CAPSTONE PROJECT- EDTH CUSTOMER CHURN

DTH Company dataset analysis to develop a churn prediction model for this company and provide business recommendations on the campaign

Hithesh Devadiga

CONTENT

Introduction

Problem Statement

In the current market, DTH providers face intense rivalry, making it difficult to hang on to their current clientele. As a result, the business is trying to create a model that will allow them to predict account churn and send targeted offers to those who might consider cancelling. Because a single account can have several customers, account churn is a big concern for this organisation. Therefore, the organisation may lose more than one customer by losing one account.For this corporation, you have been tasked with creating a churn prediction model and offering business suggestions for the campaign.Because your suggestion will be reviewed by the campaign's decision-makers, it should be distinct and explicit about the campaign offer.

Importance of analysing the dataset:

The following machine learning initiatives will help us reduce the churn rate. We can increase the number of customers by decreasing the churn rate.

Potential clients have the ability to build a robust community for that specific e-commerce website.

Also, the business uses internet and social media to only engage potential customers that machine learning has identified.

Exploring the dataset

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

EDA and Business Implication

A Glance on dataset

AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	I
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	

Size of the dataset:

The dataset is of the shape (1525 *19) ,which means there are 1525 rows and 19 columns.

A Check on Duplication:

There is no duplication of entries in the given dataset. Each entry is a unique record of customer_id.

A Check on balance of the dataset:

```
0    9364
1    1896
Name: Churn, dtype: int64
```

The dataset consists of 9.3K entries relating to customers who have not churned and 1.8K entries relating to churned customers. The ratio between the unchurn and churn being 83:17. Hence we can conclude that the dataset is unbalanced.

Checking the null Values and datatypes in dataset:

```
AccountID                  0
Churn                      0
Tenure                   102
City_Tier                112
CC_Contacted_LY          102
Payment                  109
Gender                   108
Service_Score             98
Account_user_count       112
account_segment           97
CC_Agent_Score           116
Marital_Status           212
rev_per_month            102
Complain_ly              357
rev_growth_yoy             0
coupon_used_for_payment    0
Day_Since_CC_connect     357
cashback                 471
Login_device             221
dtype: int64
```

```
 #   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
```

By the analysis we could see that 14/18 columns have null values in this dataset.

And the datatypes of the datasets being 2/18 are int, 12/18 are object and 5/18 are float

Summary of the dataset

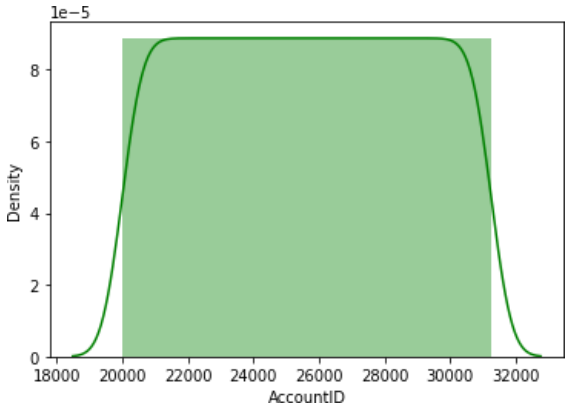
	count	mean	std	min	25%	50%	75%	max
AccountID	11260.0	25629.500000	3250.626350	20000.0	22814.75	25629.5	28444.25	31259.0
Churn	11260.0	0.168384	0.374223	0.0	0.00	0.0	0.00	1.0
City_Tier	11148.0	1.653929	0.915015	1.0	1.00	1.0	3.00	3.0
CC_Contacted_LY	11158.0	17.867091	8.853269	4.0	11.00	16.0	23.00	132.0
Service_Score	11162.0	2.902526	0.725584	0.0	2.00	3.0	3.00	5.0
CC_Agent_Score	11144.0	3.066493	1.379772	1.0	2.00	3.0	4.00	5.0
Complain_ly	10903.0	0.285334	0.451594	0.0	0.00	0.0	1.00	1.0

