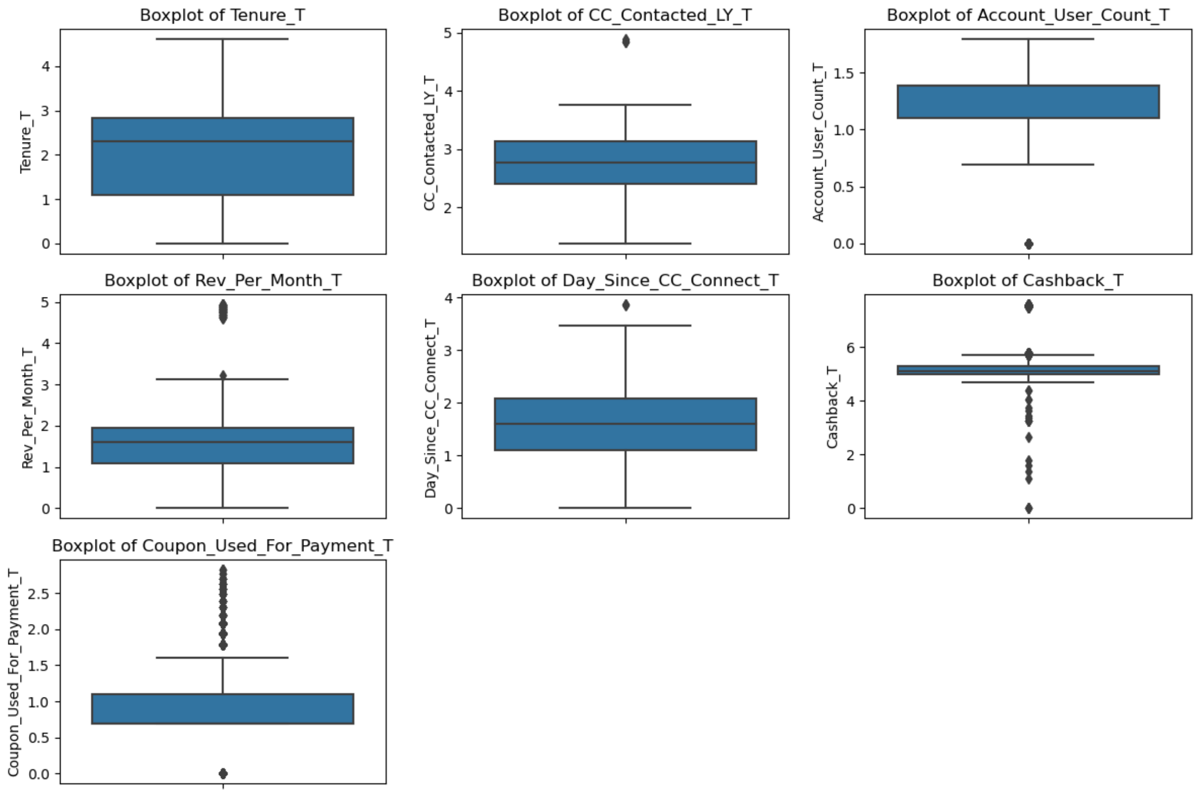


Figure 23: Boxplots after log transformation

The number of outliers has decreased in all variables except for Coupon_Used_For_Payment. There has been a substantial rise in the number of outliers in Coupon_Used_For_Payment. Therefore, log transformation is not the right technique for this variable.

Outlier treatment

Even after the log transformation, the dataset has some outliers. These will be removed so that performance of the models is not affected.

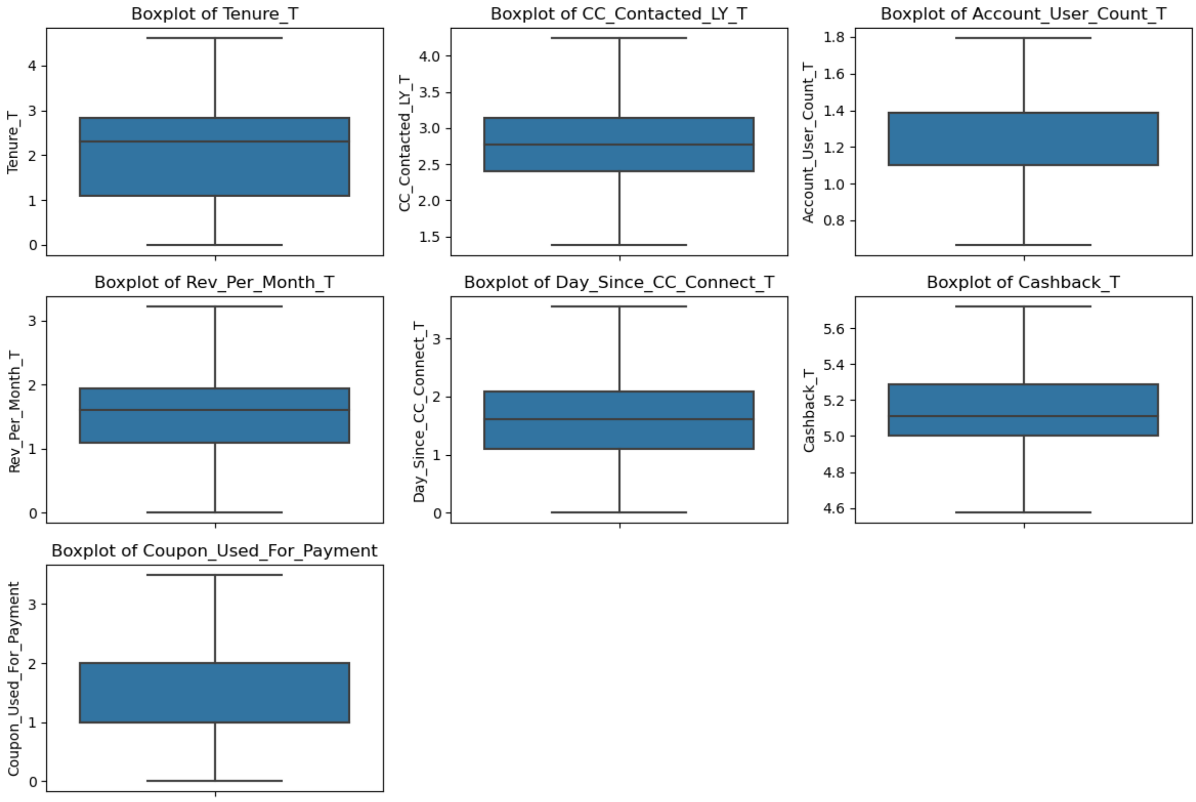


Figure 24: Boxplots post outlier treatment

It can be seen from the boxplots that the variables no longer have outliers.

Encoding categorical variables

Categorical variables cannot be used directly in linear models because the best fit line needs to fit on numerical values. In the customer churn dataset, five variables need to be encoded. The variables have been given the following labels.

Label	Payment	Account Segment	Marital Status	Gender	Login Device
0	Debit Card	Regular	Single	Male	Mobile
1	UPI	Regular Plus	Married	Female	Computer
2	Credit Card	Super	Divorced		
3	Cash on Delivery	Super Plus			
4	E wallet	HNI			

Table 7: Encoded categorical variables

Clustering

Dimensionality reduction

It is important to reduce dimensions before segmenting the customers. Otherwise, the “curse of dimensionality” will kick in. In case of too many variables, machine learning (ML) models face difficulty in working with the data. When more dimensions (features) are added, the minimum data requirements also increase rapidly. Therefore, it is imperative to reduce features. **Variance Inflation Factor (VIF), which helps us to detect multicollinearity in the regression model, is one of the techniques which helps in dimensionality reduction.**

The VIF measures the inflation in the variances of the regression parameter estimates because of collinearity among independent variables. Its value lies between 1 and infinity.

<u>Rule of thumb</u>

- If VIF = 1, there is no correlation between one of the predictors, say A, and others predictors.
- If VIF > 5, there exists moderate multicollinearity.
- If VIF > 10, there exists high multicollinearity.

For our problem at hand, we keep the threshold of the VIF = 5. Any variable having a VIF over 5 will be dropped.

Variables	VIF
Cashback_T	99.436573
CC_Contacted_LY_T	33.595026
Account_User_Count_T	28.103411
Service_Score	20.711083
Rev_Growth_YoY	19.564668
Rev_Per_Month_T	7.916130
Day_Since_CC_Connect_T	6.747939
CC_Agent_Score	6.043496
Tenure_T	5.456592
City_Tier	4.877168
Account_Segment	4.651575
Coupon_Used_For_Payment	3.454239
Marital_Status	2.667577
Payment	2.259462
Gender	1.666775
Complain_LY	1.396586
Login_Device	1.396078

It can be seen from the table that some variables have VIF over 5. These features must be dropped one by one before building a regression model.

Table 8: VIF values

After having dropped the variables, we get the following VIF values.

Variables	VIF
CC_Agent_Score	4.289174
Account_Segment	4.232290
City_Tier	4.230582
Tenure_T	4.215912
Coupon_Used_For_Payment	2.788097
Marital_Status	2.519953
Payment	2.237076
Gender	1.640579
Login_Device	1.355707
Complain_LY	1.345843

Table 9: VIF values after dropping variables

In all, seven variables were dropped. Now, each variable has a VIF less than 5.

As a result of dropping seven more variables, **we are left with 11 columns, including target variable Churn.**

K-Means clustering

To understand the customers' buying pattern, it is important to cluster the clients so that targeted offers and advertisement campaigns can be designed for them.

With the help of K-Means clustering, we can form different segments of clusters.

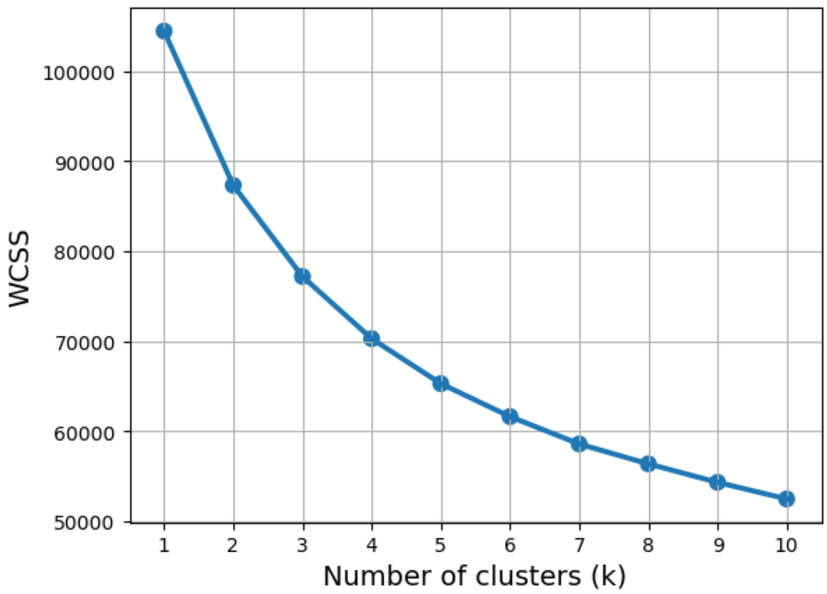


Figure 25: Elbow plot

The elbow plot shows the drop within cluster sum of squares (WCSS).

