Customer Churn Prediction
Business Report - 3

By: Ambrish Verma

Date: 16-June-2024

Contents

Content - Figures and tables

1. **Introduction**:

1.1. Problem Statement Definition:

- Based on the independent columns provided in the dataset for the e-commerce company, predict the value for the target column: 'Churn'.
- Identify the segments on which retention offers can be rolled out.
- Based on the predictions for the segments, provide the business recommendations for the campaign the company may utilize to minimize the number of customer/account churn, while avoiding over-subsidization of the offerings.

1.2. Need of the study:

The current e-commerce market is highly competitive. Customer retention is necessary for continued revenue growth. Detailed analysis of the provided dataset is required to extract maximum information of the customer's impression about the company and its performance. Moreover, there is a need to categorize the customers in different segments so that we can recognize the pattern and correctly predict if a customer can churn soon.

1.3. Understanding Business/Social Opportunity:

If the predictions are accurate there are multiple potential benefits:

- The customer attrition rate can be reduced, thereby benefiting the company.
- The company can showcase its competitiveness and frugality to the shareholders and investors in maintaining an edge over its competitors.
- Low attrition rate of existing customers can server as an attractive yardstick for potential new customers.

2. Exploratory Data Analysis:

2.1 Uni-variate analysis: Boxplot

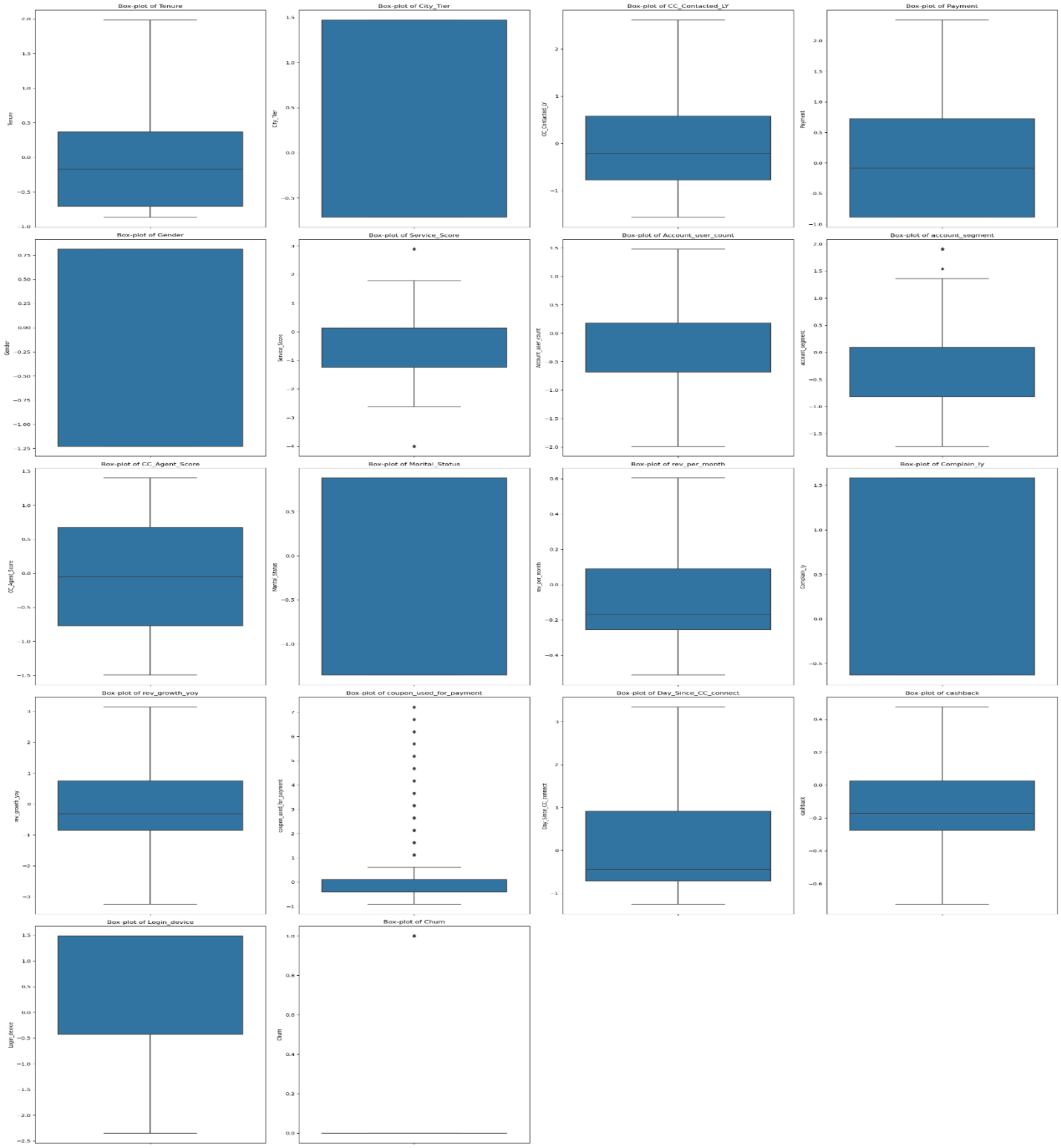

Figure 1 EDA: boxplot

Analysis:

- The attributes: Gender, City_tier, Marital_status, complain_ly do not have any outliers present and the entire data fits inside the Inter quantile range.
- Very few records have the Churn=1. This indicates highly unbalanced data.
- The attributes: Cashback, service_score, rev_growth_yoy and account_user_count follow normal distribution with low skewness.

2.2 Bi-variate analyses: Distribution plot

1. Cashback versus Churn:

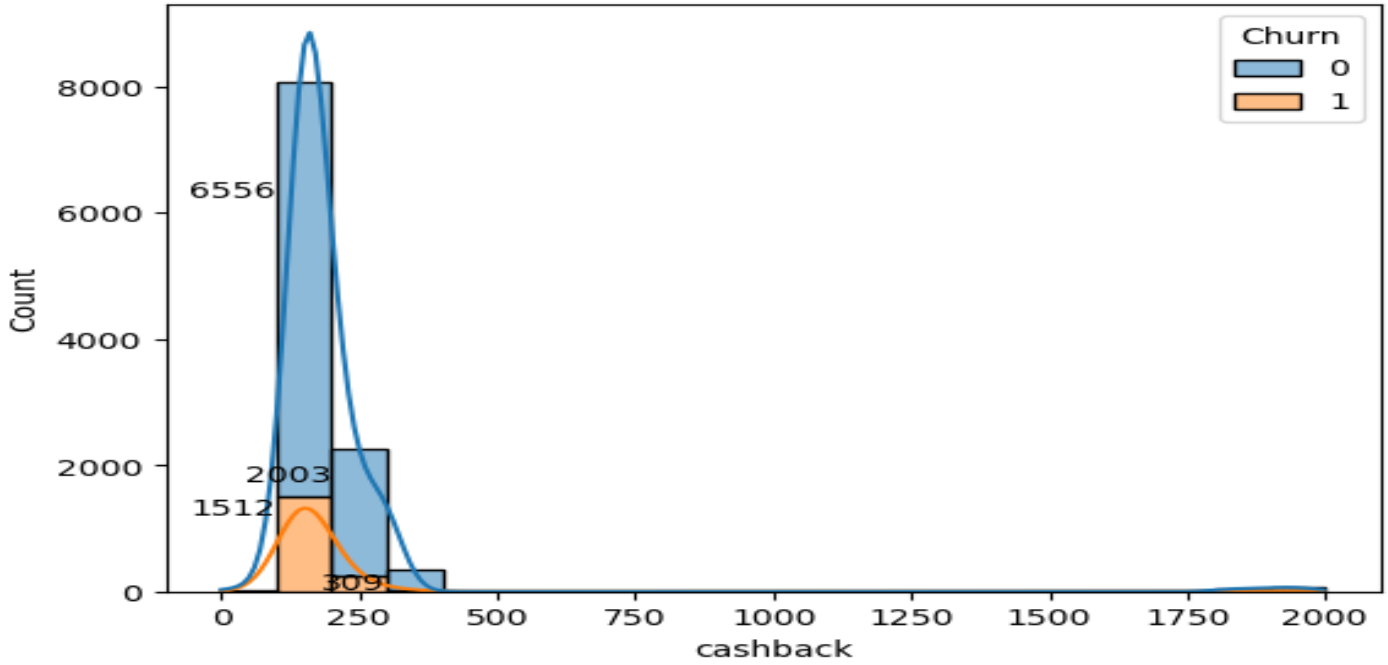

Figure 2 Cashback versus Churn

- Analysis: Average cashback of INR.175 is the most prevalent. The Churn rates: 0 and 1 follow normal distribution with Churn=1 being right skewed. It reaches a peak at around INR. 150, drops to insignificant levels by around INR. 350.

