CHURN

ABSTRACT

The objective of this project is to predict customer churn in an E-Commerce company using machine learning techniques. Customer churn, or the rate at which customers discontinue their services, is a major challenge for the company. Predicting which customers are likely to churn can help the company take proactive steps to retain them.

Balasubramaniyam, R.

PG - Data Science and Business Analytics

1. Introduction

- The E-Commerce company is facing a lot of competition in the current market and facing difficulties with retaining the existing customers.
- The company wants to develop a model to do churn prediction of the accounts and provide segmented offers to the potential churners.
- Customer churn refers to the rate at which customers stop using a company's products or services over a particular period.
- The account churn could result in the loss of customers and potentially increase the cost of acquiring new customers.
- The task is to develop a churn prediction model for this company and provide optimal and cost-effective business recommendations for a successful campaign.

a) Need for Study

Customer churn is a critical metric for any business that relies on a recurring revenue model, including our e-commerce company.

It is essential to study customer churn for several reasons:

1. Cost of customer acquisition
2. Revenue impact
3. Customer loyalty
4. Competitive advantage

Studying customer churn can provide valuable insights that our business can use to improve customer retention, reduce costs, and gain a competitive advantage.

b) Social Opportunity

There are several social opportunities associated with customer churn study.

1. Improving customer satisfaction
2. Building trust
3. Reducing waste
4. Supporting local economies

Overall, the customer churn study has significant social opportunities that extend beyond the individual business and can contribute to building more positive and sustainable communities.

2. EDA and Business Implication

a) Understanding how data was collected in terms of time, frequency, and methodology

- The data is gathered by using the customer account as a unique identifier to distinguish between individual customers or groups of customers who use our service.
- The dataset includes customer accounts belonging to both new and long-standing customers, spanning a period of up to 31 months.
- The data is sourced from customers residing in Tier-1, Tier-2, and Tier-3 cities. Monthly data is collected on the average revenue generated by each account over the last 12 months, and the revenue growth percentage of the account over the last 12 months as compared to the period between 24 and 13 months ago is also recorded.

b) Visual inspection of data (rows, columns, descriptive details)

	Churn	City_Tier	CC_Contacted_LY	Service_Score	CC_Agent_Score	Complain_ly
count	11260.000000	11148.000000	11158.000000	11162.000000	11144.000000	10903.000000
mean	0.168384	1.653929	17.867091	2.902526	3.066493	0.285334
std	0.374223	0.915015	8.853269	0.725584	1.379772	0.451594
min	0.000000	1.000000	4.000000	0.000000	1.000000	0.000000
25%	0.000000	1.000000	11.000000	2.000000	2.000000	0.000000
50%	0.000000	1.000000	16.000000	3.000000	3.000000	0.000000
75%	0.000000	3.000000	23.000000	3.000000	4.000000	1.000000
max	1.000000	3.000000	132.000000	5.000000	5.000000	1.000000

2.1. Descriptive summary

The dataset contains 11260 rows and 18 columns. ‘Churn’ is the target variable and other features are the predictor variables.

c) Understanding of attributes (variable info, renaming if required)

Variable	Description
AccountID	account unique identifier
Churn	account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_L12m	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_l12m	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 month)
coupon_used_l12m	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_l12m	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

2.2. Data Dictionary

- The Churn column in our dataset serves as the target variable, indicating whether a customer has churned (1) or remained with the company (0).
- All the variables except Churn variable are predictor variables which can be used for building our predictive model.
- Since the column names do not contain any special characters, we can keep them as it is.
- Dataset comprises of the following data types: float64 (5), int64 (1), and object (12).

#	Column	Non-Null Count	Dtype
0	Churn	11260 non-null	int64
1	Tenure	11158 non-null	object
2	City_Tier	11148 non-null	float64
3	CC_Contacted_LY	11158 non-null	float64
4	Payment	11151 non-null	object
5	Gender	11152 non-null	object
6	Service_Score	11162 non-null	float64
7	Account_user_count	11148 non-null	object
8	account_segment	11163 non-null	object
9	CC_Agent_Score	11144 non-null	float64
10	Marital_Status	11048 non-null	object
11	rev_per_month	11158 non-null	object
12	Complain_ly	10903 non-null	float64
13	rev_growth_yoy	11260 non-null	object
14	coupon_used_for_payment	11260 non-null	object
15	Day_Since_CC_connect	10903 non-null	object
16	cashback	10789 non-null	object
17	Login_device	11039 non-null	object

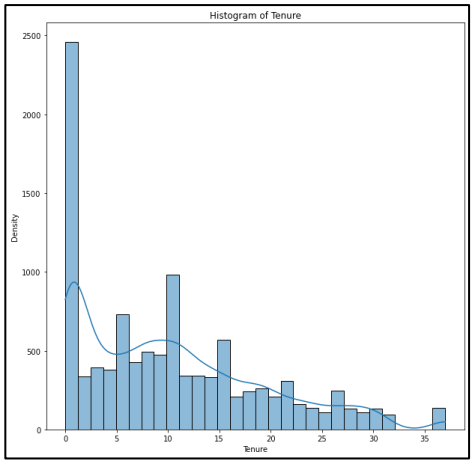
2.3. Data Information

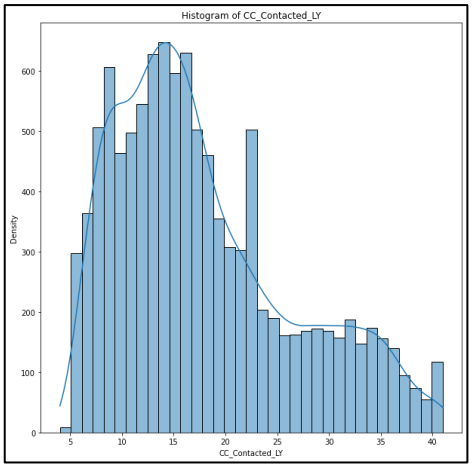
d) <u>Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)</u>

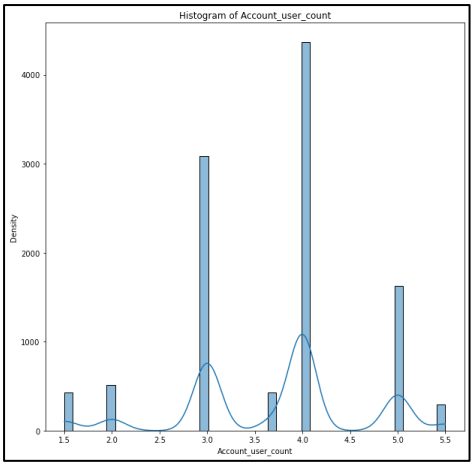
- Univariate analysis assists us in visually analysing the distribution of the data and assessing the significance of an individual feature.
- Let us understand the dataset with the help of following plots,
  1. Histogram
  2. Bar Chart
  3. Correlation Matrix

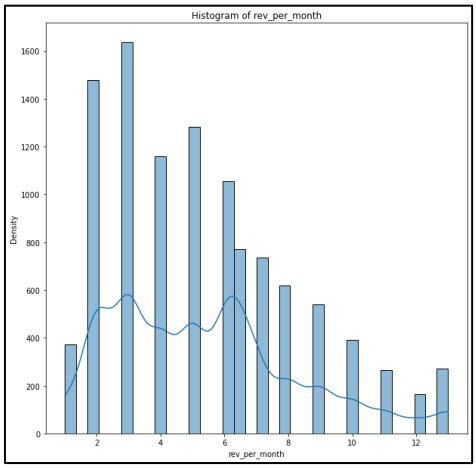
e) <u>Is the data unbalanced? If so, what can be done?</u>

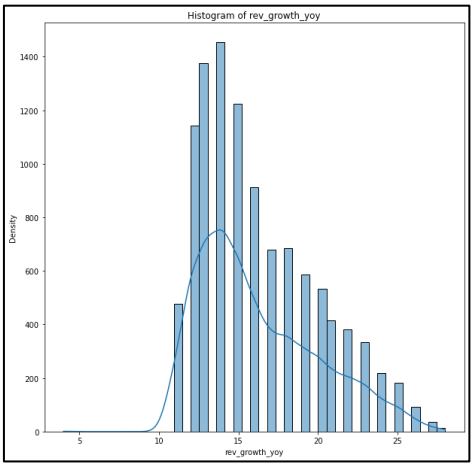
- The total number of churns is not equal to the number of customers who stayed with the company, the number of churned customers is much smaller than the number of non-churned customers because of this our data can be considered unbalanced.
- This unbalanced data can be corrected by using Oversampling or Under sampling of the target variable using algorithm such as Synthetic Minority Oversampling Technique (SMOTE).
- This unbalanced data can cause a bias in the predictive model and perform poorly on the minority class. This could result in a model which will poorly predict that a customer is loyal but, they could churn from the company.
- The imbalance in the churn status could be because the dataset contains relatively more new customers, who have not been with the company for a long period of time, and hence have not had the opportunity to churn yet.