From k = 1 to k =3, there is a significant drop in the WCSS.

From k = 3 to k = 5, there is a gradual, but significant drop in the WCSS.

However, **there is no clear break in the elbow plot.** In other words, the elbow plot does not help us to determine the number of clusters.

Silhouette score is another method to determine the number of clusters.

Average silhouette score for 2 clusters is 0.15659.

Average silhouette score for 3 clusters is 0.14625.

Average silhouette score for 4 clusters is 0.15067.

Average silhouette score for 5 clusters is 0.14017.

Average silhouette score for 6 clusters is 0.13631.

Average silhouette score for 7 clusters is 0.1374.

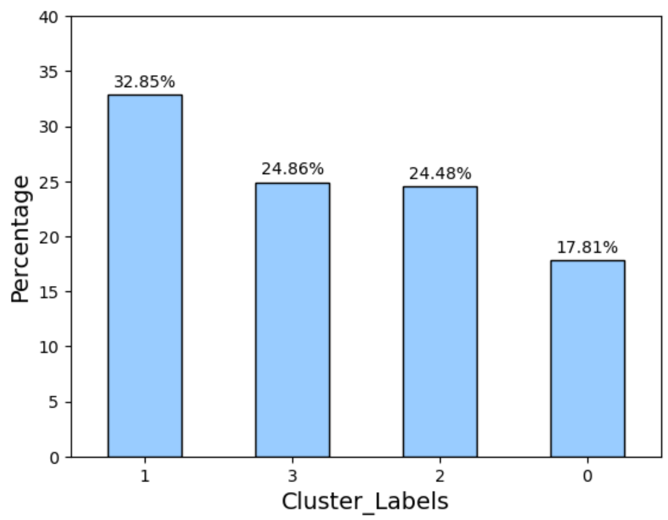
Average silhouette score for 8 clusters is 0.13915.

Average silhouette score for 9 clusters is 0.13358.

Average silhouette score for 10 clusters is 0.13669.

Silhouette score is the maximum for k = 2. However, only two clusters would not be the right choice. Silhouette score for k = 4 is slightly less than the one for k = 2. Therefore, **we take the optimum number of clusters to be four.**

Exploratory Data Analysis on clusters



Cluster 1 has the highest number of records followed by cluster 3.

Cluster 0 has the least records.

Figure 26: Count plot for clusters

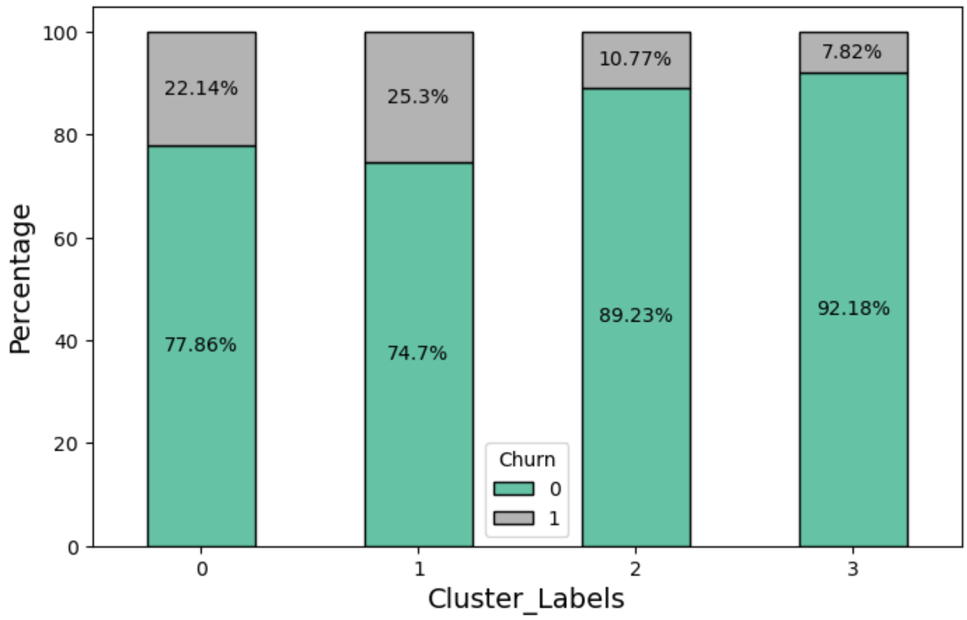


Figure 27: Churn rate by clusters

Churn rate is the highest in Cluster 1 followed by Cluster 0. Cluster 3 has the least churn rate.

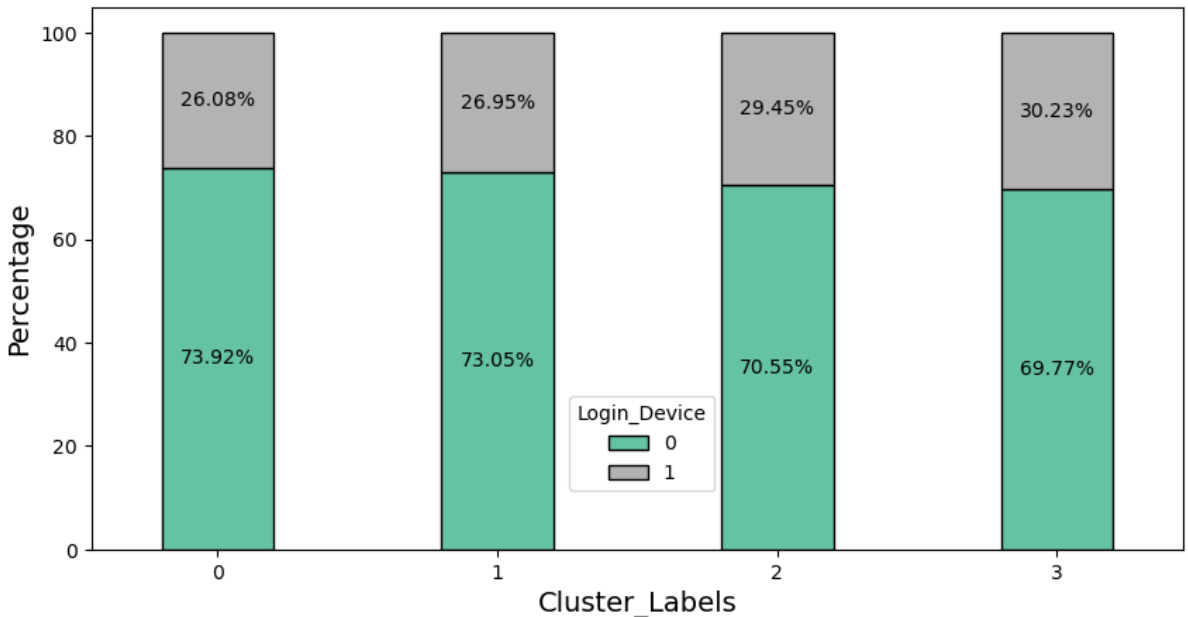


Figure 28: Clusters by Login Device

Cluster 3 has the highest number of computer users (label 1), whereas Cluster 0 has the highest number of mobile users (label 0).

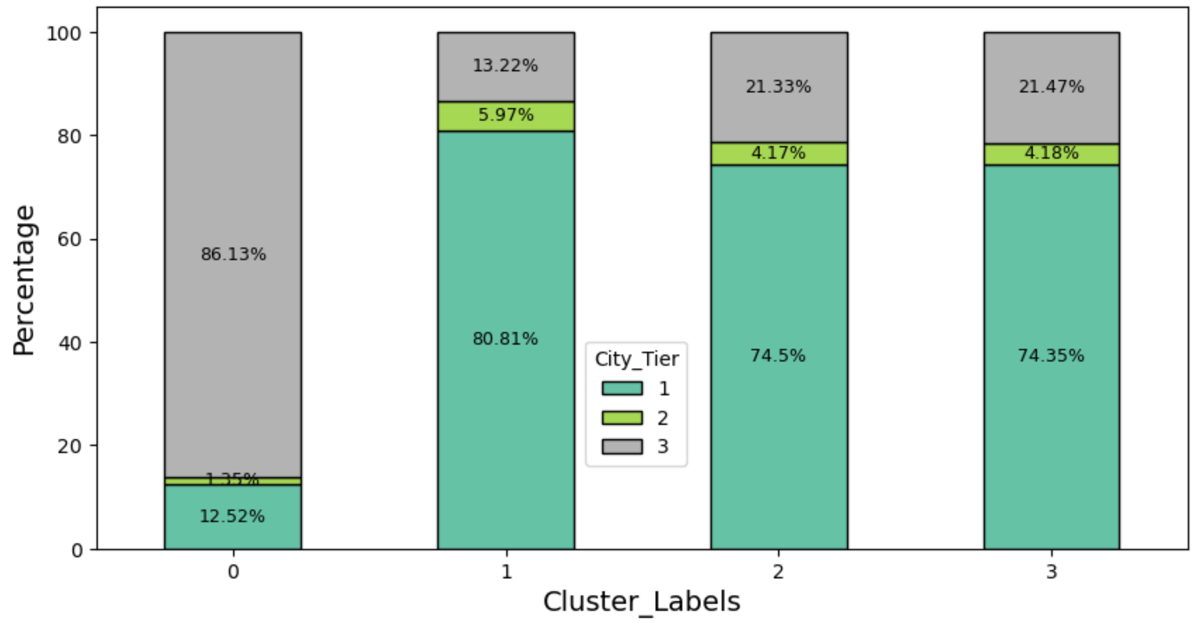


Figure 29: Clusters by City Tier

Cluster 0 is dominated by customers from Tier 3 cities, whereas the other three clusters are dominated by users from Tier 1 cities. Cluster 1 has the highest number of City Tier 1 customers. Tier 2 cities have a minimal presence in all four clusters.