2. Tenure versus Churn:

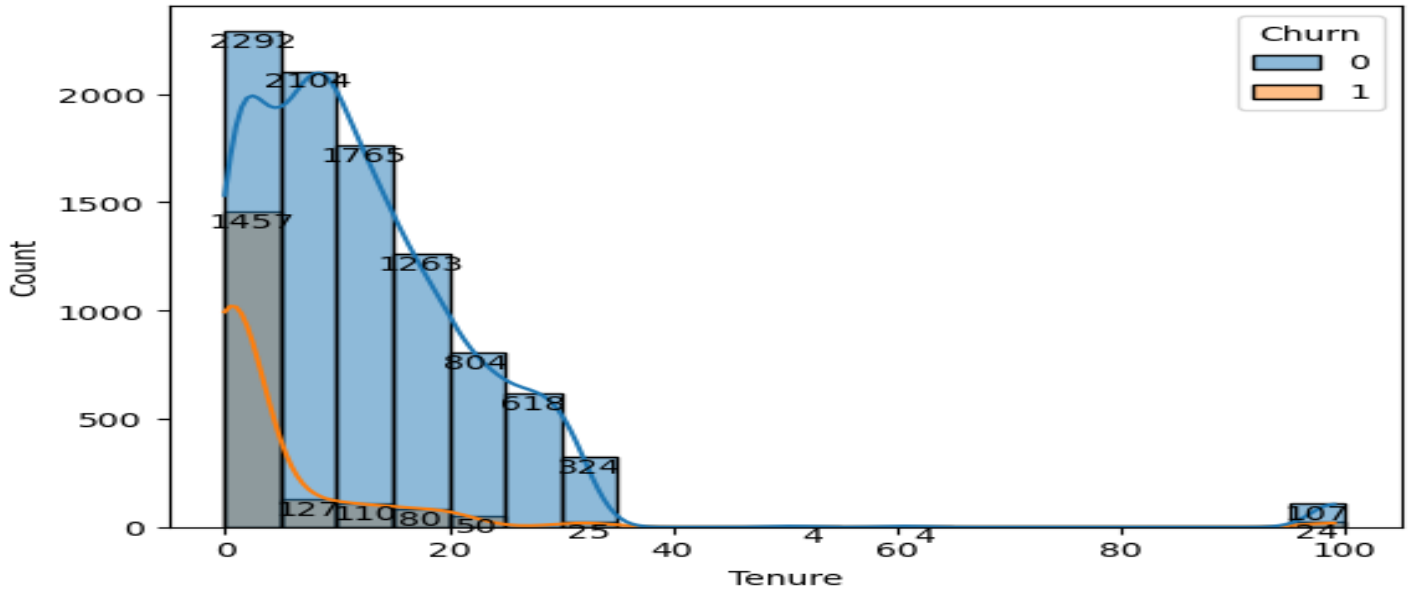

Figure 3 Tenure versus Churn

- Analysis: In the first week of customer onboarding, the Churn is very high (almost 40% of all customers onboarded). After the first 20 days, the Churn rate decreases substantially. This has been illustrated in the below table:

Attribute	Value	Total	Count: Churn =0	Count: Churn=1	Churn =1 (% of total)
Tenure	5	3749	2292	1457	38.86369699
	10	2231	2104	127	5.692514567
	15	1875	1765	110	5.866666667
	20	1343	1263	80	5.956813105
	25	857	807	50	5.834305718
	30	618	618	0	0
	35	349	324	25	7.163323782

Table 1 Churn distribution versus Tenure

3. Days_since_CC_connect versus Churn:

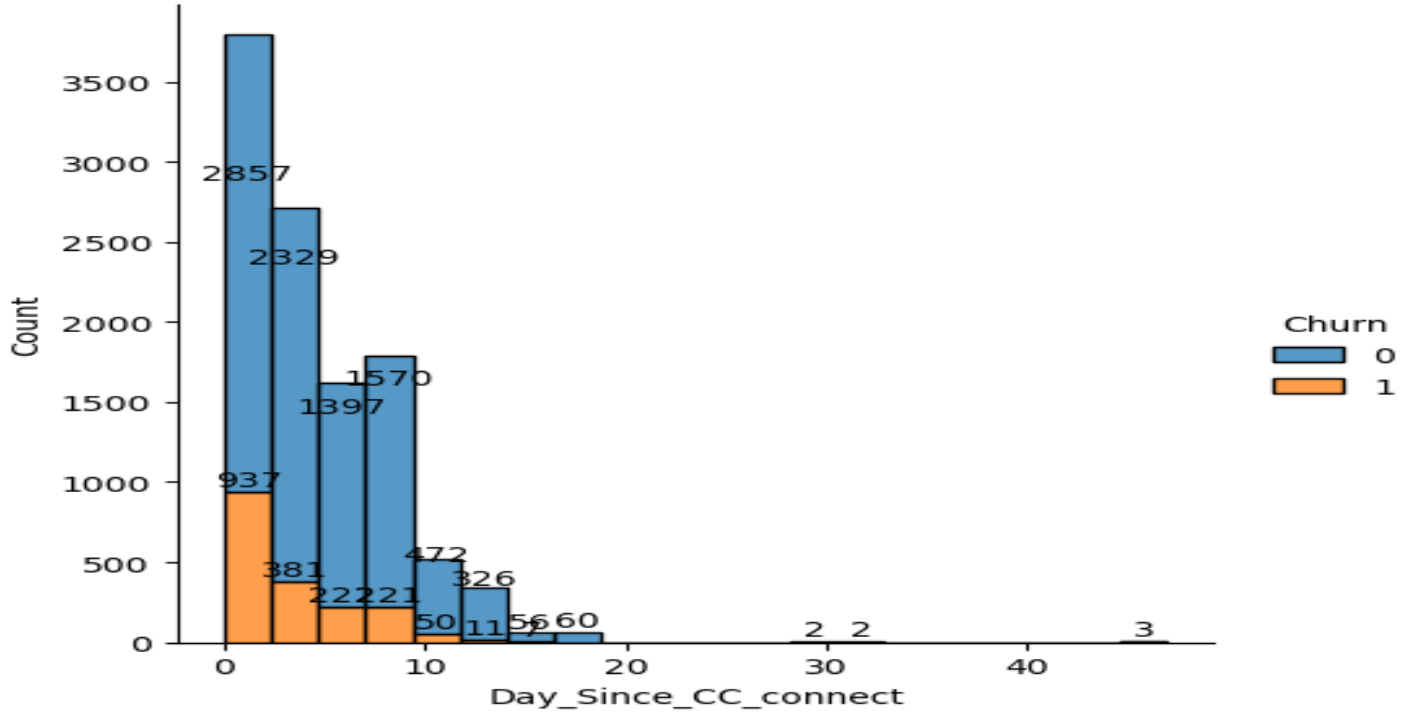

Figure 4 Days_since_CC_connect versus Churn

- Analysis:
  - The Churn is particularly high within the first few days of the Customer connecting with the customer care. Details are mentioned below:

Attribute	Value	Total	Count: Churn =0	Count: Churn=1	Churn =1 (% of total)
Day_Since_CC_Connect	1	964	638	326	33.81742739
	2	1256	880	376	29.93630573
	3	1574	1339	235	14.93011436
	4	1817	1552	265	14.58447991
	5	893	777	116	12.98992161
	6	479	424	55	11.4822547
	7	229	197	32	13.97379913
	8	911	776	135	14.81888035
	9	1169	1012	157	13.43028229
	10	622	558	64	10.28938907