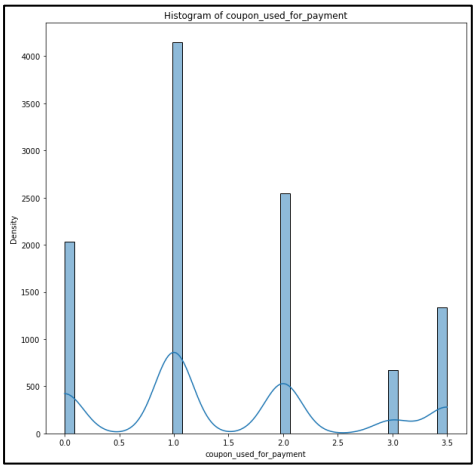


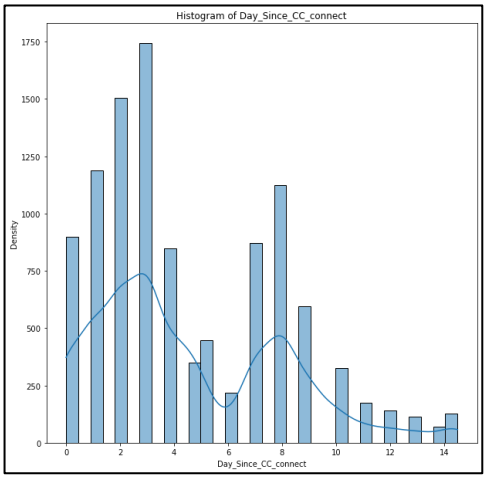


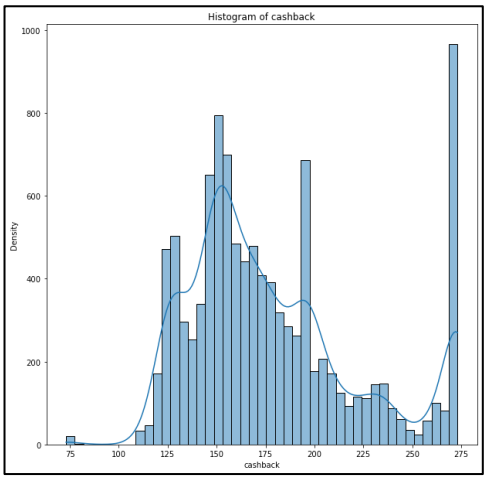






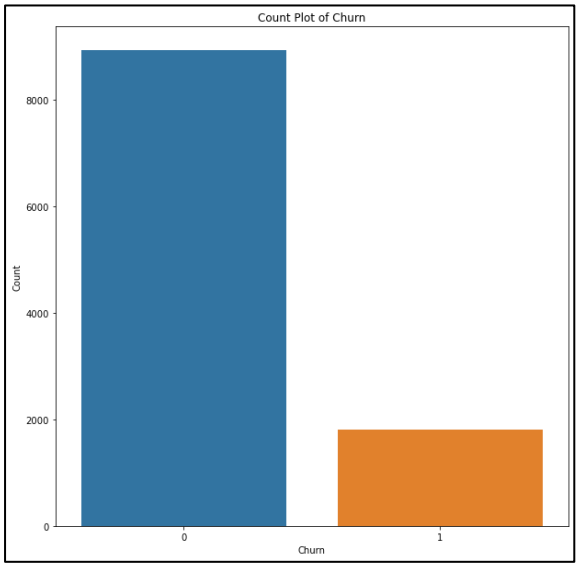


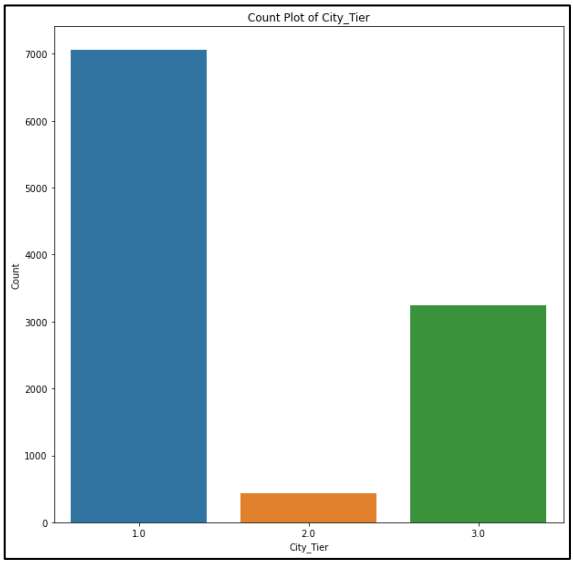




2.4. Histogram of Numerical Data

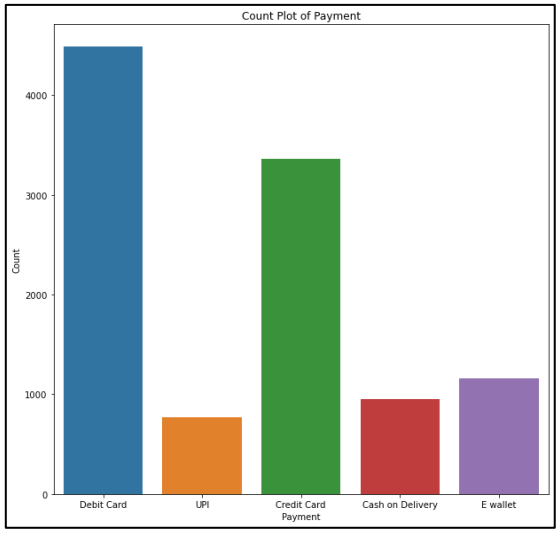
- The dataset's numerical features do not exhibit a normal distribution. Additionally, the company appears to have acquired a larger number of new customers than existing ones, as evidenced by the distribution being skewed towards new customers.
- Most of the customer accounts appear to be shared among 3 to 4 customers.

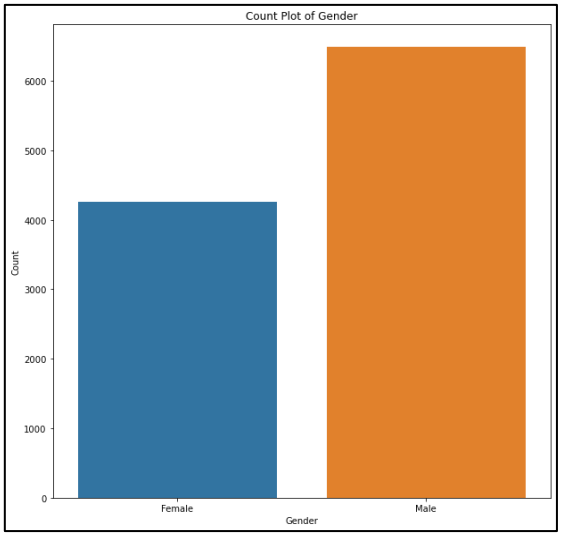




2.5. Bar chart of Churn and City Tier

- About (1852)16.8% of the customers have churned from our company and the remaining (9149) 83.2% customers have stayed.
- Many customers are from Tier-1 city followed by Tier-3 and Tier-2 cities.