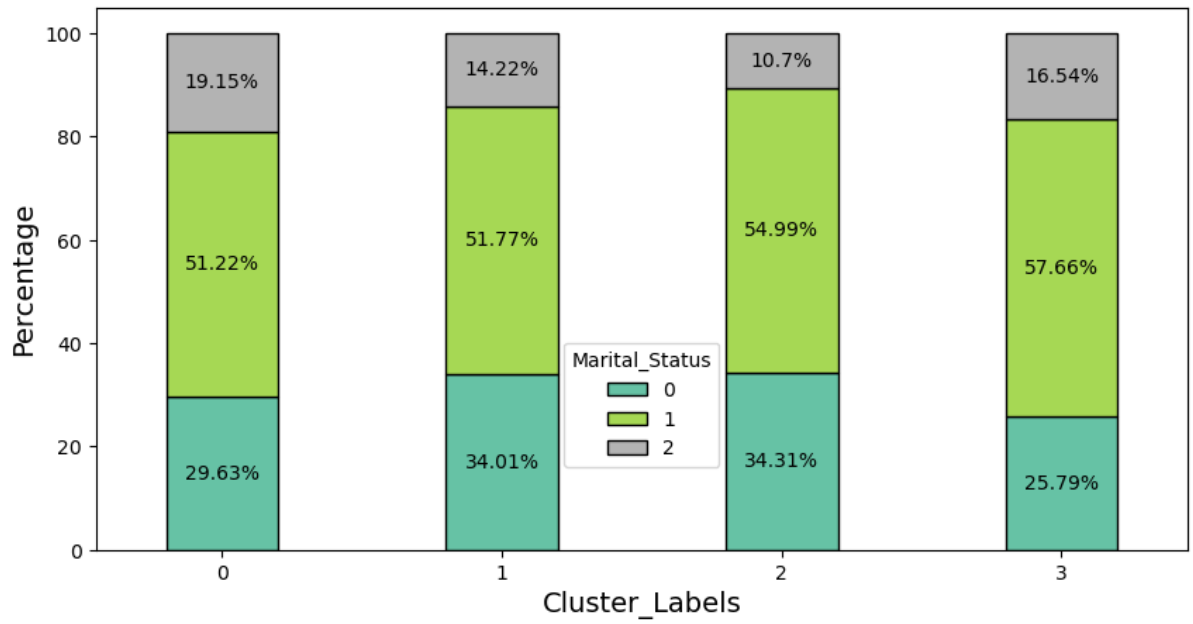


Figure 30: Clusters by Marital Status

The number of married persons (label 1) is the highest in all four clusters. The number of single persons (label 0) is the highest in Cluster 2. The number of married persons (label 1) is the highest in Cluster 3. Cluster 0 has the highest number of divorced persons.

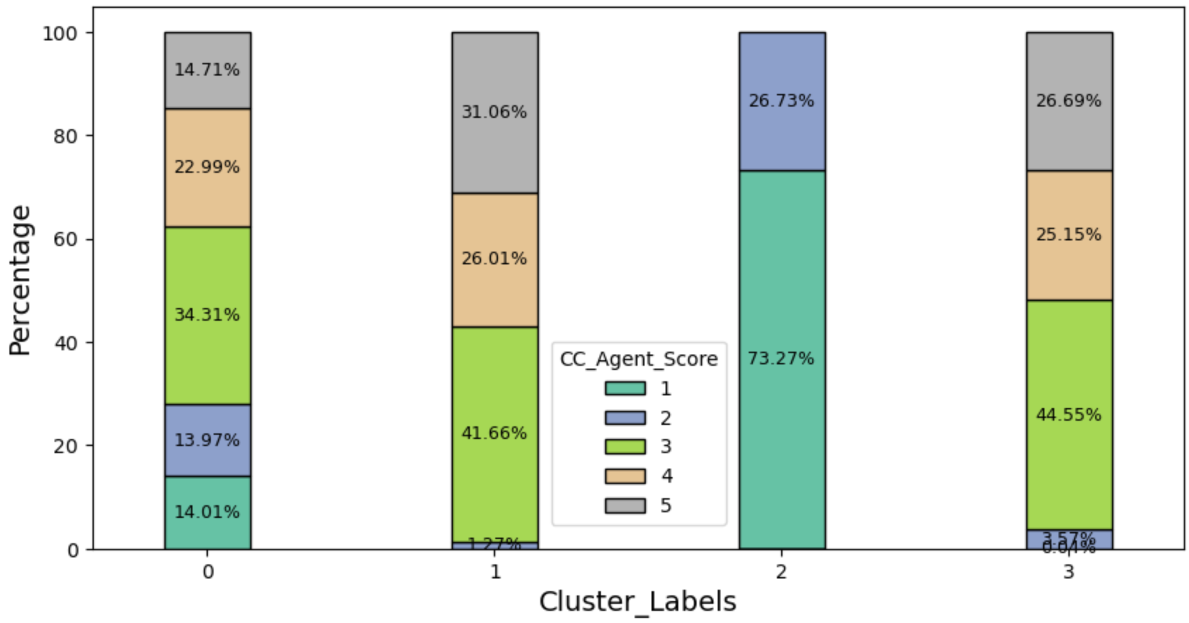


Figure 31: Clusters by CC Agent Score

Customers in Cluster 2 give only low ratings to customer care agents, whereas the majority of subscribers in Cluster 3 don't give low scores. On a scale of 5, they give a rating of 3 and above. The number of users who give the rating of 5 is the highest in Cluster 1 followed by Cluster 3.

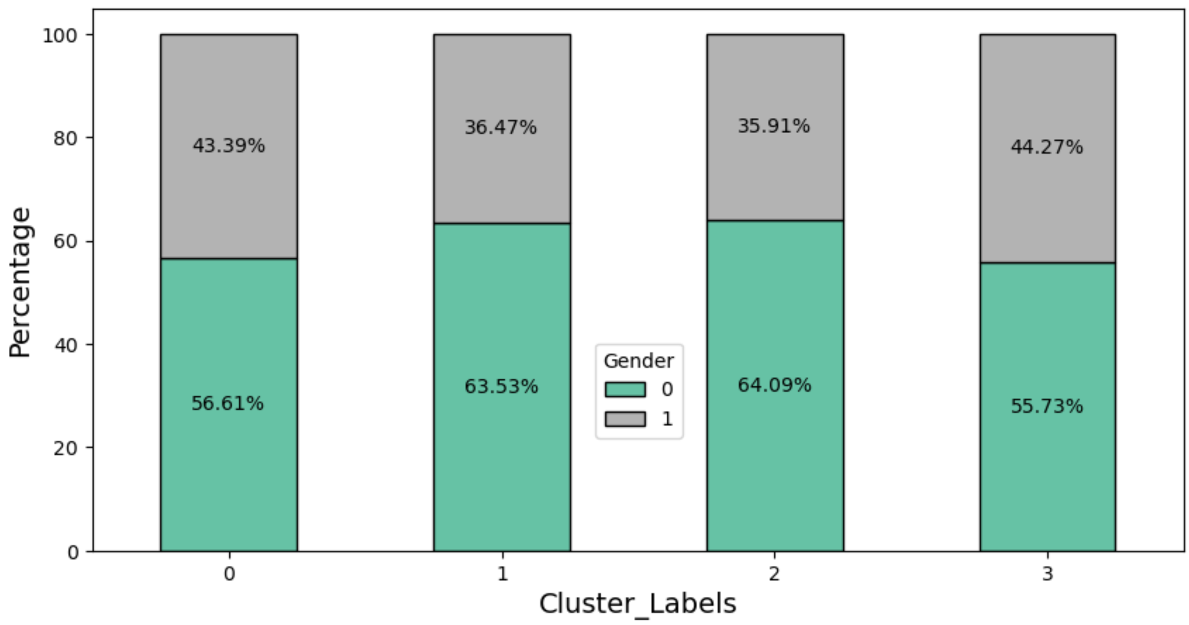


Figure 32: Clusters by Gender

Among the four clusters, Cluster 3 has the highest number of females. The number of males is the highest in Cluster 2.

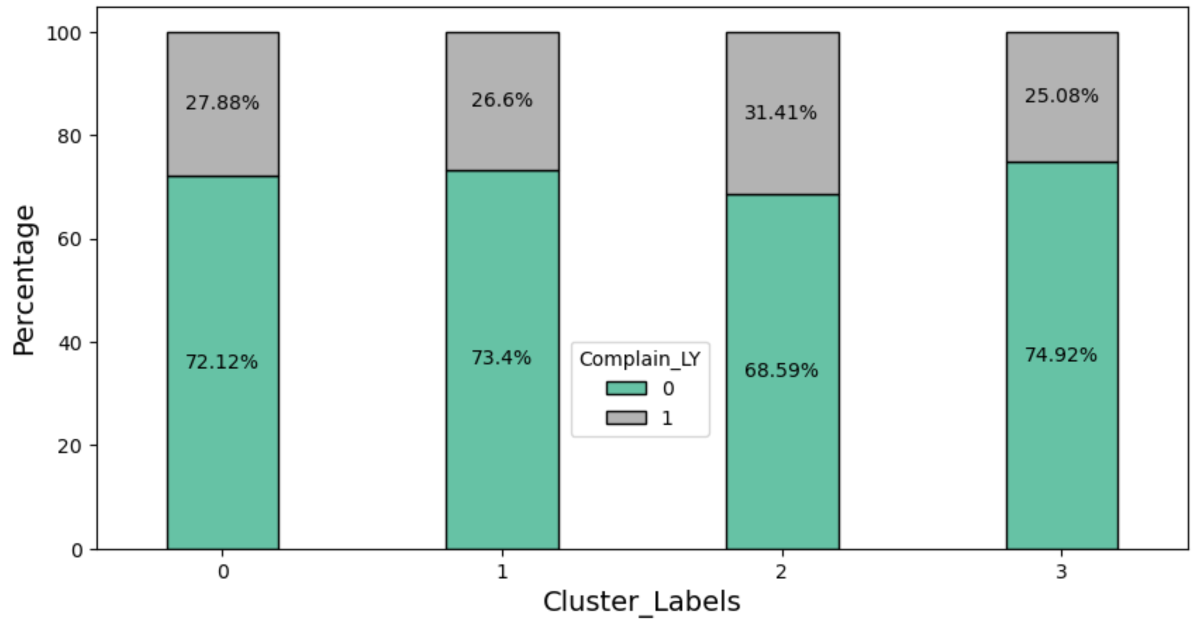


Figure 33: Clusters by Complain Last Year

The highest number of customers who have not complained (label 0) in the past one year is in Cluster 3, while Cluster 2 has the highest number of users who have complained (label 1).