Table 2 Churn distribution with respect to Days_since_CC_connect

4. CC_Agent_Score versus Churn:

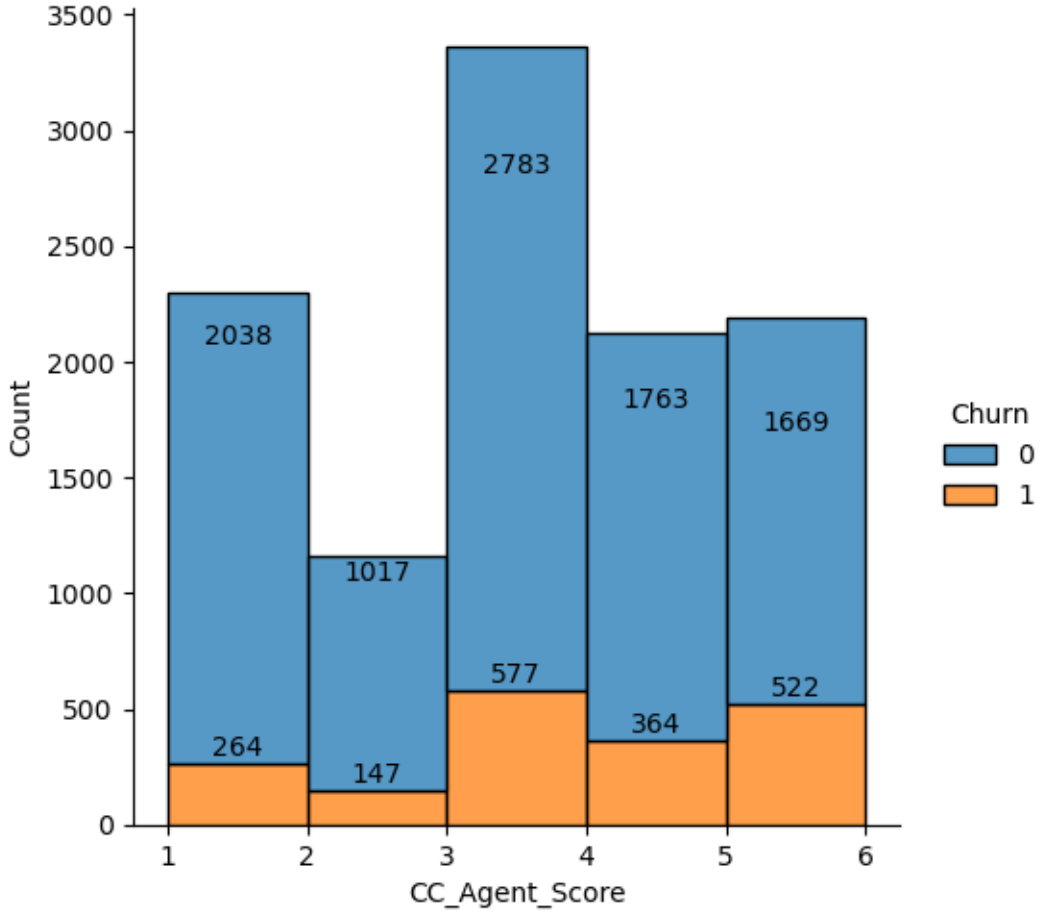

Figure 5 CC_Agent_Score versus Churn

- Analysis:
  - The Churn rate consistently increases from score =1 to score =5.

Attribute	Value	Total	Count: Churn =0	Count: Churn=1	Churn =1 (% of total)
CC_Agent_Score	1	2302	2038	264	11.46828844
	2	1164	1017	147	12.62886598
	3	3360	2783	577	17.17261905
	4	2127	1763	364	17.11330512
	5	2191	1669	522	23.82473756