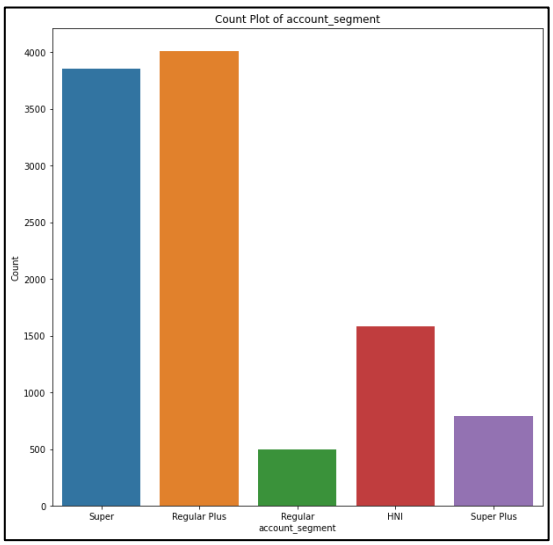


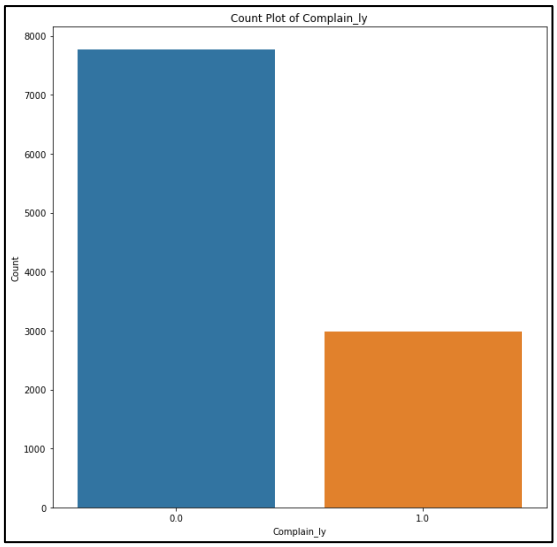


2.6. Bar chart of Payment mode and gender

- It appears that the most preferred mode of payment among customers is Debit card, followed by Credit card, E-wallet, Cash on delivery, and UPI.
- The payment mode information suggests that customers may prefer the convenience and security of using electronic payment methods over traditional cash-based transactions.
- It appears that most of our customers are male, accounting for 60% of the total customer base, while female customers account for 40%.

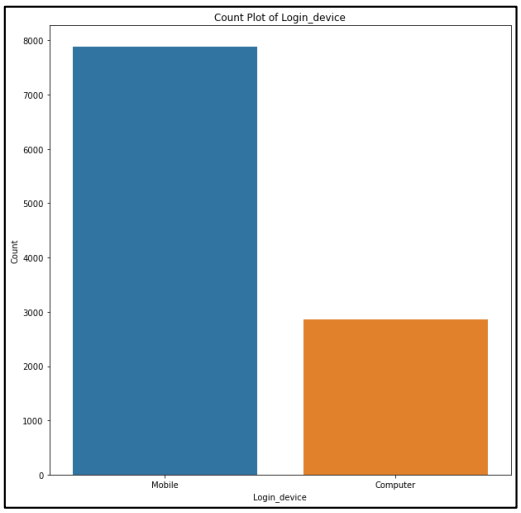
- The gender information suggests that our business may have a higher appeal to men than women or that your marketing efforts have been more successful in attracting male customers.

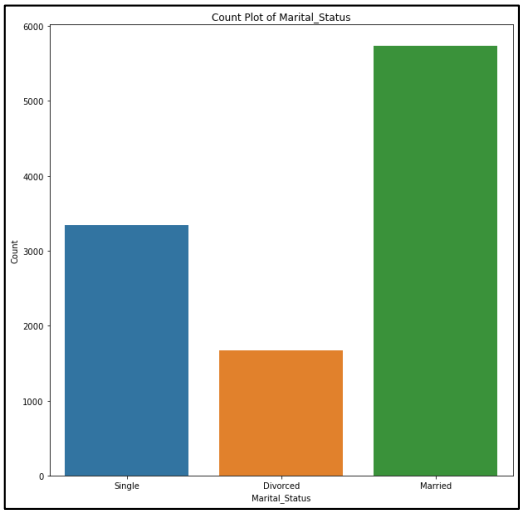




2.7. Bar chart of Account segment mode and yearly complaint

- It appears that our customer base is segmented into several categories, with most customers falling into the Regular Plus (37.3%) and Super (35.8%) segments, followed by HNI (14.7%) and Super plus (7.3%) segments, and the smallest segment being Regular (4.6%). These segments likely represent different levels of spending, engagement, and loyalty among your customer base.
- Additionally, it seems that a significant portion of your customer base has not made any complaints in the past 12 months, accounting for 72.3% of your customers.
- This suggests that your business may have a strong track record of delivering high-quality products or services and addressing any issues promptly.

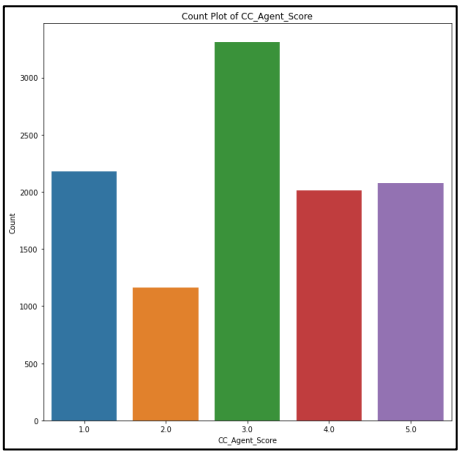


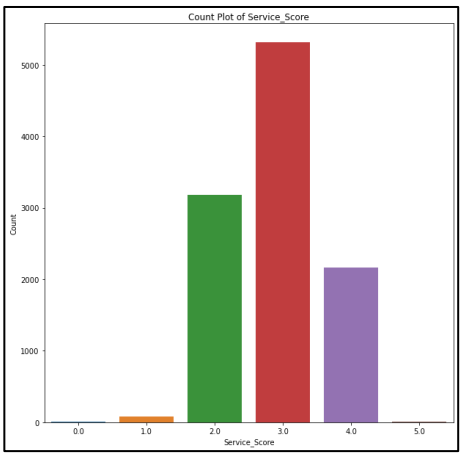


2.8. Bar chart of login device and marital status

- It appears that our customers prefer using mobile devices over computers, which could be due to the convenience and accessibility of using mobile devices on-the-go.

- Additionally, it seems that a significant portion of your customer base is married, followed by single customers, and few are divorced.

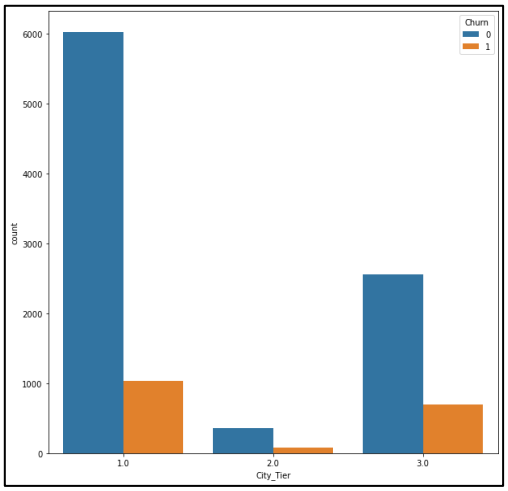


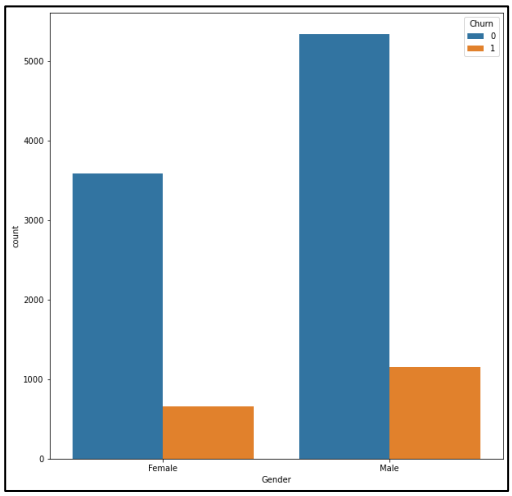


2.9. Bar chart of CC Agent score and service score

- It appears that the satisfaction score given by customers for the customer care service provided by your company is mostly around 3, accounting for 30.8% of the scores, and the least common score is 2, accounting for 10.8% of the scores.
- This information suggests that there may be room for improvement in the customer care service provided by your company to increase customer satisfaction.
- On the other hand, the satisfaction score given by customers for the service provided by your company is mostly around 3, accounting for 49.4% of the scores, while the least common score is 0, accounting for only 0.07% of the scores.
- This information suggests that the service provided by your company is generally satisfactory to customers, although there may still be areas for improvement.

f) Bivariate analysis (relationship between different variables, correlations)

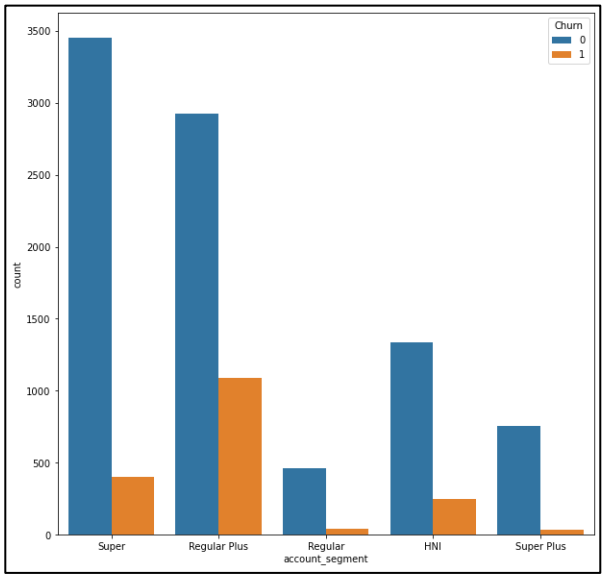




2.10. Churn Vs City, Gender

- It appears that many customers in tire-1, tire-2, and tire-3 cities are staying with the company, indicating a strong customer base in urban and semi-urban areas.