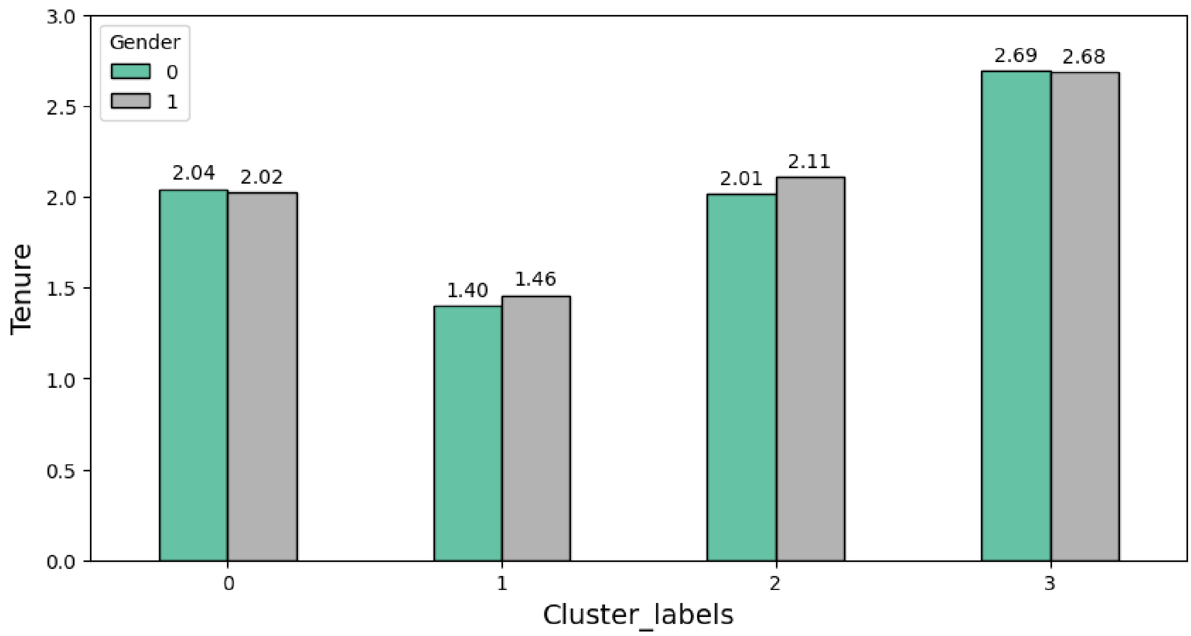


Figure 34: Clusters by Gender & Tenure

Customers in Cluster 3 have the longest tenure, while users in Cluster 1 have the shortest tenure. Females (label 1) in Clusters 1 and 2 have a longer tenure than males (label 0).

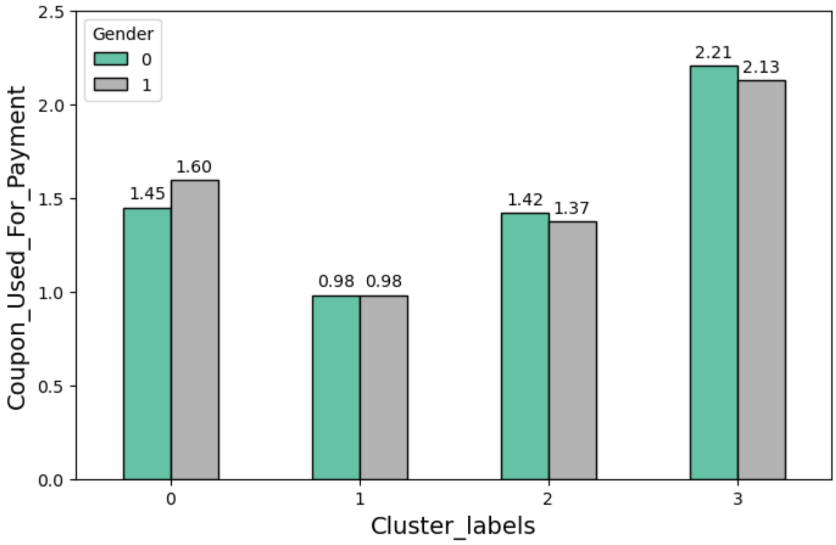


Figure 35: Clusters by Gender & Coupon Used

Subscribers in Cluster 3 use coupons the most, while customers in Cluster 1 use coupons the least. Among them, males (label 0) use coupons more number of times. Females (label 1) in Cluster 0 use coupons more number of times than males.

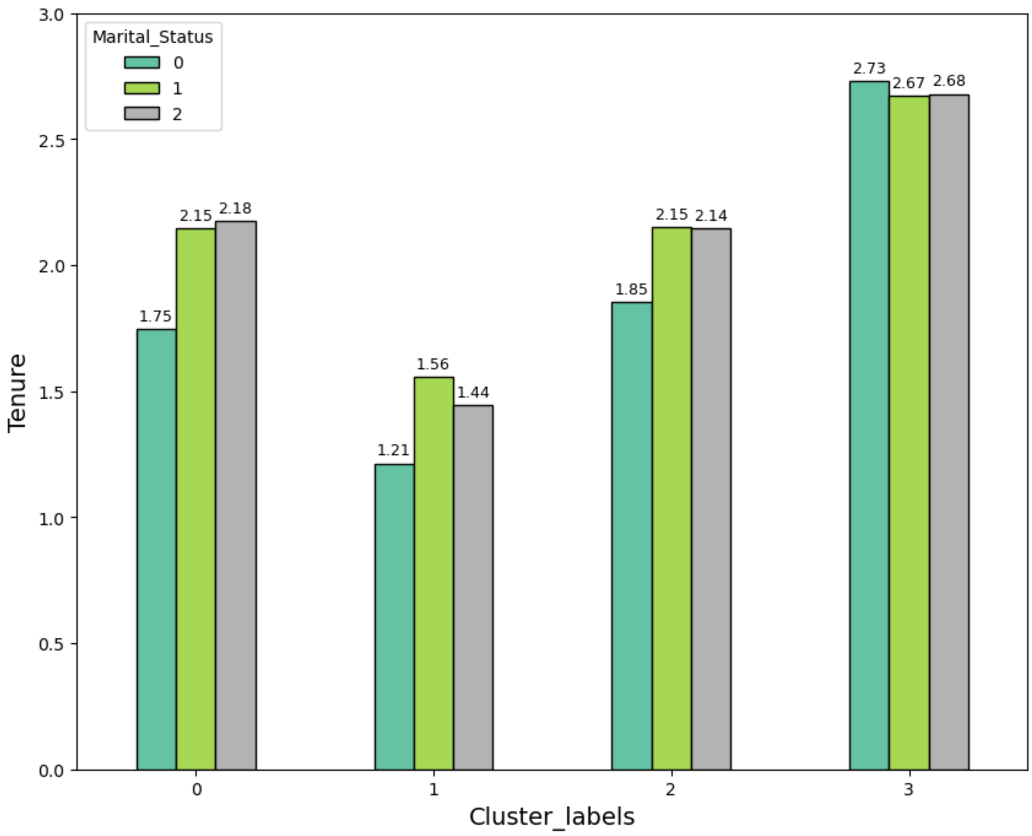


Figure 36: Clusters by Marital Status & Tenure

Single persons (label 0) have the longest tenure in Cluster 3.

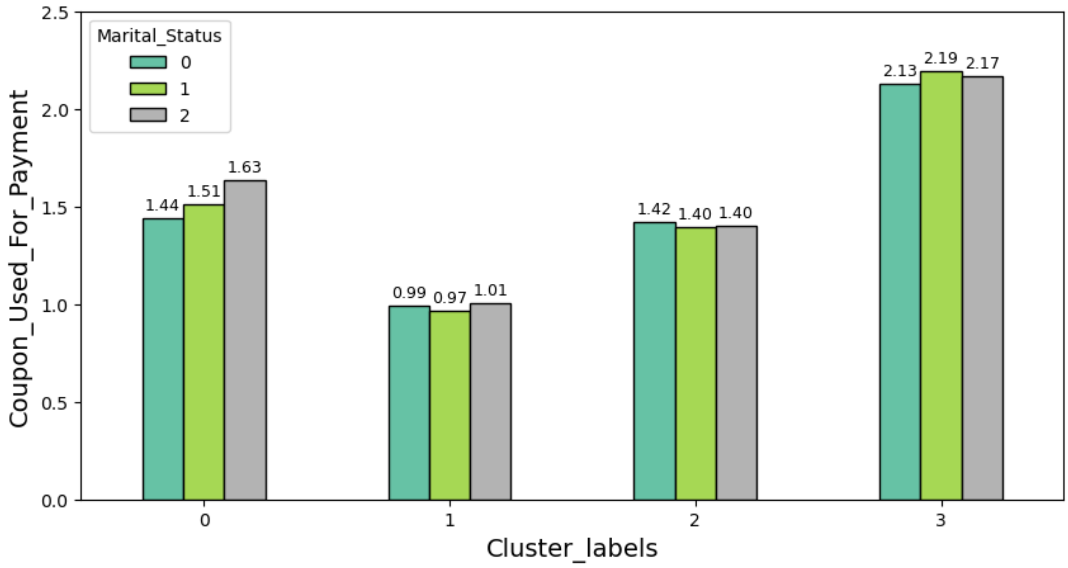


Figure 37: Clusters by Marital Status & Coupon Used

Customers in Cluster 3 use coupons the most. Among them, married persons (label 1) use coupons the most.

Business insights

Clusters should help the company in targeting the customers with high probability of churn. Instead of having a generic campaign for all customers that will hit profits, segment-based campaigns should be run. If the customer is not going to churn, no coupons or discounts should be offered to that client.

The following cluster-wise insights can be drawn from the EDA.

Cluster 0

- Second-highest churn rate.
- Least number of customers.
- Highest number of mobile users.
- Highest number of customers from Tier 3 cities.
- Highest number of divorced persons.

Cluster 1

- Highest churn rate.
- Highest number of customers.
- Highest number of customers from Tier 1 cities.
- Highly value the services of customer care agents.
- Use coupons the least number of times.
- Customers with the shortest tenure.

Cluster 2

- Highest number of single persons.
- Only low ratings to customer care agents.
- Highest number of males.
- Highest number of users who have complained.

Cluster 3

- Churn rate is the least.
- Highest number of females.
- Highest number of computer users.
- Highest number of married persons.
- Only high ratings to customer care agents.
- Customers with the longest tenure.
- Use coupons the most number of times.
- Satisfied with the company's service as a majority of users have not complained.