Table 3 Churn distribution versus CC_Agent_Score

5. Marital Status versus Churn:

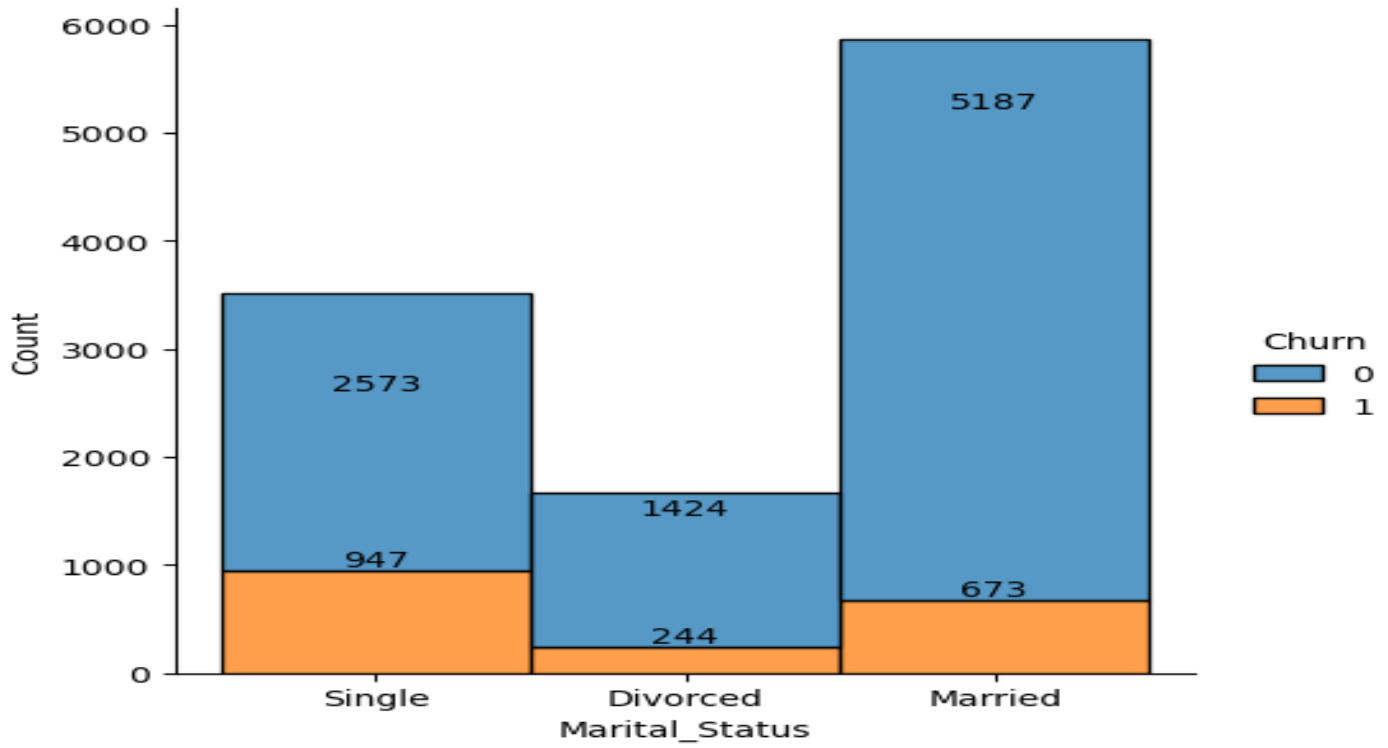

Figure 6 Marital Status versus Churn

6. Gender versus Churn

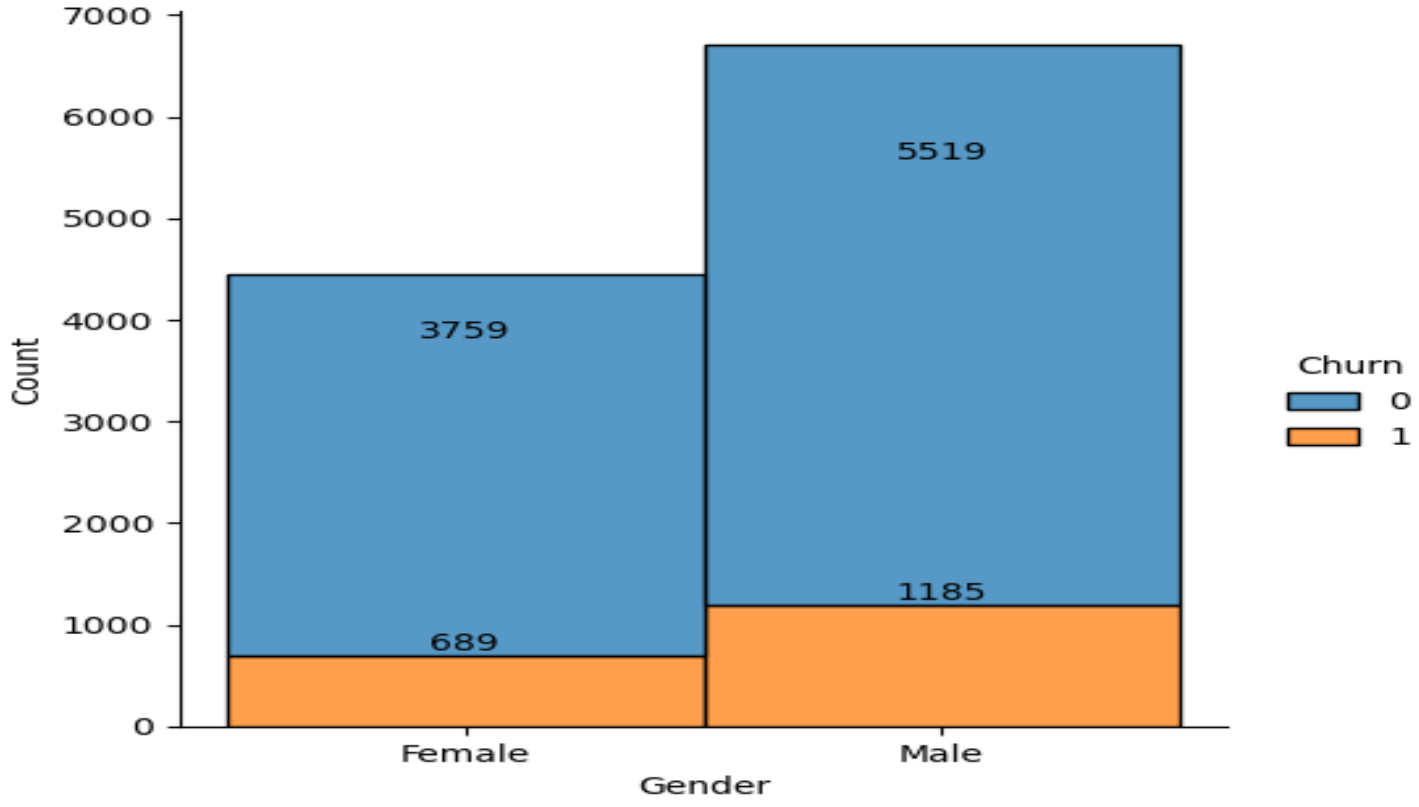

Figure 7 Gender versus Churn

7. City_Tier versus Churn

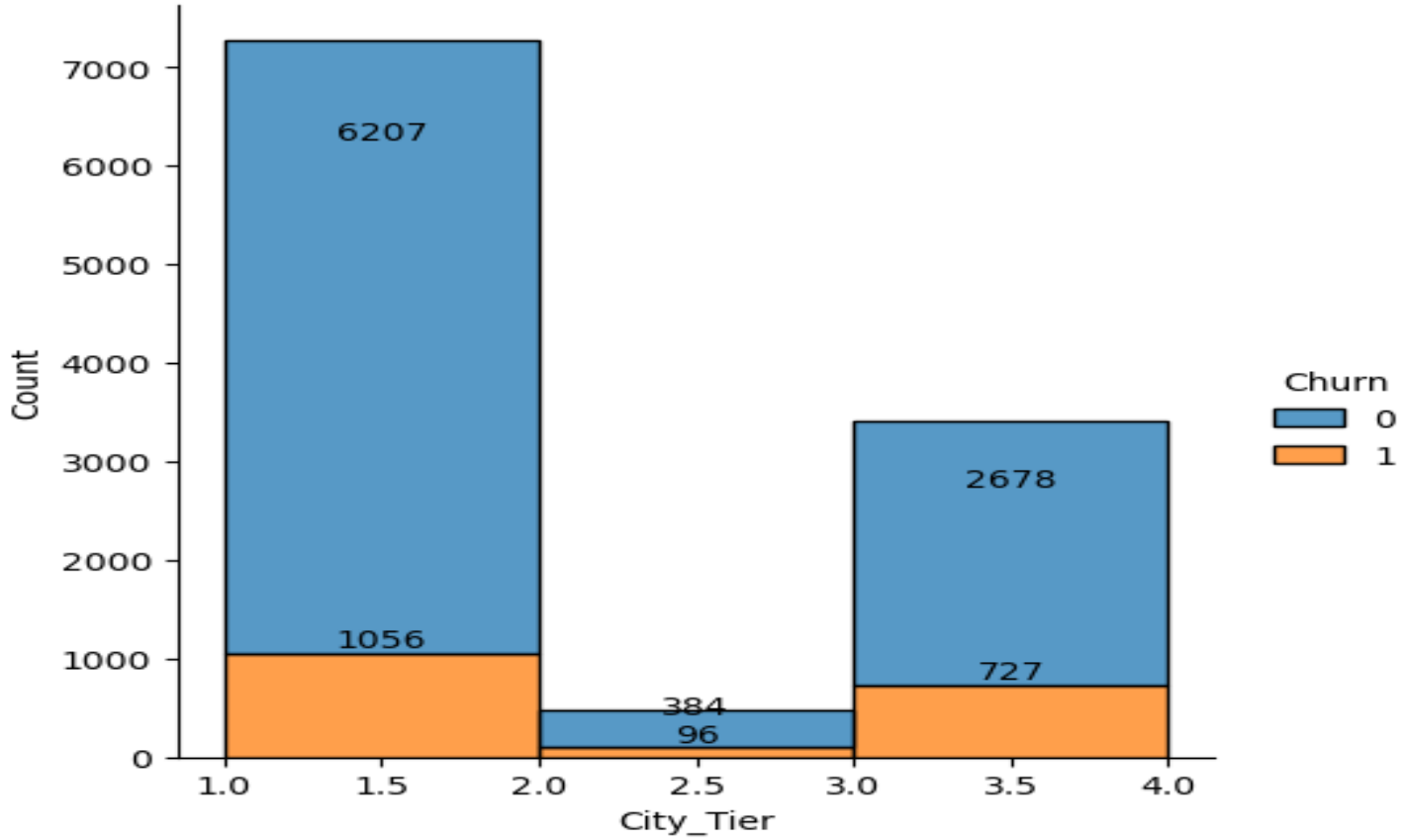

Figure 8 City Tier versus Churn

- Analysis:
  - If the customer is single, the churn rate is significantly higher. It then reduces if the person is divorced and decreases further if he/she is married.
  - Churn rate is not varying much based on the Gender of the customer.
  - In the tier 2 and tier 3 cities, the Churn rate jumps significantly compared to tier 1 cities.

Attribute	Value	Total	Count: Churn =0	Count: Churn=1	Churn =1 (% of total)
Marital_Status	Single	3520	2573	947	26.90340909
	Divorced	1668	1424	244	14.62829736
	Married	5860	5187	673	11.48464164
Gender	Female	4448	3759	689	15.49010791
	Male	6704	5519	1185	17.67601432
City_Tier	1	7263	6207	1056	14.53944651
	2	480	384	96	20
	3	3405	2678	727	21.35095448