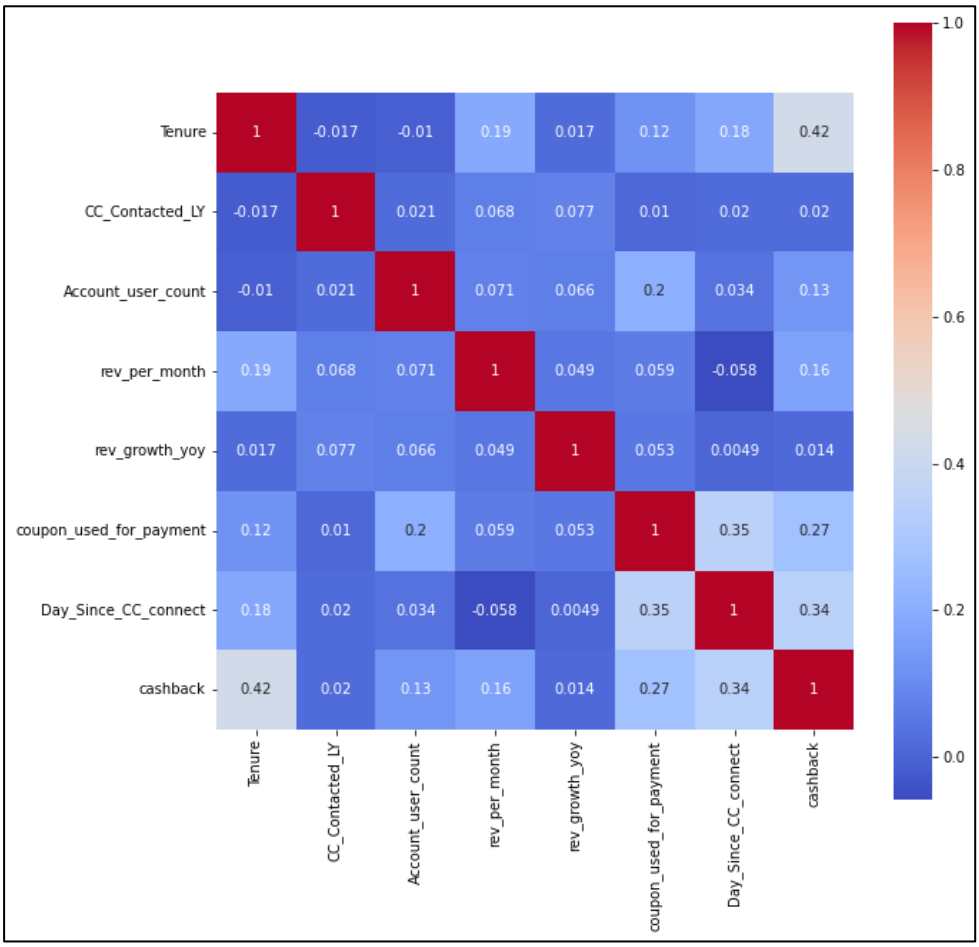
- This information suggests that your business may have a competitive advantage in these areas or that companies marketing and distribution strategies are effective in reaching customers in these regions.
- Additionally, while male customers may account for most of the customer base, it seems that they have churned at a rate equivalent to female customers. This suggests that factors beyond gender, such as product quality, customer service, or pricing, may be more critical drivers of customer churn.



2.11. Churn Vs customer segment

- It appears that customers in the Regular Plus segment are churning more than other segments. To address this issue, offering a trial run of the Super segment features to these customers could be an effective strategy to reduce their churn possibility.
- This approach could help to demonstrate the value of upgrading to a higher tier and may incentivize customers to stay with our company.

g) Multivariate analysis



2.12. Correlation matrix

- There is 42% positive correlation with tenure and cashback, 35% positive correlation with the coupon used for payment and the day since the customer has contacted the customer care.
- Tenure of the customer could be influenced by giving the customer with a cashback offer on recharges made using debit or credit card as they are the preferred payment mode.

3. Data Cleaning and Pre-processing

a) Removal of unwanted variables

- All the features of our dataset are kept for building the model. There were some special characters present in (Tenure, Account_user_count, rev_per_month, rev_growth_yoy, coupon_used_for_payment, Day_Since_CC_connect, cashback, Login_device) columns.
- These special characters are dropped and replaced with null values.
- The null values are later imputed with mean, and modes based on the data type.

b) Missing Value treatment

- Null values are present in the dataset and these null values are imputed with mean value for numerical variables and mode value for categorical variables.
- Our Churn Column doesn't have any missing values.

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	211
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	760

Tenure	0
CC_Contacted_LY	0
Account_user_count	0
rev_per_month	0
rev_growth_yoy	0
coupon_used_for_payment	0
Day_Since_CC_connect	0
cashback	0
Churn	0
City_Tier	0
Payment	0
Gender	0
Service_Score	0
account_segment	0
CC_Agent_Score	0
Marital_Status	0
Complain_ly	0
Login_device	0

3.1. Before and after missing value treatment

c) Outlier treatment

- An outlier is an observation of data that does not fit the rest of the data. It is sometimes called an extreme value.
- Such Outliers are present in the Tenure, CC_Contacted_LY, Account_user_count, rev_per_month, coupon_used_for_payment, Day_Since_CC_connect, cashback columns of our dataset.
- The outliers are imputed with upper limit values for the extreme values exceeding the upper limit, the lower limit value is used for extreme values less than the lower limit.

d) Variable transformation

- The data type of the Churn, City_Tier, Service_Score, CC_Agent_Score, Complain_ly columns are changed to object datatype as they are categorical in nature.
- Some of the features can be changed to integer or category datatype and encoded based on the model selected for prediction.

e) Addition of new variables

- No new variables were added. Instead, all existing variables except for the account ID are considered for model building.
- New variables might be added in the model building process if the current features are found not enough for building an optimum prediction model.

4. Model building

- Our customer churn prediction is a ‘classification problem’ in machine learning.
- The output variable here is binary, with the two classes being ‘churned-1’ and ‘not churned-0.’
- A ‘supervised learning’ approach is used, the model is trained on this labeled dataset containing information on whether customers churned.

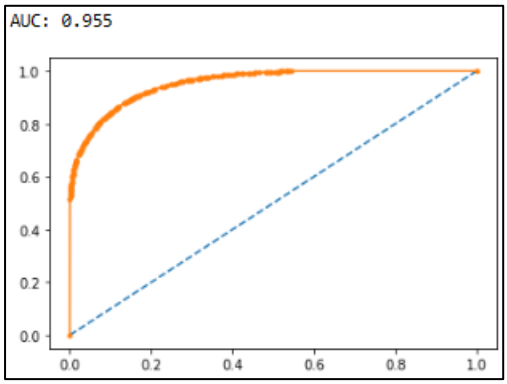
Algorithms used:

- Logistic regression,
- Decision trees,
- Random Forests,
- Ada Boosting,
- Gradient Boosting
- K-Nearest Neighbors.

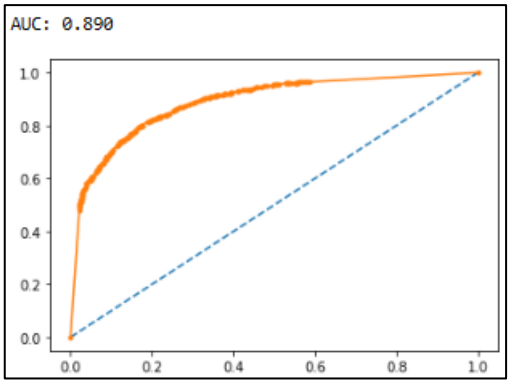
4.1.Model building and interpretation.