Model building

Model-building approach

Model-building is an important step to predict the churn rate, as it will help the e-commerce company to target the customers in a better manner. Mathematical models will also help the company to devise strategies and come up with specific insights on how to reduce the churn rate.

The following four approaches will be undertaken to build several models.

1. **Model-building with imbalanced data:** The dataset that we have been provided with is imbalanced. In other words, there is a class imbalance when it comes to the target column, Churn. Therefore, performance metrics such as precision, recall and F1 score will be looked at to evaluate the performance of the models.
2. **Model-building with balanced data:** The Sampling Minority Oversampling Technique (SMOTE) will be employed to deal with the class imbalance. Accuracy score will be of our interest in this case.
3. **Model-building with hypertuning parameters:** Hypertuning parameters will be used to improve the performance of the models. Later, the performance of the models will be compared to find out the best one.
4. **Ensemble modelling:** Different ensemble techniques such as bagging and boosting will be used to build machine learning models.

Train-test split

Before building any model, it is important to split the dataset into training and test sets.

- **Training data**: A training dataset is used to fit the models and estimate the parameters.
- **Test data**: A test data is the unseen data. It is used to assess the performance of the model.

We split the dataset into 70:30 ratio – 70 per cent of the data as the training set and 30 per cent as the test set. The **training dataset has 7,882 rows and 10 columns**. The **test dataset has 3,378 rows and 10 columns**.

Both training and test sets have a **class imbalance**. The target column, Churn, has two classes – Class 0 and Class 1. **The class of interest is label 1** (customers who have churned). The distribution of the two classes is as follows:

Class 0	**:**	**83%**
Class 1	**:**	**17%**

In all, seven parametric and non-parametric models have been built for the imbalanced data. The seven models are:

1. Logistic Regression
2. Linear Discriminant Analysis (LDA)
3. Naïve Bayes
4. Support Vector Machine (SVM)
5. K-Nearest Neighbour (KNN) algorithm
6. Random Forest
7. Decision Tree Classifier

The same number of models have been built with balanced data. **After applying SMOTE only on the training set, we have 13,110 rows and 10 columns.**

Three non-parametric models have been tuned to see whether or not the performance of the models improves. The three models are:

1. K-Nearest Neighbours (KNN) algorithm
2. Random Forest
3. Decision Tree Classifier

As for the ensemble modelling, three models each for Ada Boost, Gradient Boost and Bagging Classifier have been built. In all, **nine ensemble ML models have been built** by employing the following approach.

- Model with imbalanced data
- Model with balanced data
- Model with hypertuning parameters

In all, 26 ML models have been built for the customer churn prediction.

Important features

Logistic Regression (Logit) models

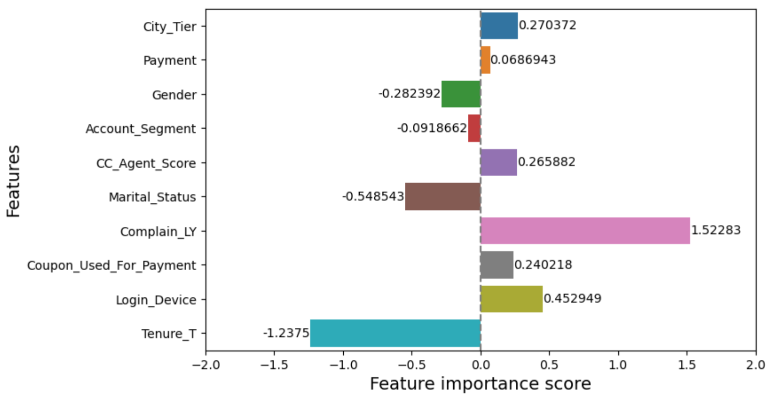


Figure 38: Important features (Logit)

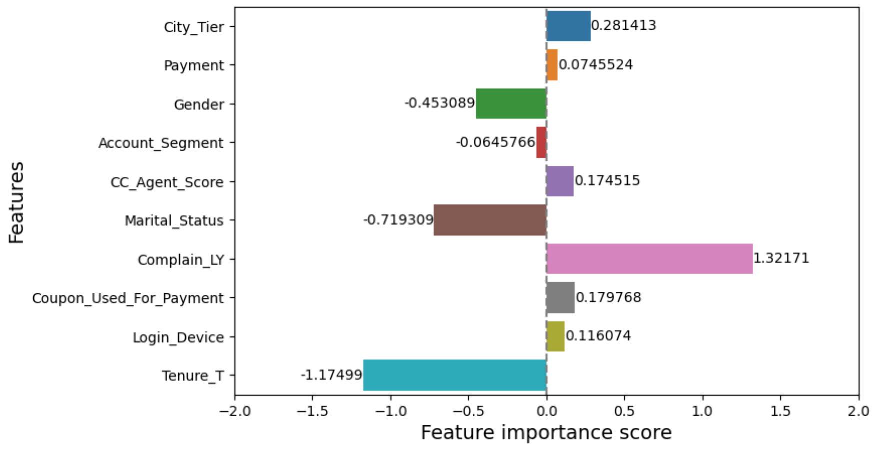


Figure 39: Important features (Logit – SMOTE)

Figures 38 and 39 show that 'Complain_LY' is the most important feature followed by 'Tenure' and 'Marital_Status'. Payment is the least important feature.

Linear Discriminant Analysis models

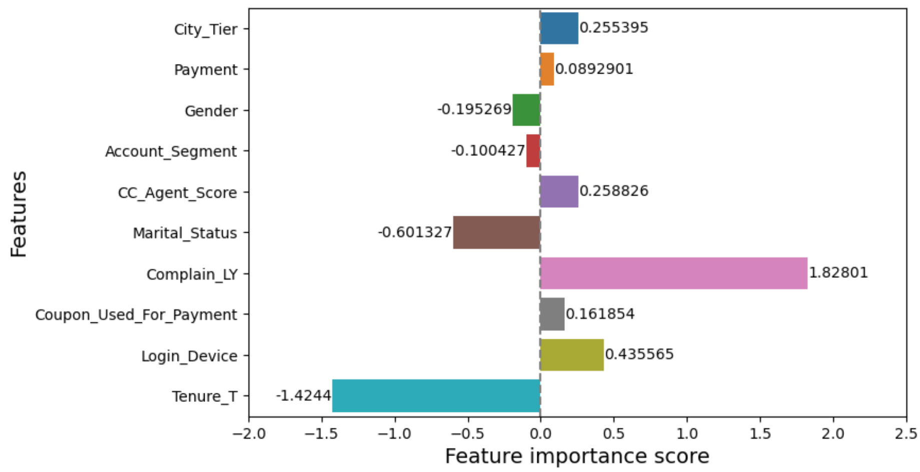


Figure 40: Important features (LDA)

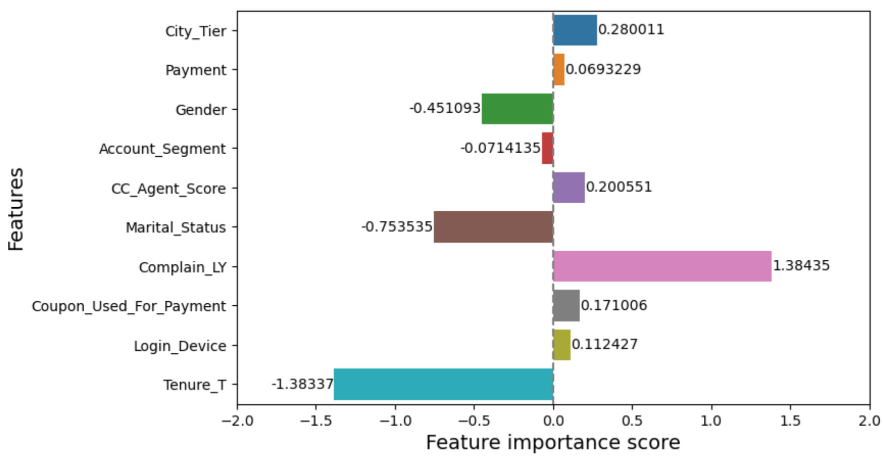


Figure 41: Important features (LDA – SMOTE)

Figures 40 and 41 show that 'Complain_LY' and 'Tenure' are the most important features. 'Account_Segment' and 'Payment' are the least important features.

Random Forest models

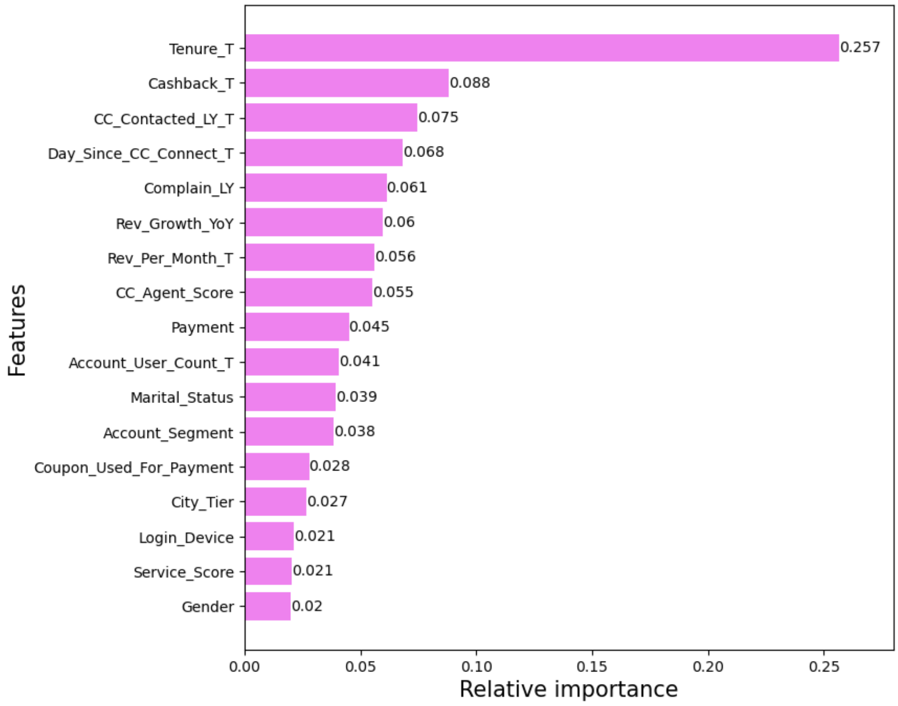


Figure 42: Important features (Random Forest)