Table 4 Churn distribution with columns: Gender, Marital Status, City Tier

8. Login_device versus Churn

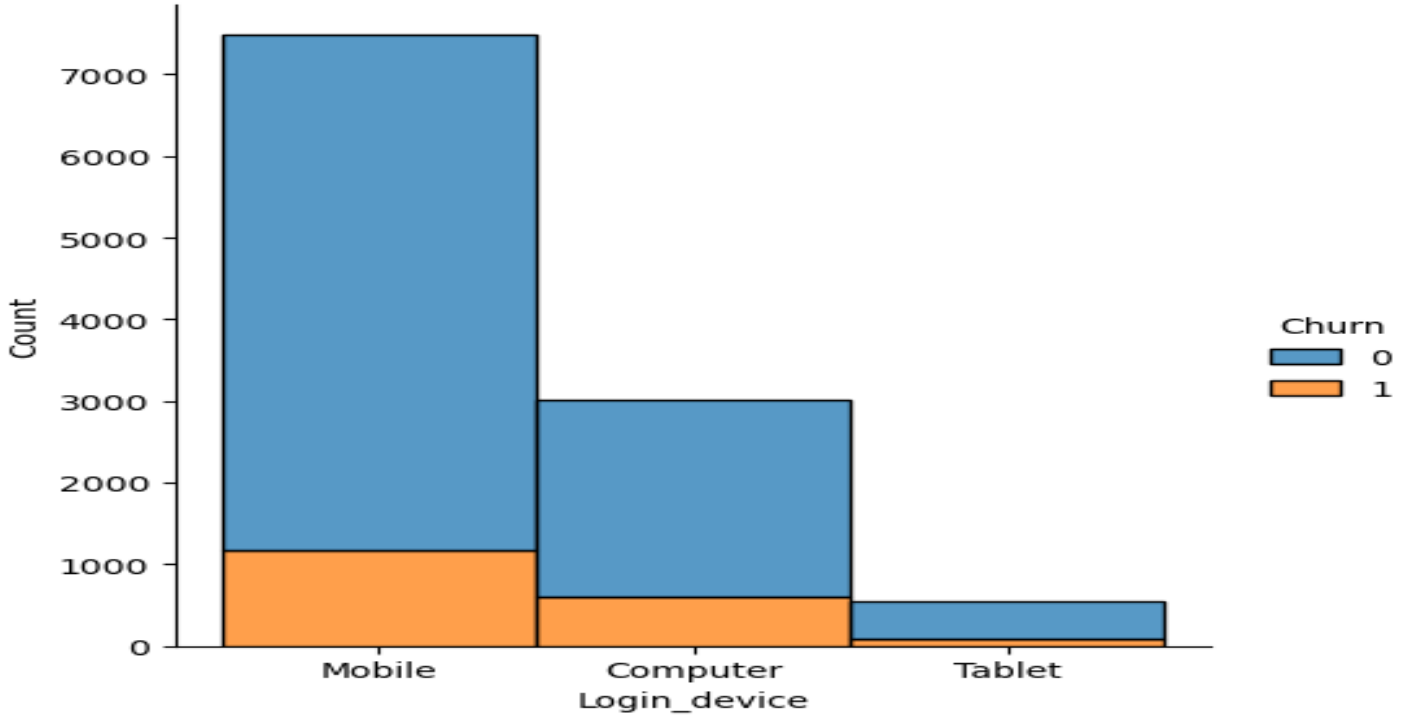

Figure 9 Login Device versus Churn

9. Complain_ly versus Churn:

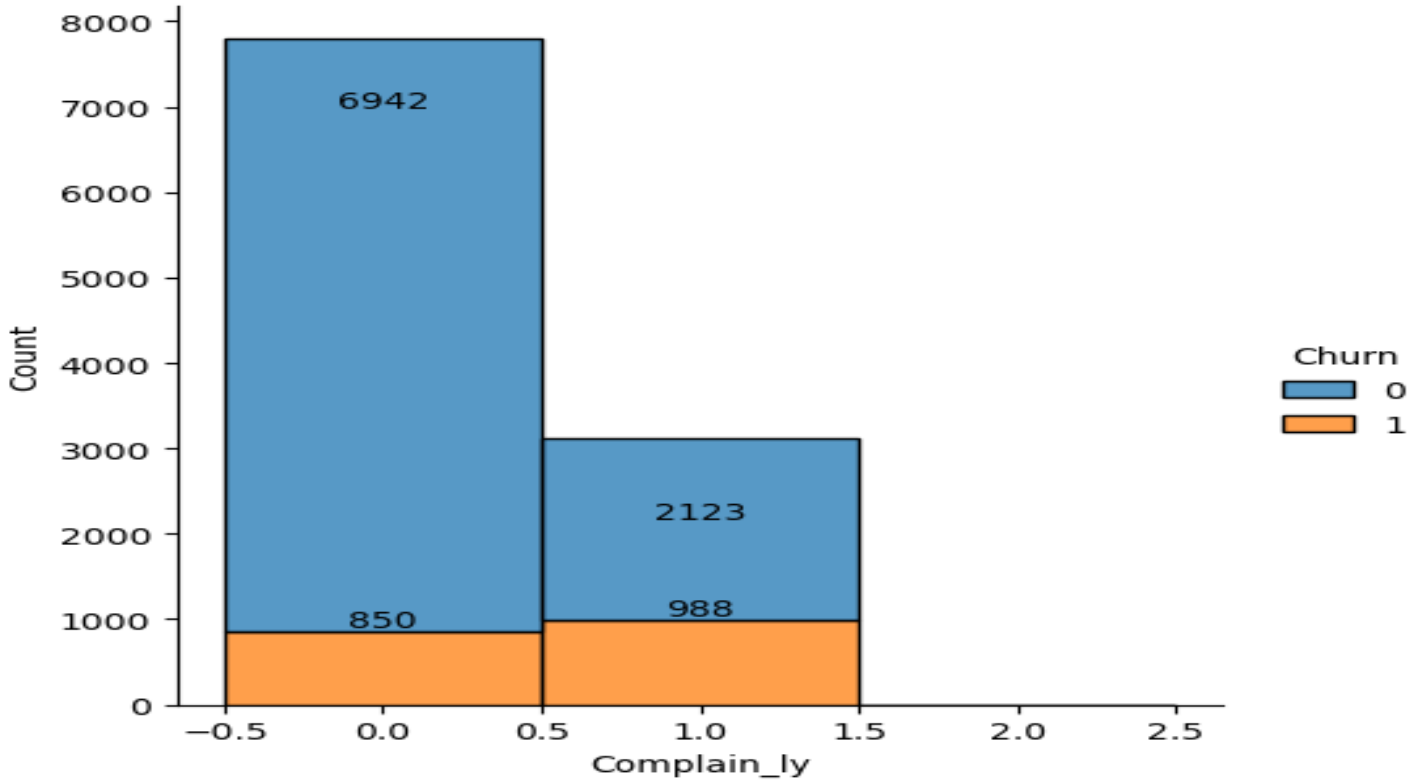

Figure 10 Complain_ly versus Churn.

- Analyses:
    - The Churn rate for mobile and tablet is similar whereas, when the login device is computer, the churn rate is higher.
    - If there was a complaint raised in the last 1 year, the chances of Customer churn are almost 3 times compared to churn when no complaints were being raised in the last 1 year.

Attribute	Value	Total	Count: Churn =0	Count: Churn=1	Churn =1 (% of total)
Login Device	Mobile	7482	6310	1172	15.66426089
	Computer	3018	2421	597	19.78131213
	Tablet	539	454	85	15.76994434
Complain_ly	0	7792	6942	850	10.90862423
	1	3111	2123	988	31.75827708

2.3 Multi-variate analysis: Correlation matrix

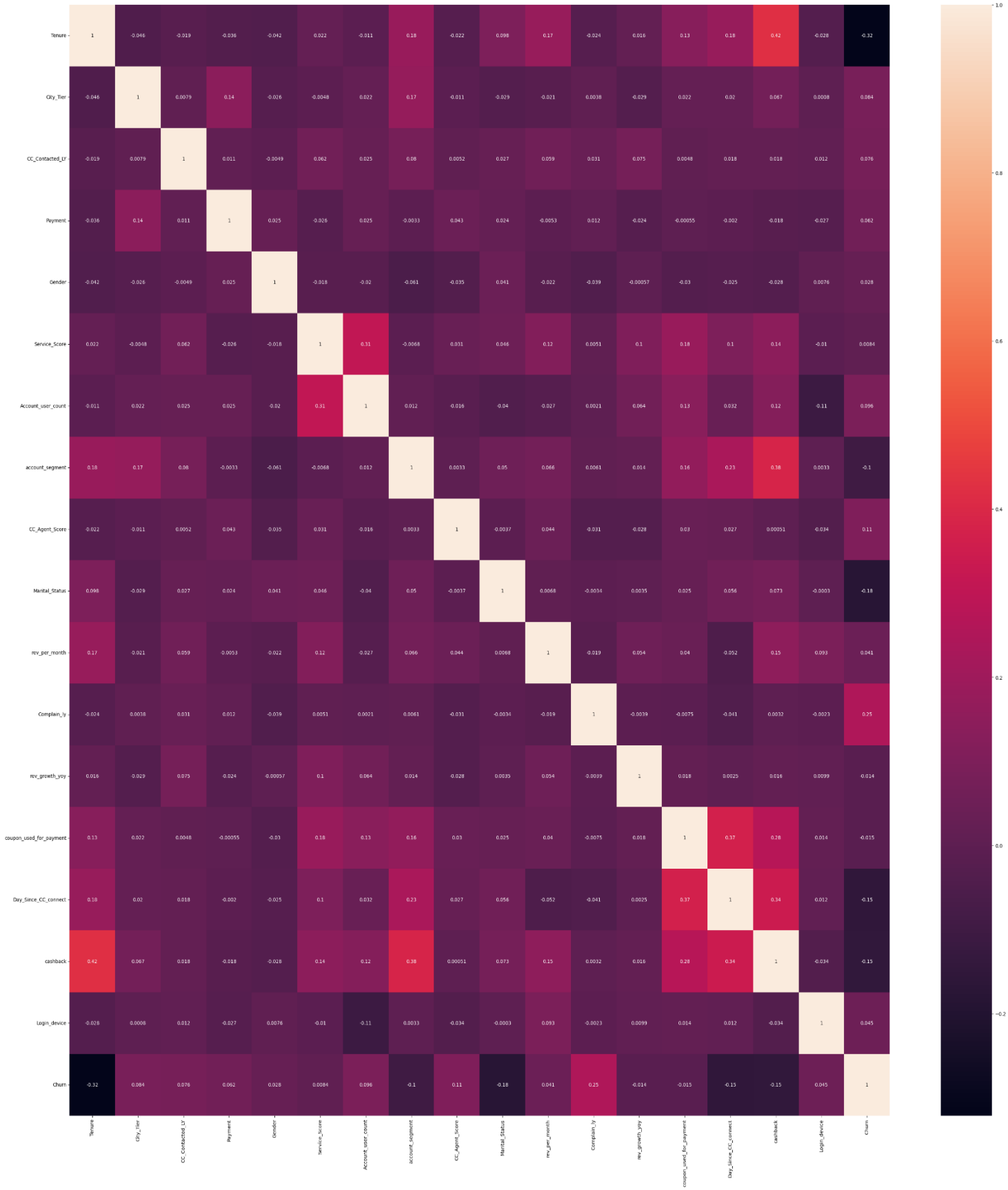

Figure 11 Correlation matrix

- Analyses:
    - The attributes ‘Tenure’ and ‘Cashback’ have a relatively higher degree of correlation.
    - The attributes ‘Cashback’ and ‘Account segment’ are positively correlated.
    - The attributes ‘Cashback’ and ‘Days_since_CC_Connect’ are also, relatively correlated.
    - The attributes ‘Coupon_used_for_payment’ and ‘Days_since_cc_connect’ are also correlated.
    - ‘Tenure’ and ‘Churn’ have a high negative correlation.
    - ‘Marital_status’ and ‘churn’ are negatively correlated.