I. Decision Tree Model

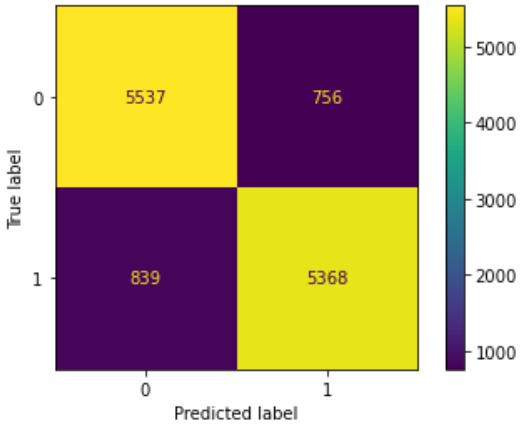
1.1. AUC and ROC for the training data



AUC and ROC for the test data



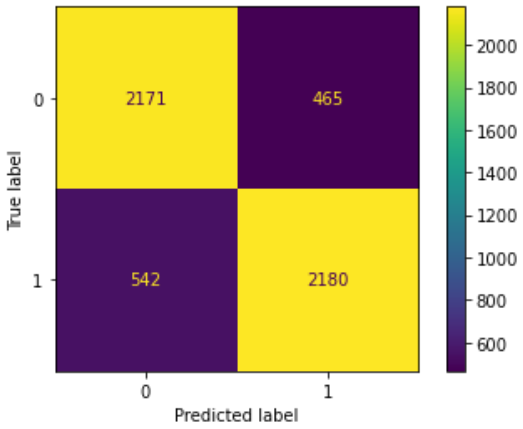
1.2. Confusion Matrix for the training data



Classification report for the train data

	precision	recall	f1-score	support
0	0.87	0.88	0.87	6293
1	0.88	0.86	0.87	6207
accuracy			0.87	12500
macro avg	0.87	0.87	0.87	12500
weighted avg	0.87	0.87	0.87	12500

1.3. Confusion Matrix for the test data



Classification report for the test data

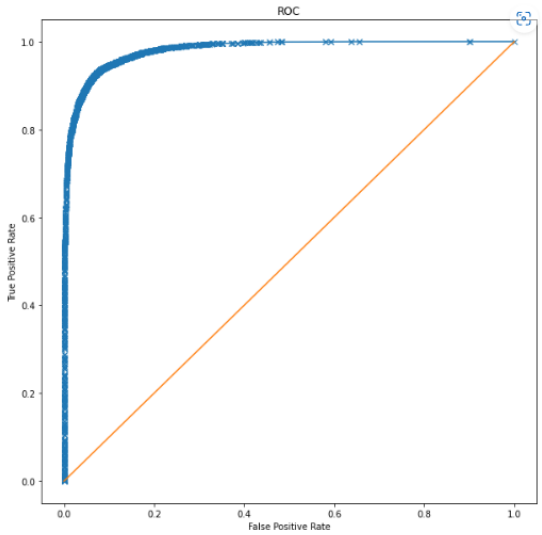
	precision	recall	f1-score	support
0	0.80	0.82	0.81	2636
1	0.82	0.80	0.81	2722
accuracy			0.81	5358
macro avg	0.81	0.81	0.81	5358
weighted avg	0.81	0.81	0.81	5358

- The decision tree model is not overfitted/underfitted.
- From Confusion matrix, the model is good at identifying the churn customers and non-churn customers.
- The recall is 80% for test data and 86% for train data. The model can identify 80% of churns correctly.
- The precision is 82% for test data and 88% for train data. The model can identify 82% of non-churns correctly.

II. Random Forest model

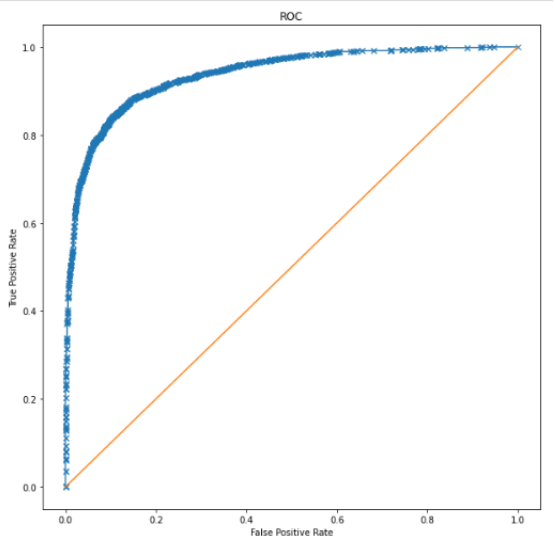
- Random Forest is a supervised machine learning algorithm made up of decision trees. Random Forest is used for both classification and regression—for example, classifying whether a customer is “Churn” or “not Churn.”

2.1. AUC and ROC for the training data



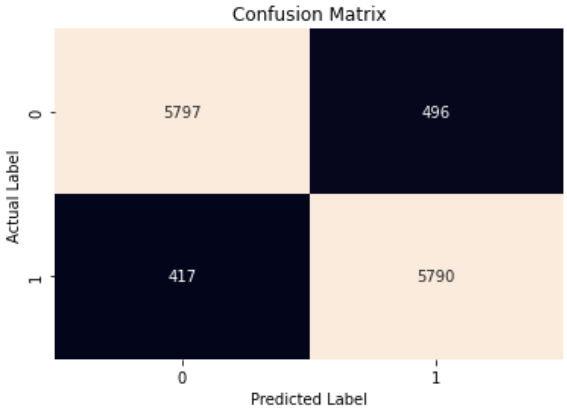
Area under Curve is 0.9809780692083191

AUC and ROC for the test data



Area under Curve is 0.9375711479218952

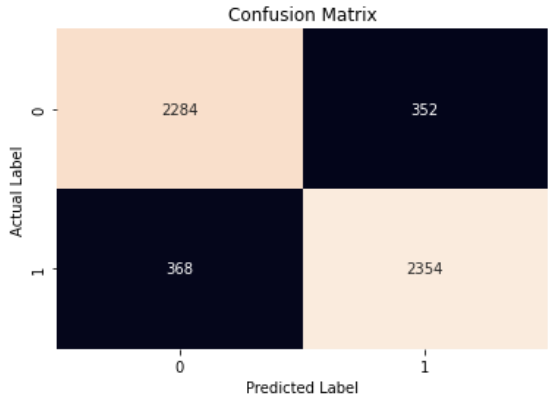
2.2. Confusion Matrix for the training data



Classification report for the train data

	precision	recall	f1-score	support
0	0.93	0.92	0.93	6293
1	0.92	0.93	0.93	6207
accuracy			0.93	12500
macro avg	0.93	0.93	0.93	12500
weighted avg	0.93	0.93	0.93	12500

2.3. Confusion Matrix for the test data



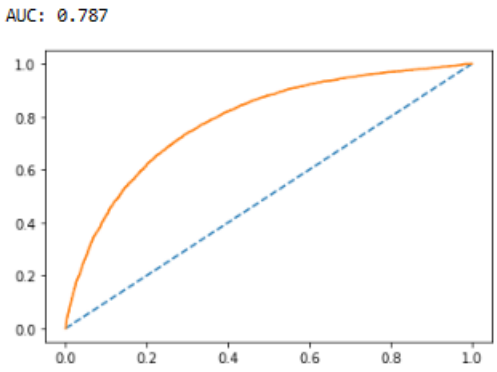
Classification report for the test data

	precision	recall	f1-score	support
0	0.86	0.87	0.86	2636
1	0.87	0.86	0.87	2722
accuracy			0.87	5358
macro avg	0.87	0.87	0.87	5358
weighted avg	0.87	0.87	0.87	5358

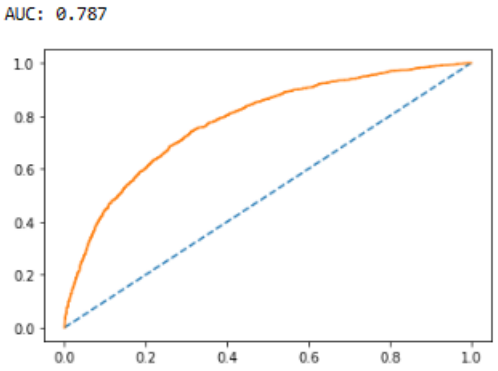
- The Random Forest model is not overfitted/underfitted.
- From Confusion matrix, the model is good at identifying the churn customers and non-churn customers.
- The recall is 86% for test data and 93% for train data. The model can identify 86% of churns correctly.
- The precision is 87% for test data and 92% for train data. The model can identify 87% of non-churns correctly.