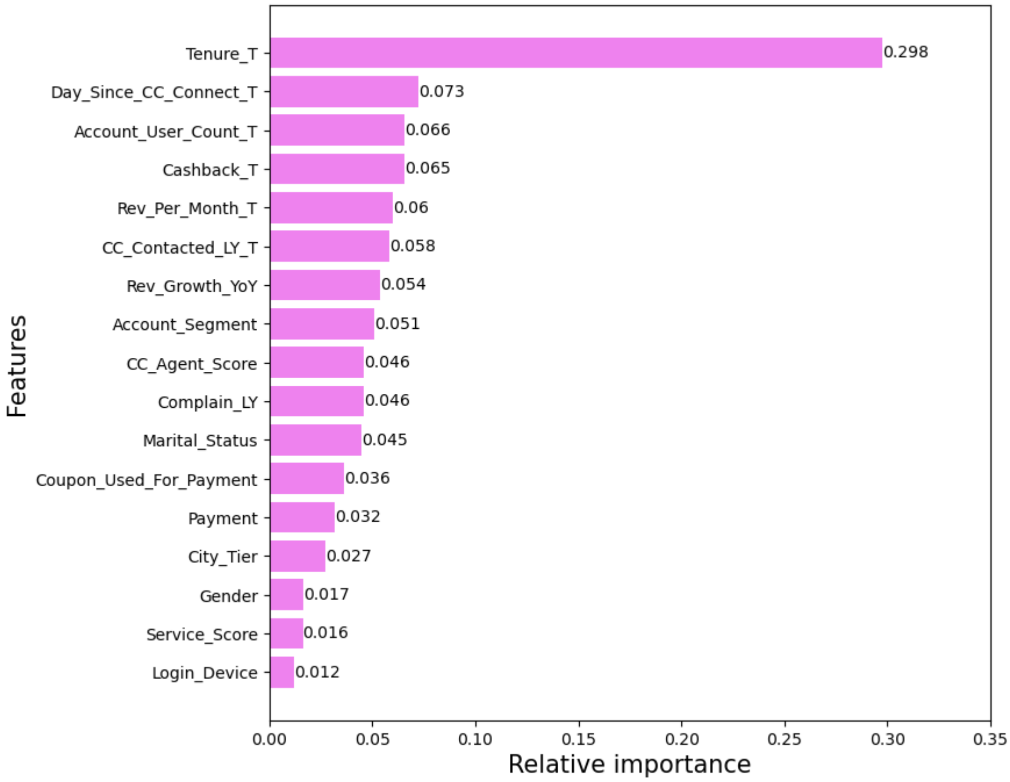


Figure 43: Important features (Random Forest – SMOTE)

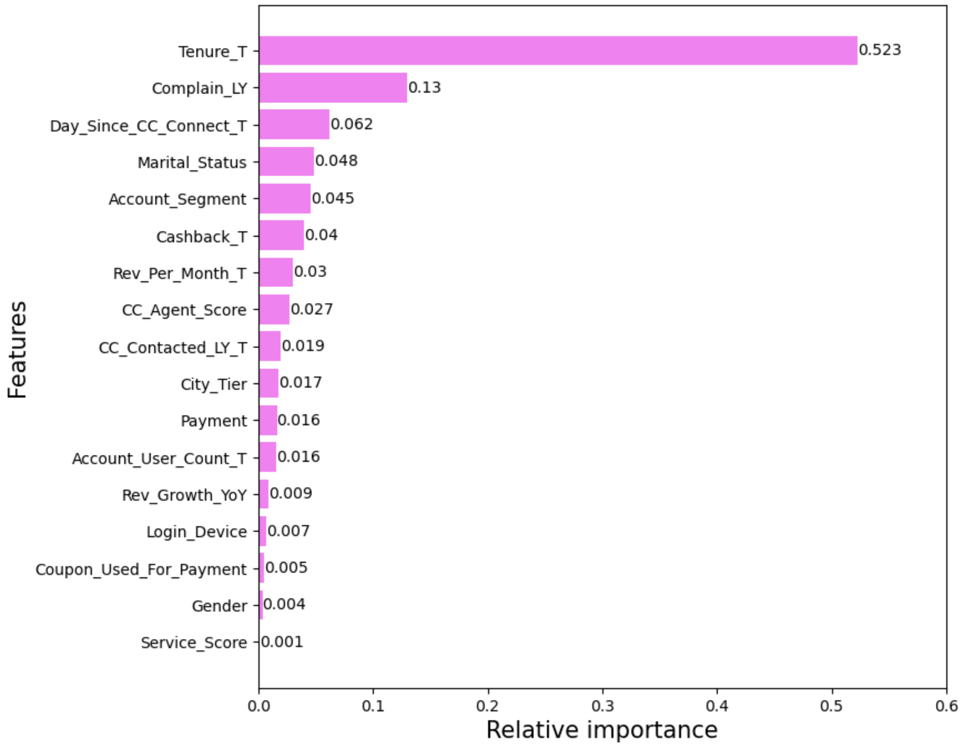


Figure 44: Important features (Tuned Random Forest)

Figures 42, 43 and 44 show that 'Tenure' is the most important feature in the Random Forest models. Other important features are 'Cashback' and 'Day_Since_CC_Connect'.

Decision Tree Classifier

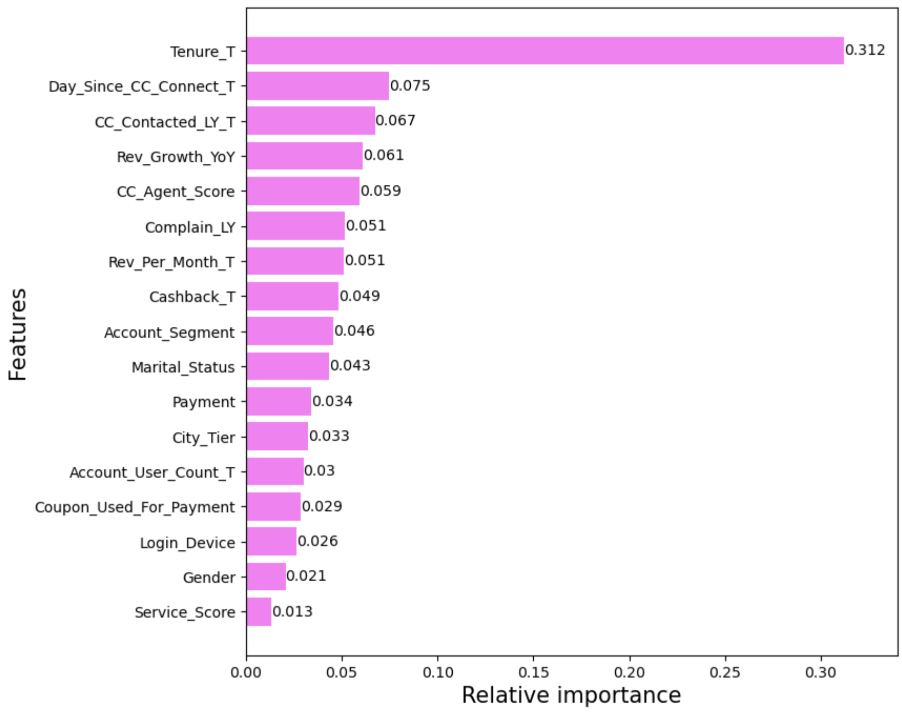


Figure 45: Important features (Decision Tree)

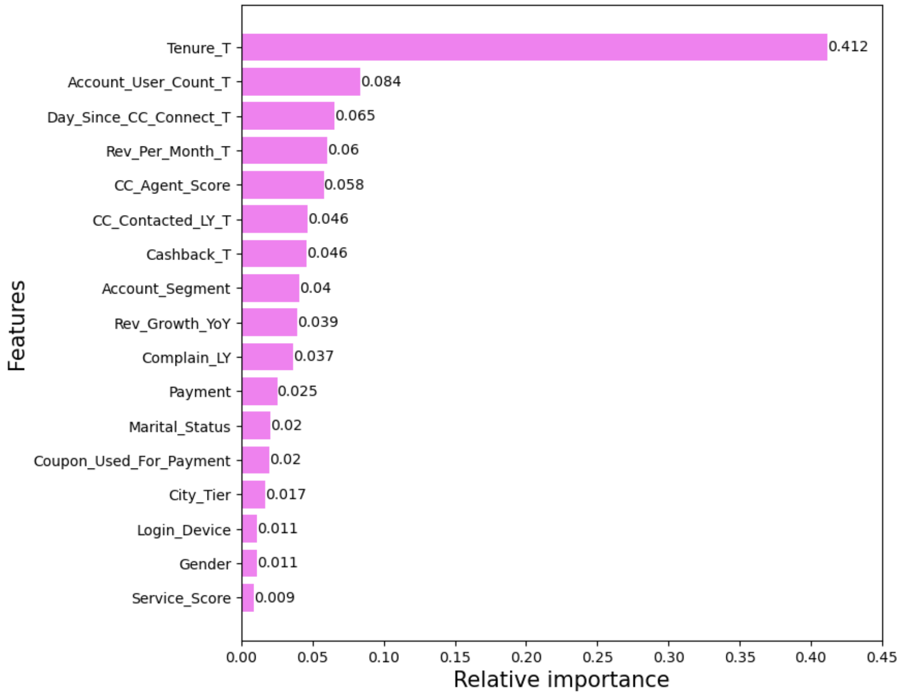


Figure 46: Important features (Decision Tree – SMOTE)

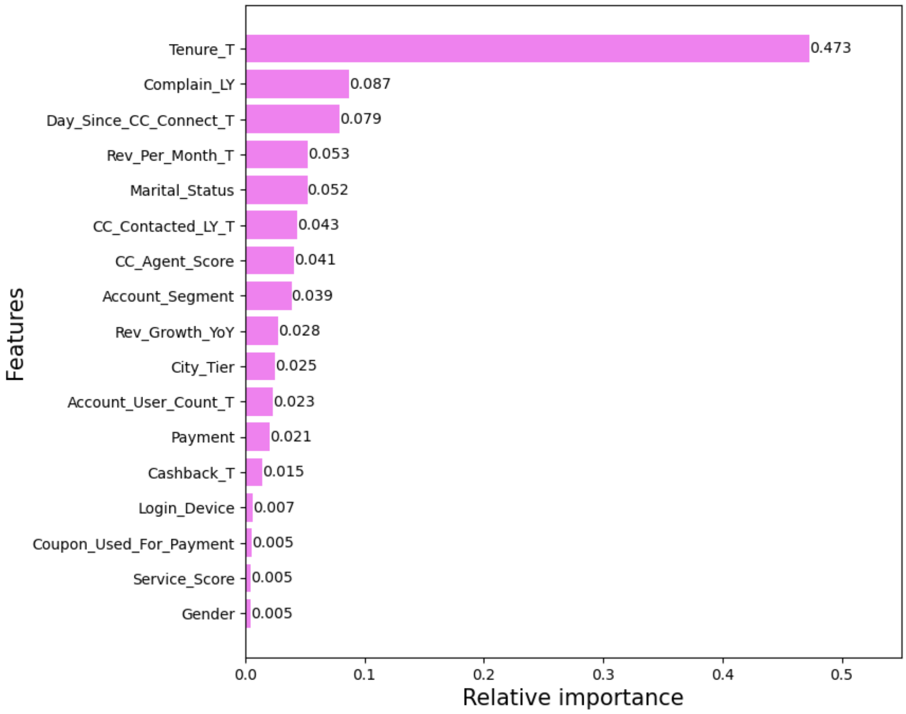


Figure 47: Important features (Tuned Decision Tree)

Figures 45, 46 and 47 show that 'Tenure' is the most important feature in Decision Tree models. 'Day_Since_CC_Connect' and 'CC_Contacted_LY' are the other key variables.