3. **Data Cleaning and Pre-processing**

- Total records in the dataset: 11260
- Total attributes present in the dataset: 19.
- The data for Churn was found to be highly unbalanced. Ratio of Churn =0 to Churn =1 was found to be 83:17.
- The attributes, their datatypes and number of non-null values present are shown in the below table:

Attributes	Non-null Count	Null Count	Numeric datatype?
AccountID	11260	0	Y
Churn	11260	0	Y
Tenure	11158	102	N
City_Tier	11148	112	Y
CC_Contacted_LY	11158	102	Y
Payment	11151	109	N
Gender	11152	108	N

Service_Score	11162	98	Y
Account_user_count	11148	112	N
account_segment	11163	97	N
CC_Agent_Score	11144	116	Y
Marital_Status	11048	212	N
rev_per_month	11158	102	N
Complain_Ly	10903	357	Y
rev_growth_yoy	11260	0	N
coupon_used_for_payment	11260	0	N
Day_Since_CC_connect	10903	357	N
cashback	10789	471	N
Login_device	11039	221	N

Table 5 Attributes Information

- 1.2% of all data was found to be null.
- Moreover, there was data quality issues identified in multiple columns. They were fixed as shown in the following table:

Attribute	Original Value	Replacement Value
Gender	M	Male
	F	Female
Account_user_count	@	7
account_segment	Regular +	Regular Plus
	Super +	Super Plus
coupon_used_for_payment	#	1
	$	1
	*	1
Login_device	&&&&	Tablet
Tenure	#	32
Day_Since_CC_connect	$	3
rev_growth_yoy	$	14
rev_per_month	+	0

Table 6 Data fix performed.

- Outlier treatment: Outlier treatment was performed once the non-numeric attributes were converted to numeric ones after the data fix as shown above.
    - The lower quantile(Q1) and upper quantile(Q3) were calculated and defined to be 25 percentile and 75 percentiles of the dataset.
    - Interquartile range (IQR) was defined as Q3-Q1.
    - Lower Level that an attribute data was made to fall under was defined as Q1-(1.5*IQR)

    - Upper level that an attribute data was made to fall under was defined as Q3+(1.5*IQR)
- Not all attributes were made to undergo the outlier treatment because of the unbalanced dataset. The attributes for which outlier treatment was performed were: ‘Tenure','Day_Since_CC_connect','CC_Contacted_LY', 'Account_user_count','rev_per_month', and 'cashback'.
- Before any null imputation, encoding of data was performed as illustrated in the table below:

Attribute	Value	Encoded value
Payment	Debit Card	0
	Credit Card	1
	E wallet	2
	Cash on Delivery	3
	UPI	4
Gender	Female	0
	Male	1
account_segment	Regular	0
	Regular Plus	1
	Super	2
	Super Plus	3
	HNI	4
Marital_Status	Single	0
	Divorced	1
	Married	2
Login_device	Tablet	0
	Mobile	1
	Computer	2

Table 7 Attributes’ Encoded values

- Prior to imputing null values, the dataset was scaled. The target column: ‘Churn’ was removed, and the null imputation was performed using KNN (K – Nearest neighbour) imputation. This was done because KNN imputation should not be performed on the target column. The data was scaled, and then null values were imputed.
- The dataset was scaled using standardscalar library in python.
- KNN imputation was performed on the independent columns and then the target column was joined to the treated dataset.

4. Model building

- Prior to machine learning model building, following steps were performed:
    - Identification of attributes with high Variance Inflation Factor (VIF) to identify if there were any attribute who had high degree of multicollinearity and could be dropped. A threshold of VIF =5 was chosen and none of the available attributes were found to have VIF>5. The highest VIF recorded was 1.6.
    - The VIF values are displayed below:

Attribute	VIF
cashback	1.605351
Tenure	1.323856
Day_Since_CC_connect	1.298326
Churn	1.298149
coupon_used_for_payment	1.228979
account_segment	1.224701
Service_Score	1.17124
Account_user_count	1.159163
rev_per_month	1.134649
Complain_ly	1.071332
City_Tier	1.069213
Marital_Status	1.043022
Login_device	1.031131
Payment	1.03
CC_Contacted_LY	1.026568
CC_Agent_Score	1.026151
rev_growth_yoy	1.021302
Gender	1.013575

Table 8 Attributes - Variance Inflation Factor

    - The dataset was split into 'Train' and 'Test' data sets. The model built would then be trained on the 'Train' dataset and then validated on the 'Test' dataset.
    - The ratio of 'Train' and 'Test' data was decided to be 70:30.
    - Since the business problem is to identify Churn = 0 or 1, only classification algorithms are chosen for models building.
- The base models assessed to solve the problem were:
    - Logistic Regression
    - LDA (Linear Discriminant Analysis)
    - Decision Tree
    - Random Forest

- Naïve Bayes
- KNN (K – Nearest Neighbour)

- The evaluation of the above models using 'Train' is provided below:

Model	Accuracy	Precision (pred = 0)	Precision (Pred = 1)	Recall (Pred = 0)	Recall (Pred = 1)	Incorrect pred = 0	Incorrect Pred =1	Correct pred =0	Correct Pred=1
Logistic Regression	0.879	0.89	0.75	0.97	0.43	763	190	6365	564
LDA	0.877	0.89	0.75	0.97	0.4	793	176	6379	534
Decision Tree(DT)	0.979	0.99	0.95	0.99	0.93	97	68	6487	1230
Random Forest(RF)	0.967	1	0.84	0.96	1	1	257	6298	1326
Naïve Bayes	0.863	0.91	0.6	0.92	0.57	575	504	6051	752
KNN	1	1	1	1	1	0	0	6555	1327

Table 9 Model evaluation metrics - Train data

- The models: Decision Tree, Random Forest and KNN were likely to have overfit. Hyper-parameter tuning was thus employed to reduce over fitting and enhance the evaluation metrics.
- Hyper-parameter tuning using Grid search was performed on the models wherever applicable. This enhanced the performance of the models: Random Forest, Decision Tree.
- For KNN, the K value was identified as 2. The same is shown in the graph below:

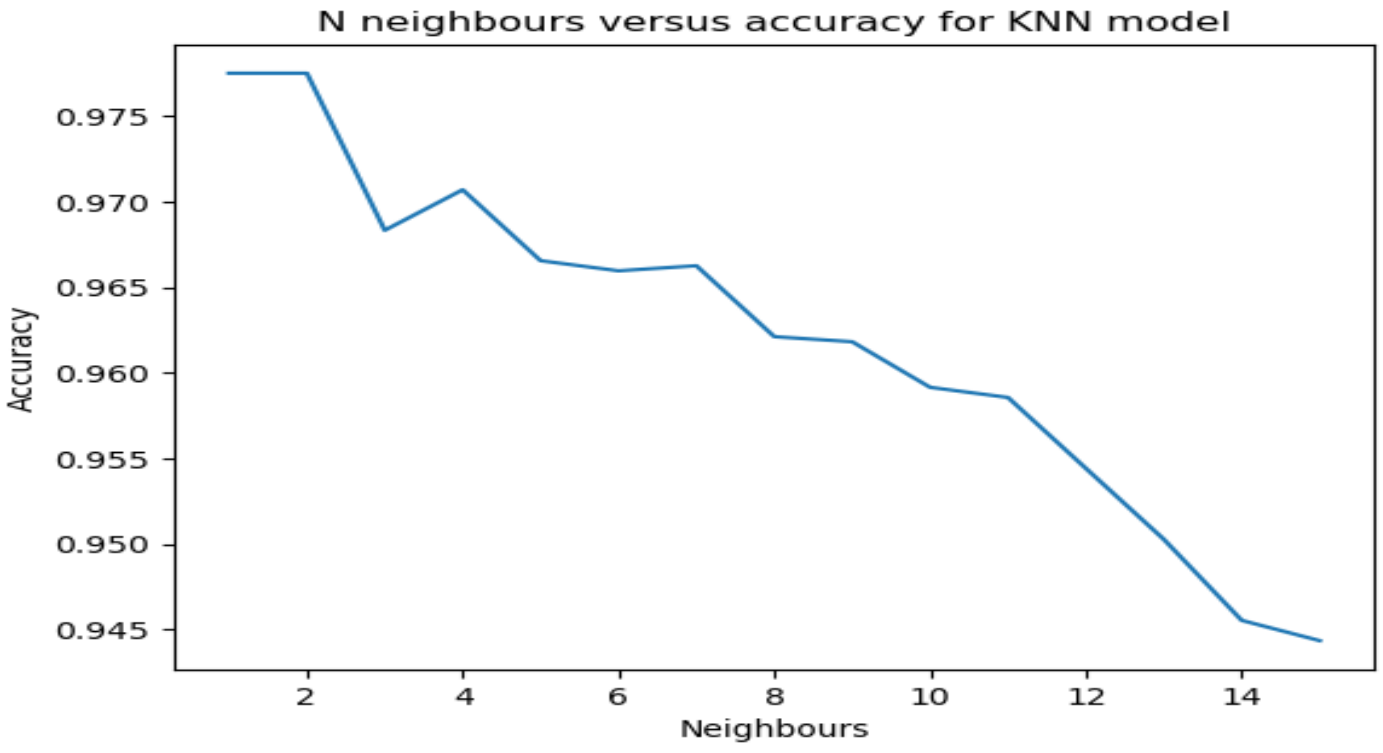

Figure 12 N-Neighbours versus accuracy

- Tuned parameters for Random forest are: 'Random Forest': RandomForestClassifier(class_weight={0: 1, 1: 10}, criterion='entropy', max_depth=16, max_features=11, min_samples_leaf=5, min_samples_split=10, n_estimators=220, random_state=42).
- The best tuned models identified were: Random Forest (Tuned) and KNN(Tuned).
- To further enhance the accuracy of the model that can be deployed, ensemble modeling technique was applied.
- Multiple iteration of ensemble of various models were tried to obtain the best combination of models. They were:
    - Ensemble 1: combination of Logistic Regression, LDA, Decision Tree(tuned), Random Forest(tuned), KNN (tuned).
    - Ensemble 2: Decision Tree(tuned), Random Forest(tuned), KNN (tuned)
    - Ensemble 3: Random Forest(tuned), KNN (tuned).

- All the above models were then evaluated using 'Test' data on the following metrics:
    - Accuracy
    - Confusion matrix
    - Recall of Churn prediction
    - AUC Curve
- The model that was identified as performing the best was **ensemble of tuned Random Forest and KNN models**.

5. **Model validation**

- All the models were trained on the 'Train' data. 7882 records were utilized to train the models. They were then evaluated on the 'Test' data, 3378 in number.
- All the models were validated using the metrics mentioned above. The relative performance metrics on the test data set obtained are shown in the table below:

Model	Accuracy	Precision (pred = 0)	Precision (Pred = 1)	Recall (Pred = 0)	Recall (Pred = 1)	Incorrect pred = 0	Incorrect Pred =1	Correct pred =0	Correct Pred=1
Logistic Regression	0.883	0.9	0.76	0.97	0.44	318	78	2731	251
LDA	0.881	0.89	0.76	0.97	0.43	327	75	2734	242
Decision Tree(DT)	0.938	0.96	0.82	0.96	0.81	109	99	2710	460
Random Forest(RF)	0.948	0.98	0.8	0.95	0.93	41	135	2674	528
Naïve Bayes	0.866	0.91	0.61	0.93	0.57	243	210	2599	326
KNN(Tuned)	0.978	0.98	0.94	0.99	0.92	45	31	2778	524
DT (Tuned)	0.918	0.94	0.8	0.97	0.68	180	96	2713	389
RF – Tuned	0.959	0.99	0.84	0.96	0.93	39	101	2708	530
Ensemble(Linear, Logit, LDA, DT, RF, NB, KNN)	0.954	0.96	0.94	0.99	0.78	127	27	2782	442
Ensemble(DT, RF, KNN)	0.972	0.98	0.93	0.99	0.9	59	36	2773	510
Ensemble (RF, KNN)	0.981	0.99	0.95	0.99	0.93	39	25	2784	530

Table 10 Models evaluation metrics - Test data

The ROC curve also was used for comparative analysis. The ROC curve for the tuned base models is shown below:

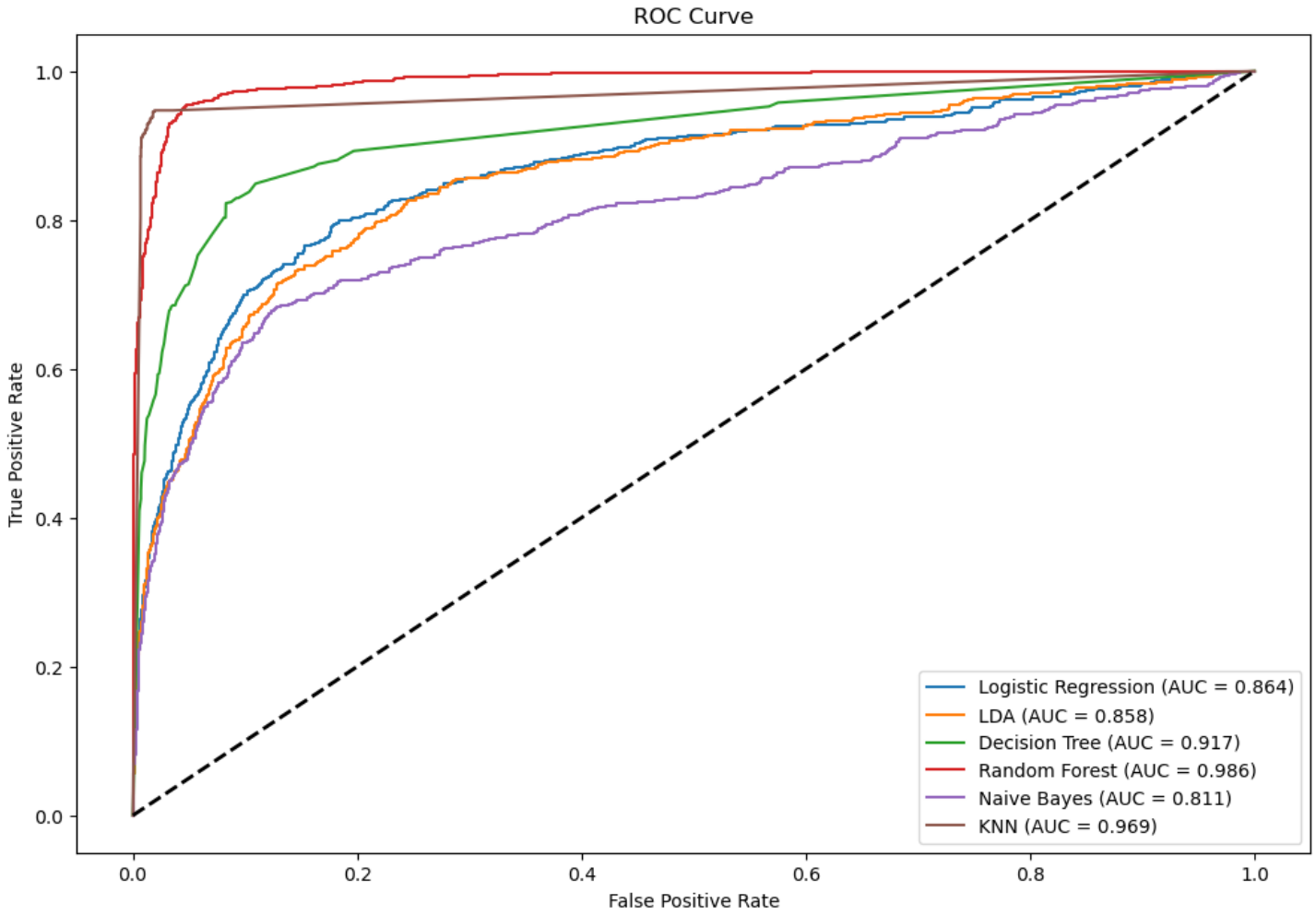

Figure 13 Models evaluation - ROC Curve

6. **Final interpretation / recommendation**

1. Feature Importance:

The best models, Tuned Random Forest, KNN and ensemble of these two models were then utilized to calculate respective feature importance that would help identify the features that are the most important for the models so that precise business actions can be taken. Following graphs show the feature importance:

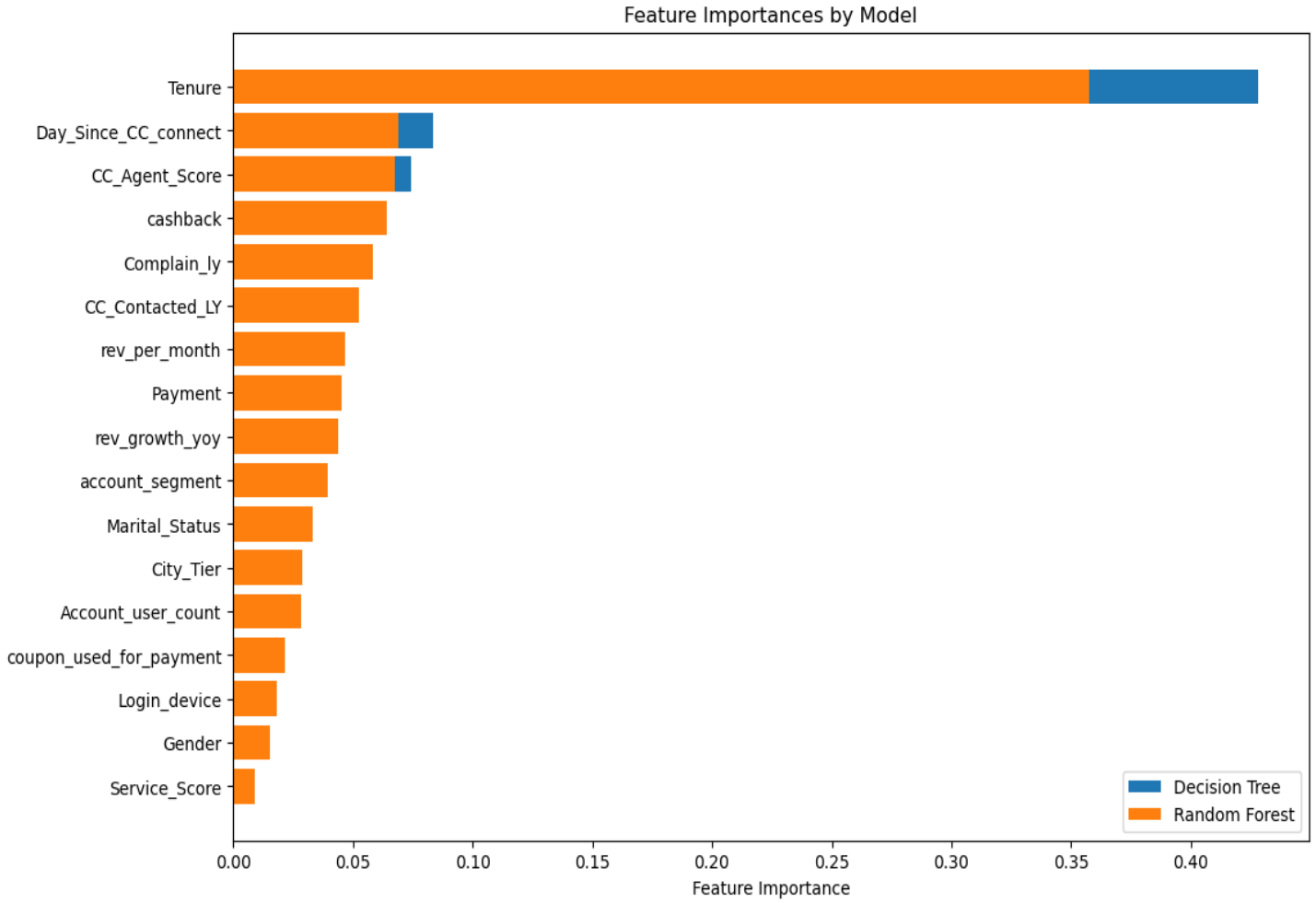

Figure 14 Feature importance Bar graph

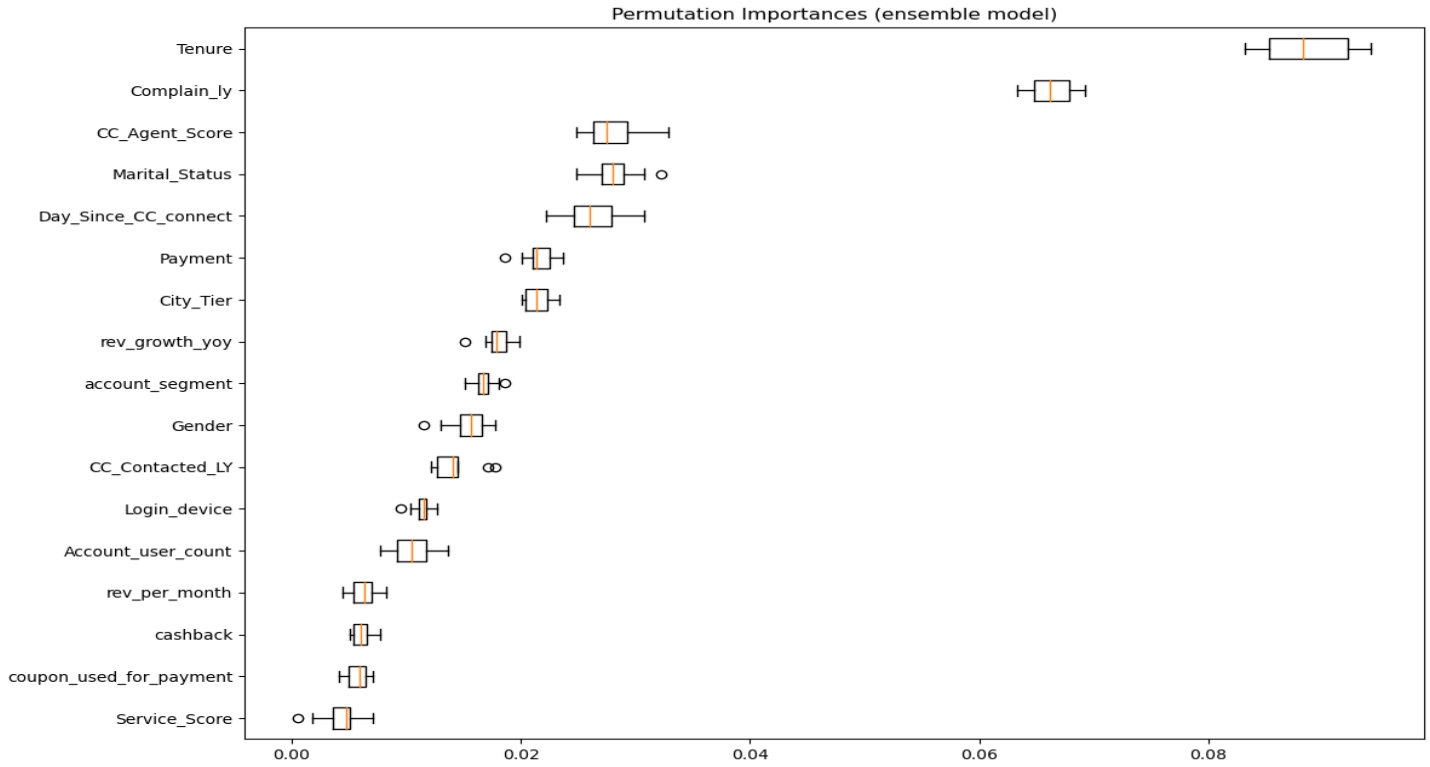

Figure 15 Permutation importance - ensemble model

- The most important attributes thus identified are:
  - Tenure
  - Days_Since_CC_Connect
  - CC_Agent_Score
  - Complain_ly
  - Marital Status
  - City_Tier
  - Cashback

2. Interpretation/ Recommendations:

Basic interpretations have been provided in the section on EDA. Based on the attributes identified by the best fit models and the EDA on those attributes, following are the recommendations:

- In the first 5 days since customer onboarding, the churn is the highest (~40%). Special campaign needs to be launched to understand and resolve the issues faced by customers. The campaign duration must be 1–2-weeks from onboarding.
- Cashback of INR.250 seems to be retaining the customers. Coupled with the campaign for new customers, it is safe to expect this cashback for new customers in the first month will retain them.
- Almost all the churners leave our platform within 10 days of interaction with the customer care agent.
    - 68% customers rated their interactions as 3-5.
    - Of all the churners, they constitute 78%.
    - This indicates customers who gave agent score of 3-5 are in high-risk category.
    - Therefore, the agent score must be carefully evaluated. Transcripts of the conversation between the customer and the agent must be monitored to understand if the concern of the customer was addressed and if there was any room for improvement.
    - If the concerns for those customers are not addressed, the issue must be immediately escalated and resolved on priority.
- Identify the customers who raised at least 1 complaint in the last 1 year. List all the issues for which the complaint was raised and resolve them. Inform the customers of the resolution and seek feedback.
- Launch campaigns targeted on a regular basis on customers who are single. Bucketing the products that are more popular among this segment may help. Also, launch such campaigns at regular intervals. Ensure that the Customer Care Agents are aware of the campaigns and can help customers.
- The company is not doing as well in tier 2 and tier 3 cities as they are in tier 1 cities. Focusing on customers belonging to such cities will help in customer retention.
- Combining all the above risk areas, the single customers belonging to tier 2 and tier 3 cities who had connected with a customer care agent and was not happy with the conversation belong to the highest risk category. Such customers must be given special attention to understand their concern and address them asap. Retention of such customers is expected to have a positive impact on our company's performance.