III. Logistic Regression model

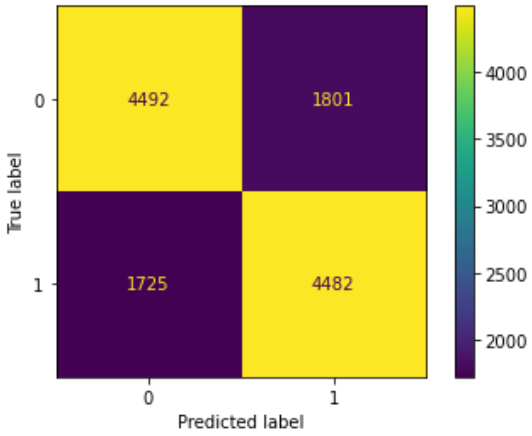
3.1. AUC and ROC for the training data



AUC and ROC for the test data



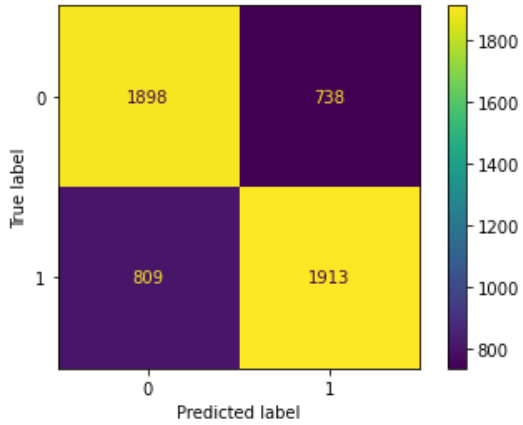
3.2. Confusion Matrix for the training data



Classification report for the train data

	precision	recall	f1-score	support
0	0.72	0.71	0.72	6293
1	0.71	0.72	0.72	6207
accuracy			0.72	12500
macro avg	0.72	0.72	0.72	12500
weighted avg	0.72	0.72	0.72	12500

3.3. Confusion Matrix for the test data



Classification report for the test data

	precision	recall	f1-score	support
0	0.70	0.72	0.71	2636
1	0.72	0.70	0.71	2722
accuracy			0.71	5358
macro avg	0.71	0.71	0.71	5358
weighted avg	0.71	0.71	0.71	5358

- The Logistic regression model is not overfitted/underfitted.
- From Confusion matrix, the model is okay at identifying the churn customers and non-churn customers.
- The recall is 70% for test data and 72% for train data. The model can identify 70% of churns correctly. The precision is 72% for test data and 71% for train data. The model can only identify 72% of non-churns correctly.

Model	AUC		Recall		Precision		F1 Score		Accuracy	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
CART	0.955	0.89	0.86	0.8	0.88	0.82	0.87	0.81	0.87	0.81
Random Forest	0.98	0.94	0.93	0.86	0.92	0.87	0.93	0.87	0.93	0.87
Logistic regression	0.787	0.787	0.72	0.7	0.71	0.72	0.72	0.71	0.72	0.71
Ada Boosting	-	-	-	-	-	-	-	-	-	-
Gradient Boosting	-	-	-	-	-	-	-	-	-	-
KNN	0.97	0.92	0.97	0.94	0.84	0.77	0.9	0.84	0.89	0.82

3.4. Model Comparison

Interpretation of models:

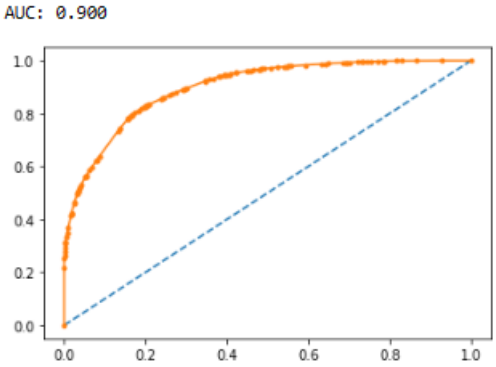
- There is no underfitting or overfitting in any of the above models.
- For this churn classification problem, recall is more important than precision, because the cost of false negatives (predicting that a customer will not churn when they actually do) is usually higher than the cost of false positives (predicting that a customer will churn when they actually do not).
- If the model has high precision but low recall, it means that the model is correctly identifying a small percentage of customers who are at risk of churning, but it is missing many customers who are actually churning. This will result in a significant loss of revenue and customer satisfaction.
- For this churn classification problem, recall is more important than precision.
- The 'Recall' value of KNN is better than Random Forest, CART, Logistic regression models.
- The 'Precision' of Random Forest is better than CART, KNN, Logistic regression models.
- Even though the KNN model is having good recall than other models, the other performance metrics like Precision, F1-Score and accuracy are optimal for the Random Forest model.
- After Random Forest model, CART is the second-best model for churn prediction.
- Logistic Regression is having least performance among the models with minimal recall, precision, accuracy.

4.2.Model Tuning

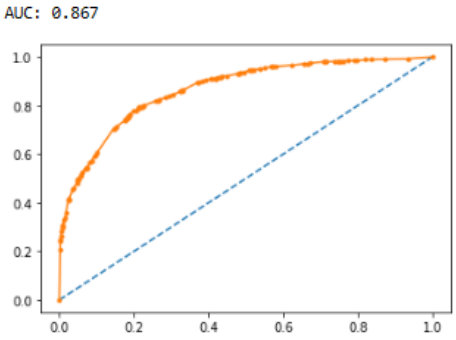
- GridSearchCV is a method in machine learning used for hyperparameter tuning, which involves selecting the best combination of hyperparameters for a particular model.
- The process involves defining a range of values for each hyperparameter and exhaustively searching all possible combinations of these hyperparameters to find the combination that gives the best performance.

I. Tuned Decision Tree Model

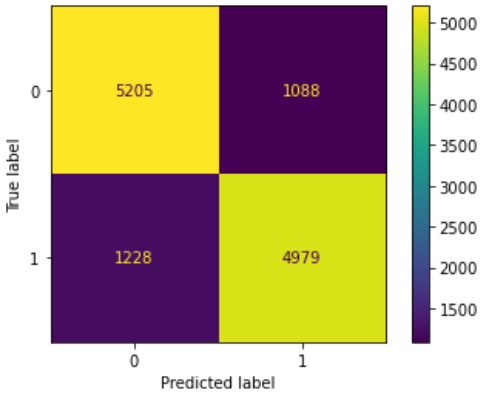
1.1. AUC and ROC for the training data



AUC and ROC for the test data



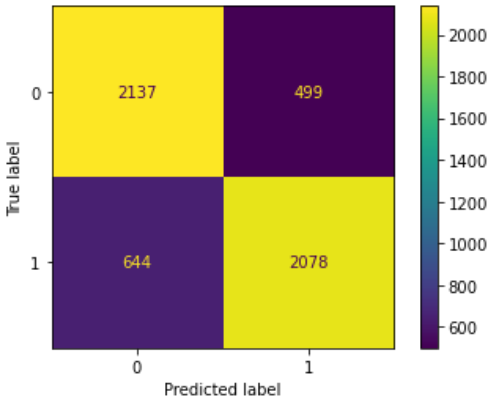
1.2. Confusion Matrix for the training data



Classification report for training data

	precision	recall	f1-score	support
0	0.81	0.83	0.82	6293
1	0.82	0.80	0.81	6207
accuracy			0.81	12500
macro avg	0.81	0.81	0.81	12500
weighted avg	0.81	0.81	0.81	12500

1.3. Confusion Matrix for the test data



Classification report for test data

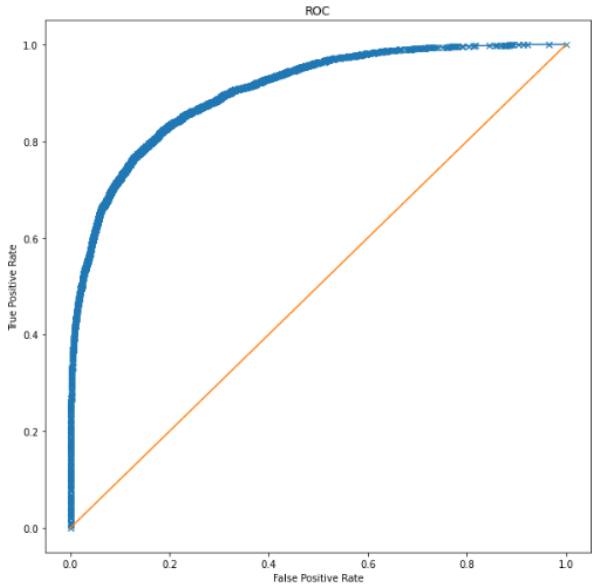
	precision	recall	f1-score	support
0	0.77	0.81	0.79	2636
1	0.81	0.76	0.78	2722
accuracy			0.79	5358
macro avg	0.79	0.79	0.79	5358
weighted avg	0.79	0.79	0.79	5358

- The decision tree model is not overfitted/underfitted.
- From Confusion matrix, the model is good at identifying the churn customers and non-churn customers.
- The recall is 76% for test data and 80% for train data. The model can identify 76% of churns correctly.