ADA Boost models

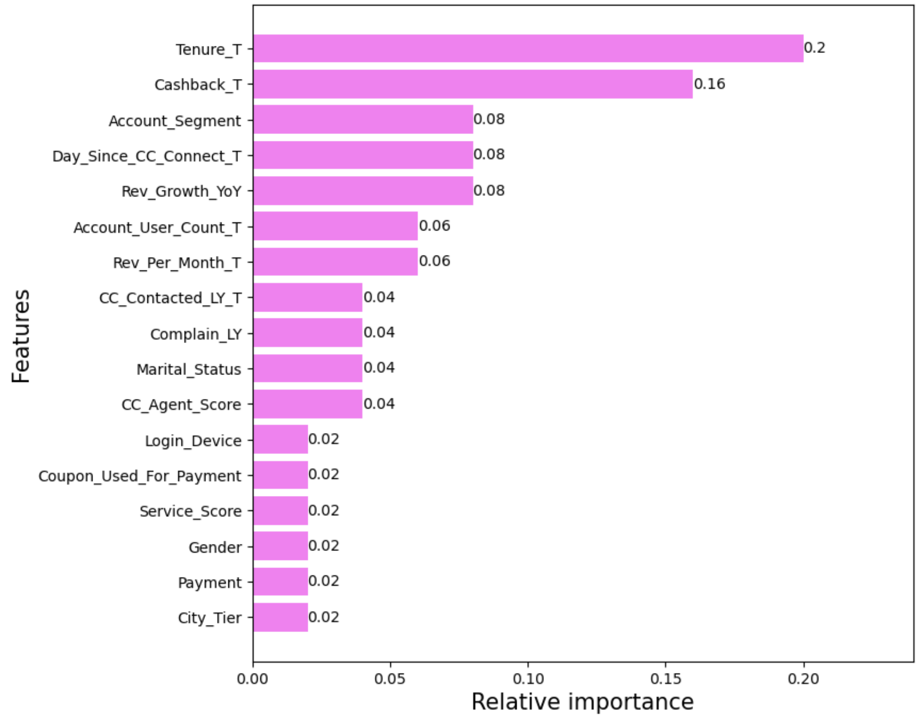


Figure 48: Important features (ADA Boost)

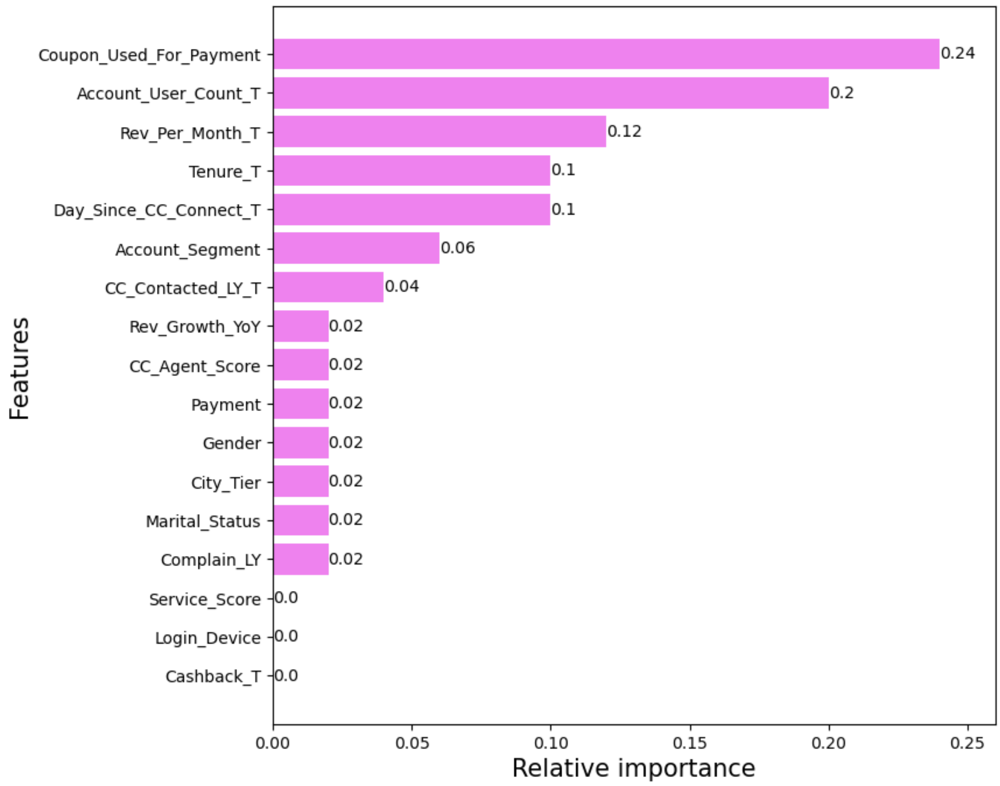


Figure 49: Important features (ADA Boost – SMOTE)

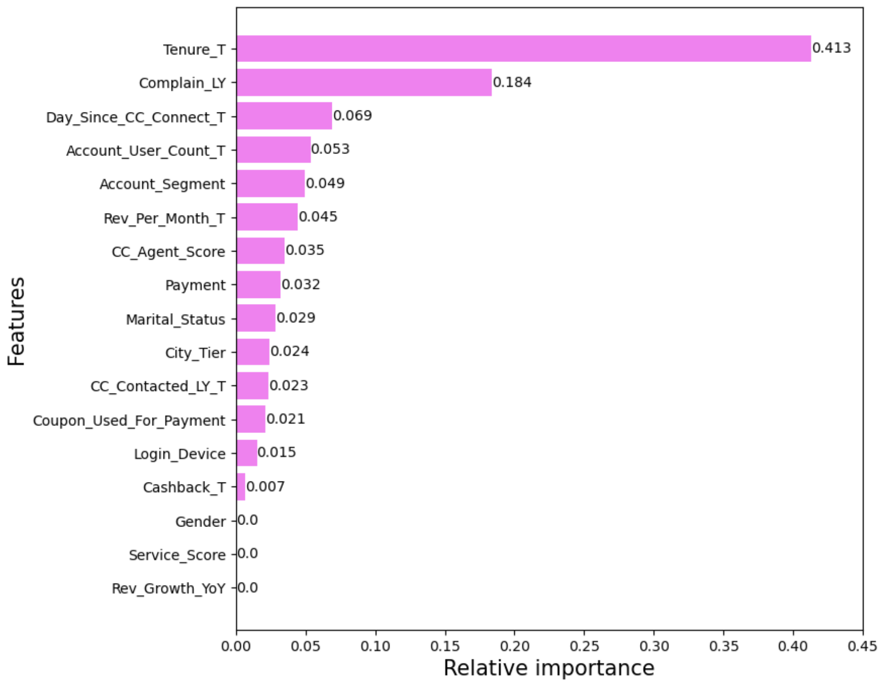


Figure 50: Important features (Tuned ADA Boost)

Figures 48 and 50 show that 'Tenure' is the most important feature in ADA Boost with imbalanced data and tuned ADA Boost. 'Day_Since_CC_Connect' is another important variable.

Gradient Boost models

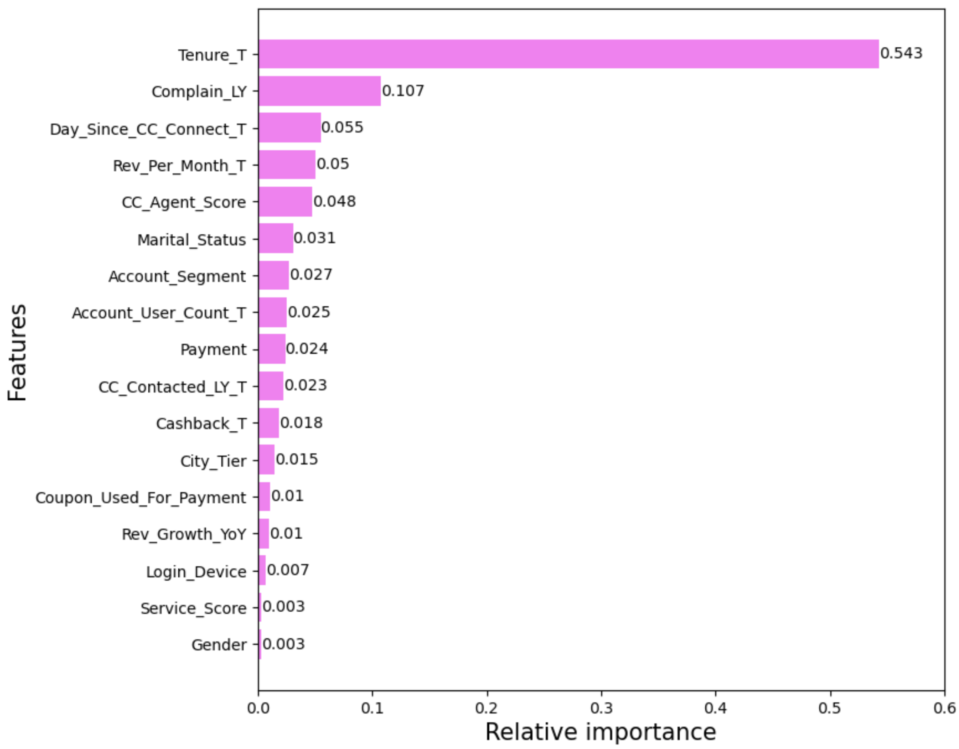


Figure 51: Important features (Gradient Boost)