- The precision is 81% for test data and 82% for train data. The model can identify 81% of non-churns correctly.

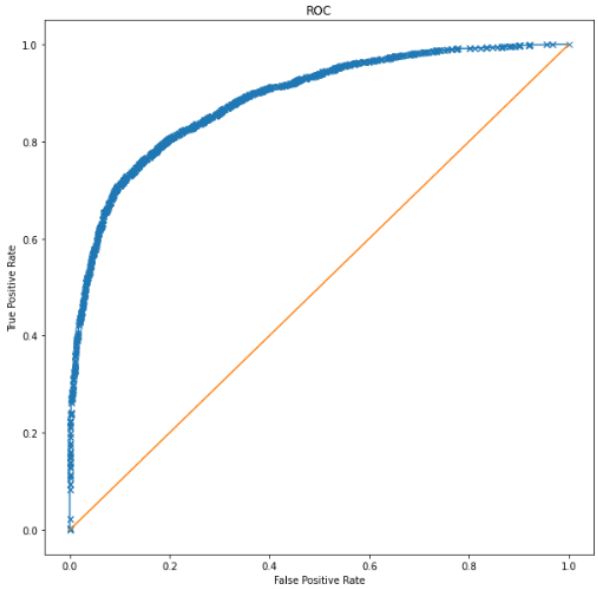
II. Tuned Random Forest Model

2.1. AUC and ROC for the training data



Area under Curve is 0.9036287183232045

AUC and ROC for the test data



Area under Curve is 0.8859204046386493

2.2. Confusion Matrix for the training data

```
([[5164, 1129],
  [1151, 5056]],
```

Classification report for training data

	precision	recall	f1-score	support
0	0.82	0.82	0.82	6293
1	0.82	0.81	0.82	6207
accuracy			0.82	12500
macro avg	0.82	0.82	0.82	12500
weighted avg	0.82	0.82	0.82	12500

2.3. Confusion Matrix for the test data

```
([[2146,  490],
  [ 563, 2159]],
```

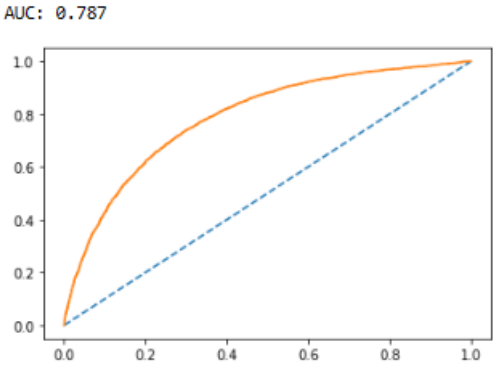
Classification report for test data

	precision	recall	f1-score	support
0	0.79	0.81	0.80	2636
1	0.82	0.79	0.80	2722
accuracy			0.80	5358
macro avg	0.80	0.80	0.80	5358
weighted avg	0.80	0.80	0.80	5358

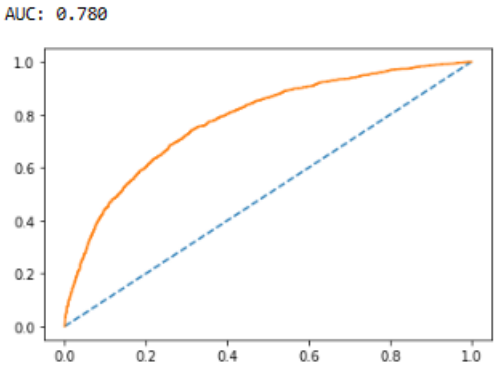
- The Random Forest model is not overfitted/underfitted.
- From Confusion matrix, the model is good at identifying the churn customers and non-churn customers.
- The recall is 79% for test data and 81% for train data. The model can identify 79% of churns correctly.
- The precision is 82% for test data and 82% for train data. The model can identify 82% of non-churns correctly.

III. Tuned Logistic regression Model.

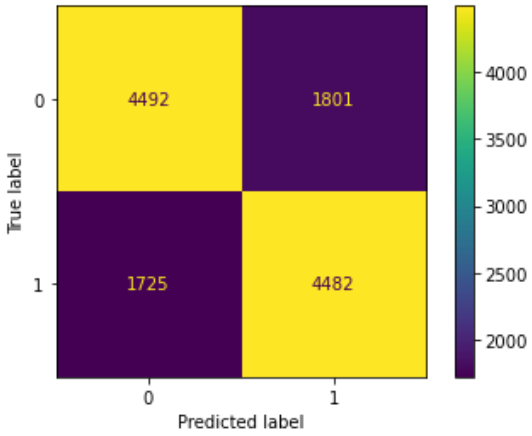
3.1. AUC and ROC for the training data



AUC and ROC for the test data



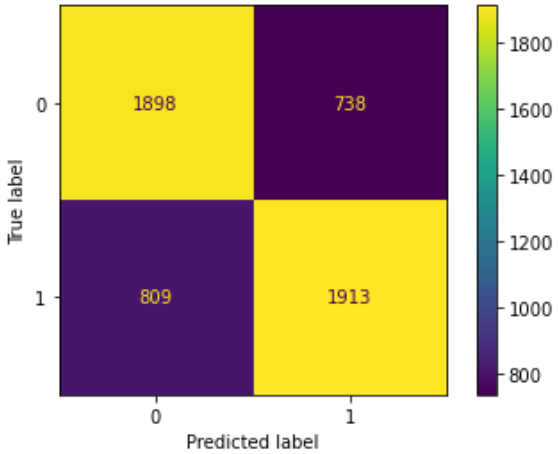
3.2. Confusion Matrix for the training data



Classification report for training data

	precision	recall	f1-score	support
0	0.72	0.71	0.72	6293
1	0.71	0.72	0.72	6207
accuracy			0.72	12500
macro avg	0.72	0.72	0.72	12500
weighted avg	0.72	0.72	0.72	12500

3.3. Confusion Matrix for the test data



Classification report for test data

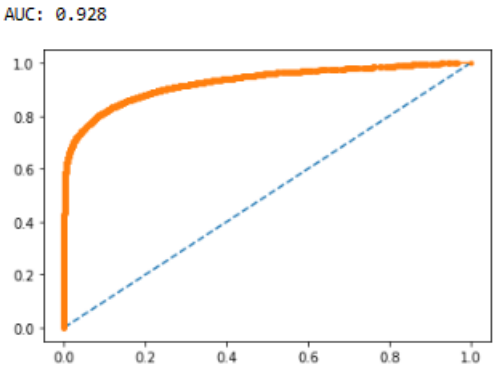
	precision	recall	f1-score	support
0	0.70	0.72	0.71	2636
1	0.72	0.70	0.71	2722
accuracy			0.71	5358
macro avg	0.71	0.71	0.71	5358
weighted avg	0.71	0.71	0.71	5358

- The Logistic regression model is not overfitted/underfitted.
- From Confusion matrix, the model is okay at identifying the churn customers and non-churn customers.
- The recall is 70% for test data and 72% for train data. The model can identify 70% of churns correctly. The precision is 72% for test data and 71% for train data. The model can only identify 72% of non-churns correctly. No improvement in model performance.

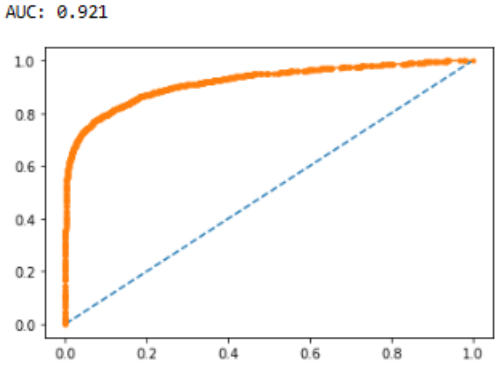
4.3.Ensemble models

Ada Boosting

4.3.1. AUC and ROC for the training data



AUC and ROC for the test data



4.3.2. Confusion Matrix for the training data, Classification report for training data

```
[[5564  729]
 [1058 5149]]
              precision    recall  f1-score   support

           0       0.84      0.88      0.86      6293
           1       0.88      0.83      0.85      6207

    accuracy                           0.86     12500
   macro avg       0.86      0.86      0.86     12500
weighted avg       0.86      0.86      0.86     12500
```

4.3.3. Confusion Matrix for the test data, Classification report for test data

```
[[2321  315]
 [ 510 2212]]
              precision    recall  f1-score   support

           0       0.82      0.88      0.85      2636
           1       0.88      0.81      0.84      2722

    accuracy                           0.85      5358
   macro avg       0.85      0.85      0.85      5358
weighted avg       0.85      0.85      0.85      5358
```

Tuned Model	AUC		Recall		Precision		F1 Score		Accuracy	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
CART	0.9	0.87	0.8	0.76	0.82	0.81	0.81	0.78	0.81	0.79
Random Forest	0.9	0.88	0.81	0.79	0.82	0.82	0.82	0.8	0.82	0.8
Logistic regression	0.787	0.78	0.72	0.7	0.71	0.72	0.72	0.71	0.72	0.71
Ada Boosting	0.928	0.921	0.83	0.81	0.88	0.88	0.85	0.84	0.86	0.85
Gradient Boosting	0.948	0.94	0.85	0.83	0.91	0.91	0.88	0.87	0.88	0.87
KNN	1	0.81	1	0.88	1	0.65	1	0.75	1	0.7

4.3.4. Tuned model Comparison.

- After applying GridSearch CV to optimize the parameters of our prediction models, we have noticed that the Gradient Boosting model (with a test recall of 83% and a test precision of 91%) outperforms the other tuned models. KNN model is overfitted in this case.

Interpretation of the most tuned model and its implication on the business:

- The most optimum model is the Gradient Boosting model, it is a powerful machine learning algorithm that can be used for our customer churn prediction. Churn prediction is the task of identifying customers who are likely to cancel or stop using a service or product.
- Gradient Boosting works by iteratively adding the weak learners to the model, each one focused on reducing the errors of the previous models. The final model is a combination of all these weak learners, and it can accurately predict whether a customer is likely to churn or not.
- There is **Overfitting** in the **KNN** model, and all other models don't show overfitted or an underfit nature.
- The Recall, Precision, Accuracy of the **Gradient Boosting model** is better than the **Ada Boosting, Random Forest, CART, Logistic regression models.**
- After tuning Logistic regression has the least performance in comparison with other models.

5. Model validation

Model	AUC		Recall		Precision		F1 Score		Accuracy	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
CART	0.955	0.89	0.86	0.8	0.88	0.82	0.87	0.81	0.87	0.81
Random Forest	0.98	0.94	0.93	0.86	0.92	0.87	0.93	0.87	0.93	0.87
Logistic regression	0.787	0.787	0.72	0.7	0.71	0.72	0.72	0.71	0.72	0.71
Ada Boosting	-	-	-	-	-	-	-	-	-	-
Gradient Boosting	-	-	-	-	-	-	-	-	-	-
KNN	0.97	0.92	0.97	0.94	0.84	0.77	0.9	0.84	0.89	0.82

Tuned Model	AUC		Recall		Precision		F1 Score		Accuracy	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
CART	0.9	0.87	0.8	0.76	0.82	0.81	0.81	0.78	0.81	0.79
Random Forest	0.9	0.88	0.81	0.79	0.82	0.82	0.82	0.8	0.82	0.8
Logistic regression	0.787	0.78	0.72	0.7	0.71	0.72	0.72	0.71	0.72	0.71
Ada Boosting	0.928	0.921	0.83	0.81	0.88	0.88	0.85	0.84	0.86	0.85
Gradient Boosting	0.948	0.94	0.85	0.83	0.91	0.91	0.88	0.87	0.88	0.87
KNN	1	0.81	1	0.88	1	0.65	1	0.75	1	0.7

5.1.1. Model Comparison (Normal and Tuned models)

- The '**Untuned Random Forest**' model has the best performance metrics for properly **identifying both the churned and non-churn** customers.
- The Untuned Random Forest model is selected for churn prediction because of the following best output,
    1. Recall-86%,
    2. Precision-87%
    3. Accuracy-87% in predicting the test dataset.

- For this churn classification problem, recall is more important than precision, because the cost of false negatives (predicting that a customer will not churn when they actually do) is usually higher than the cost of false positives (predicting that a customer will churn when they actually do not).
- If the model has high precision but low recall, it means that the model is correctly identifying a small percentage of customers who are at risk of churning, but it is missing many customers who are churning. This will result in a significant loss of revenue and customer satisfaction.
- The overall predictive performance of this Random Forest model with respect to RECALL, PRECISION, ACCURACY is better than all the other models.

Variable Importance of Random Forest model

Variables	Importance (%)
Marital_Status	21%
account_segment	16%
Account_user_count	7%
Payment	7%
coupon_used_for_payment	6%
Login_device	5%
Tenure	5%
CC_Contacted_LY	4%
cashback	4%
rev_per_month	4%
rev_growth_yoy	4%
Complain_ly	3%
Day_Since_CC_connect	3%
City_Tier	3%
CC_Agent_Score	3%
Service_Score	2%
Gender	2%

- Variable importance refers to the degree to which these input variables contribute to the output or "Churn" variable in our machine learning model.
- A high feature importance score indicates that a variable has a strong positive relationship with the Churn, while a low feature importance score indicates a negative relationship.

Top 5 Important variables,

1. Marital Status
2. Account segment
3. Account user count
4. Payment and Coupon used for Payment.

5.1.2. Variable Importance

6. Final interpretation / recommendation

Insights

- 27% of Single customers have churned from the company followed by (14.6%) of Divorced customers, Married customers have the lowest churn rate (11.5% have churned). 53.3% of our customers are married.
- 27% of Regular plus account segment have churned from the company, this segment has the maximum customer churn in the company. Super plus segment has the minimum percentage of churn in the company (4.6%).

- Accounts with user counts of (3 to 5 customers) have higher churn percentage (17.6% in this category), 12% of customers in the 6 user accounts category have churned from the company.
- Customers who used (8% of total customers) Cash on Delivery mode of payment have higher churn rate (25% of COD used customers), Customers who used (31.3% of total customers) Credit card mode of payment have the lowest churn rate (14% of Credit card used customers).
- Customers who have used the coupon (15 times) for making the payment in the last 12 months have higher churn rate (25% of the customers of this cluster have churned).
- 74% of the customers are using the mobile device, 26% of customers are using computers. 20% of computer users have churned from the company whereas only 15.7% of mobile users have churned from using our service.
- Male Customers are having higher churn rate at (17% of total male customers)

Recommendations

- Focus is need in retaining the single and divorced customers: Since single and divorced customers have a higher churn rate than married customers, provide promotional offers to these customers.
- Consider offering incentives to married customers: While married customers have the lowest churn rate, it's still important to retain these customers since they make up the majority of the customer base.
- Consider adjusting pricing or services for Regular plus account segment: If pricing or service offerings are contributing to churn in the Regular plus account segment, it may be necessary to adjust these factors to better meet the needs of these customers. This could involve offering more competitive pricing or adding new services that are tailored to this segment's needs.
- Encourage customers to switch to credit card mode of payment: Since customers who used credit card mode of payment have the lowest churn rate, it may be worthwhile to incentivize customers to switch to this mode of payment.
- This could involve offering exclusive discounts or rewards for customers who use credit card mode of payment or providing education on the benefits of using credit card mode of payment, such as improved security and faster processing times.
- COD payment modes could involve analyzing of customer feedback and conducting surveys to gain insights into customer satisfaction and preferences.
- Provide initial discounts or bundle packages to computer using customers and encourage them to switch to mobile devices. The mobile device users tend to stay longer with the company.