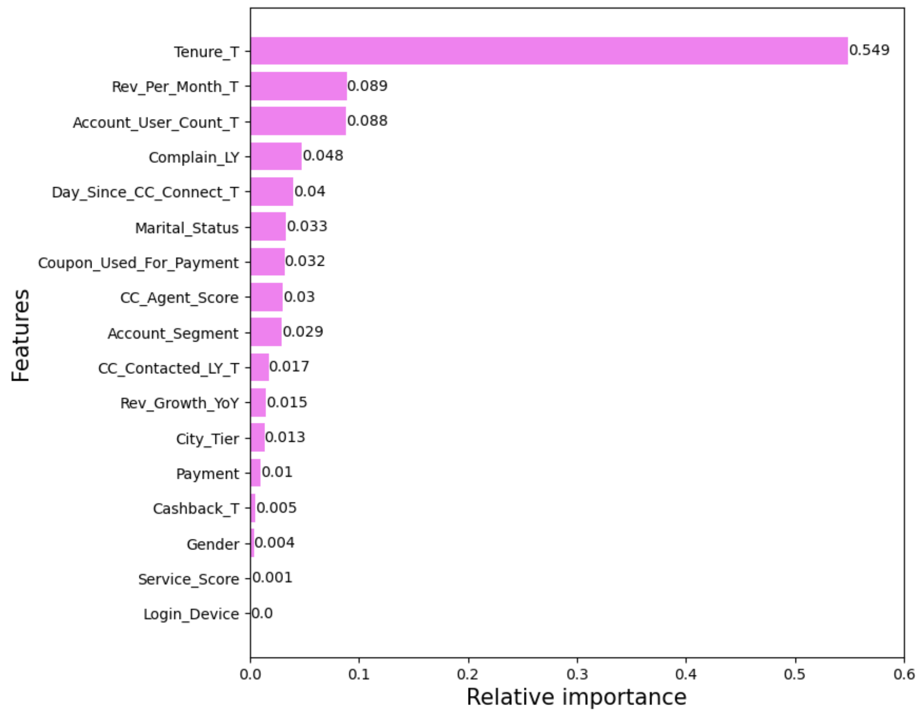


Figure 52: Important features (Gradient Boost – SMOTE)

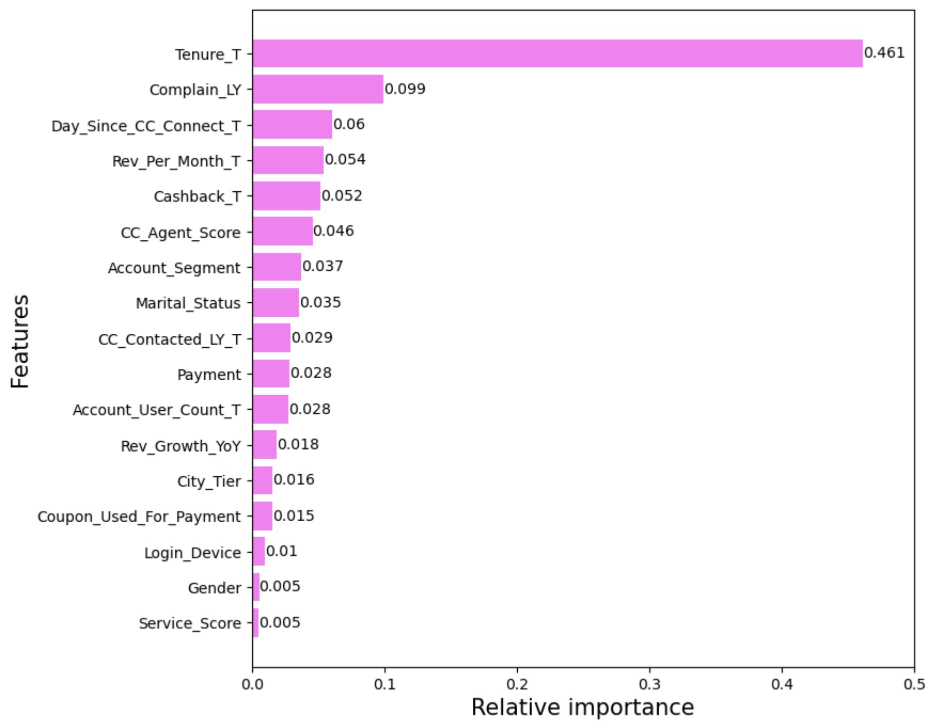


Figure 53: Important features (Tuned Gradient Boost)

Figures 51, 52 and 53 show that Tenure is the most important feature in the Gradient Boost models. 'Complain_LY' and 'Day_Since_CC_Connect' are other key variables.

Model validation

In this section, performance metrics such as accuracy, precision, recall, F1 score and AUC (Area Under Curve) score of all 26 models will be compared for Class 1 (customer has churned). On the basis of performance metrics, the best model will be chosen.

	Train accuracy	Test accuracy	Train recall	Test recall	Train precision	Test precision	Train F1	Test F1	Train AUC	Test AUC	Model overfits?
Logit model	0.88	0.88	0.47	0.44	0.73	0.75	0.57	0.55	0.86	0.85	No
LDA model	0.88	0.88	0.52	0.50	0.69	0.70	0.59	0.58	0.86	0.85	No
Naïve Bayes	0.87	0.87	0.53	0.50	0.63	0.66	0.58	0.57	0.86	0.84	No
SVM model	0.90	0.90	0.53	0.49	0.81	0.83	0.64	0.62	0.90	0.90	No
KNN model	0.97	0.95	0.87	0.77	0.96	0.92	0.91	0.84	0.99	0.98	Yes
Random Forest	1	0.97	1	0.86	1	0.98	1	0.91	1	0.99	Yes
Decision Tree	1	0.95	1	0.86	1	0.83	1	0.84	1	0.91	Yes
Logit SMOTE	0.80	0.80	0.80	0.79	0.80	0.45	0.80	0.58	0.87	0.85	No
LDA SMOTE	0.81	0.80	0.80	0.79	0.81	0.45	0.81	0.57	0.86	0.85	No
NB SMOTE	0.80	0.79	0.80	0.77	0.80	0.44	0.80	0.56	0.86	0.83	No
SVM SMOTE	0.87	0.85	0.88	0.82	0.85	0.54	0.87	0.65	0.93	0.90	Yes
KNN SMOTE	0.98	0.92	1	0.96	0.95	0.70	0.98	0.81	1	0.98	Yes
RF SMOTE	1	0.97	1	0.89	1	0.93	1	0.91	1	0.99	Yes
DT SMOTE	1	0.91	1	0.77	1	0.73	1	0.75	1	0.86	Yes
Tuned KNN	1	0.98	1	0.89	1	0.97	1	0.93	1	0.99	Yes
Tuned RF	0.90	0.89	0.48	0.46	0.83	0.84	0.60	0.59	0.93	0.91	No
Tuned DT	0.92	0.91	0.69	0.63	0.83	0.76	0.75	0.69	0.94	0.91	No
Ada Boost	0.90	0.90	0.59	0.59	0.76	0.77	0.66	0.67	0.92	0.91	No
Ada SMOTE	0.87	0.86	0.87	0.75	0.88	0.56	0.87	0.64	0.95	0.89	No
Tuned Ada	0.90	0.90	0.58	0.56	0.75	0.76	0.65	0.65	0.90	0.89	No

Gradient Boost	**0.92**	**0.91**	**0.64**	**0.60**	**0.85**	**0.83**	**0.73**	**0.70**	**0.95**	**0.93**	**No**
Gr Boost SMOTE	0.92	0.89	0.91	0.73	0.93	0.67	0.92	0.70	0.98	0.92	Yes
Tuned Gr Boost	0.94	0.92	0.74	0.66	0.90	0.85	0.81	0.74	0.97	0.95	No
Bagging model	0.99	0.96	0.97	0.79	1	0.98	0.98	0.87	1	0.99	Yes
Bagging SMOTE	1	0.96	1	0.88	1	0.87	1	0.88	1	0.99	Yes
Tuned Bagging	0.88	0.88	0.37	0.37	0.82	0.87	0.51	0.52	0.91	0.90	No

Table 10: Comparison of all ML models

It can be seen from Table 10 that some of the models such as KNN, SVM, Random Forest and Decision Tree are overfitting. It means that the model tries to capture every data point in the training set, but comes a cropper on the unseen data.

The best model is Gradient Boost. This is so because its performance on both training and test sets is consistent. Accuracy is pretty high and so is precision, F1 score and AUC score. Recall, however, is low but it is comparatively better than other models. That being said, **our focus is on high precision**.

Precision measures the percentage of predictions made by the model that are correct. In simple terms, precision is the ratio of true positives and the sum of true positives and false positives.

For the customer churn dataset, **our objective is to have fewer false positives** – data points labelled as positive that are actually negative. In other words, customers predicted as having churned actually stay.

If a model predicts high false positives, the company would direct its resources/revenue in the wrong direction. It would spend on customers who the model predicts would churn but actually they remain with the company. The firm would not want such a scenario. Therefore, **it would be in the best interest of the company if it deploys the Gradient Boost model that has high precision on both training and unseen datasets.**

Recommendations

Recommendations are based on the top five features (Figure 51) of the Gradient Boost model.

1. **Tenure:** The business team must focus on increasing the tenure of customers by offering them special pricing and long-term plans. As it has been seen that new customers churn more, they must be offered 'loyalty plans'.
2. **Complain Last Year:** The team must ensure that complaints are handled promptly. The e-commerce company must ensure that customer care service is proficient enough to deal with complaints.
3. **Days Since CC Connect:** Some users churn silently. Even if users have not contacted the customer care for days, the company must conduct telephone surveys to get the clients' feedback.
4. **Revenue Per Month:** Churned customers are generating slightly more revenue. Therefore, the focus should be on finding the root cause of their complaints so that such users can be targeted in a better manner.
5. **CC Agent Score:** The majority of the customers gives a rating of 3 to customer care agents. The agents must ask for feedback so that new policies can be devised that will enhance the customers' experience.

Some other recommendation are as follows:

- The churn rate is high in Tier 2 cities (Figure 17), where the company has the least number of subscribers (Figure 8). This is a cause for concern. The company should devise policies to make a dent in the untapped market and look into the reasons for the high churn rate.
- Since single persons are more likely to churn (Figure 13), they must be offered with subscription plans that they can use with their parents and friends.
- The Regular Plus account, which is the most popular category (Figure 10), has the potential to generate revenue, but it has the highest churn rate (Figure 15). Even HNI users have comparatively a high churn rate. The company should come up with dedicated plans to target them.

Source: Great Learning logo that has been used on the cover page has been taken from one of the monographs provided by Great Learning.