Summary of categorical variable

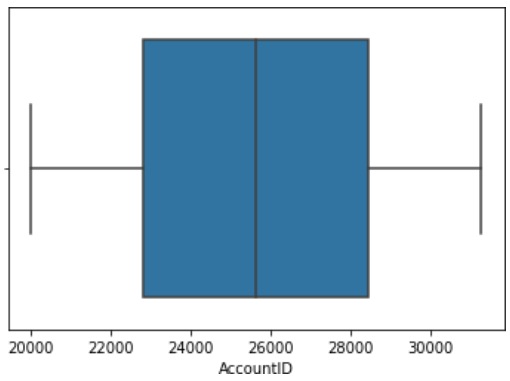
	Tenure	Payment	Gender	Account_user_count	account_segment	Marital_Status	rev_per_month	rev_growth_yoy	coupon_used_for_payment
count	11158	11151	11152	11148	11163	11048	11158	11260	11260
unique	38	5	4	7	7	3	59	20	20
top	1	Debit Card	Male	4	Super	Married	3	14	1
freq	1351	4587	6328	4569	4062	5860	1746	1524	4373

Univariate Analysis of Data

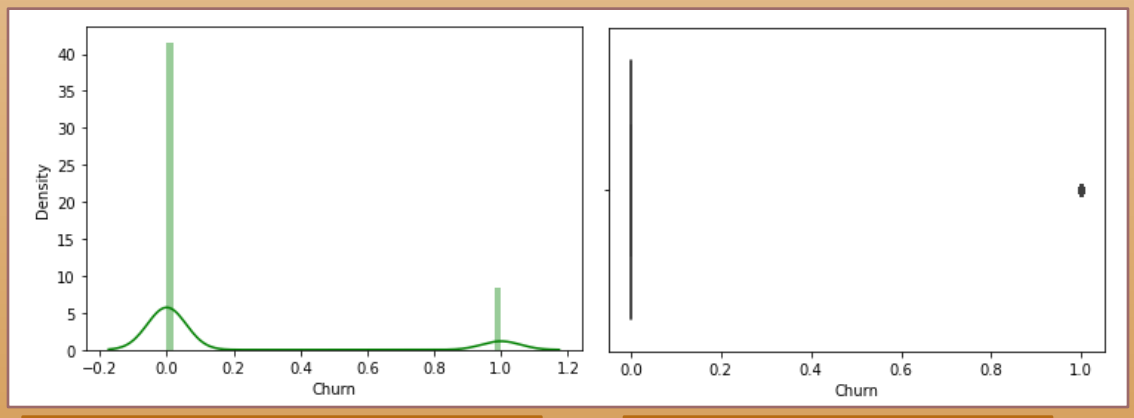
Exploratory data analysis for numerical data



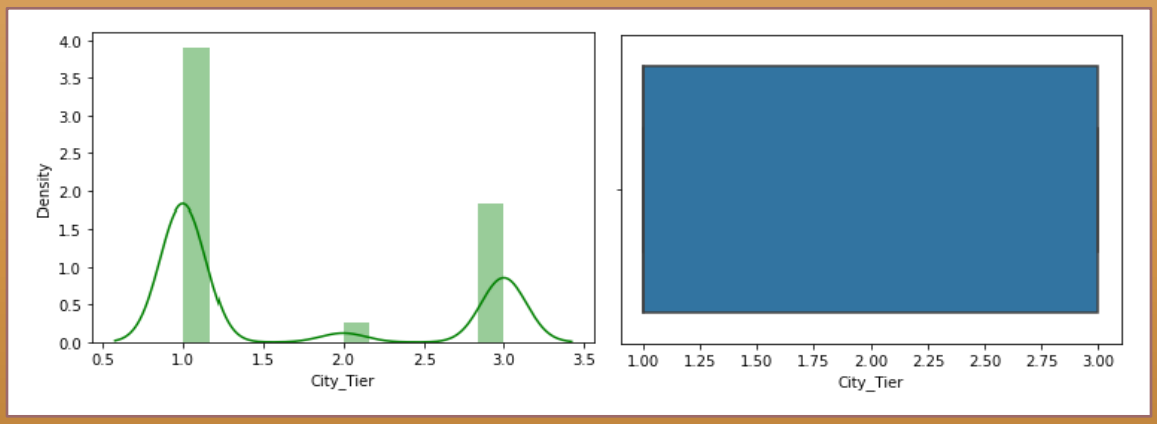
Distribution plot for Account ID



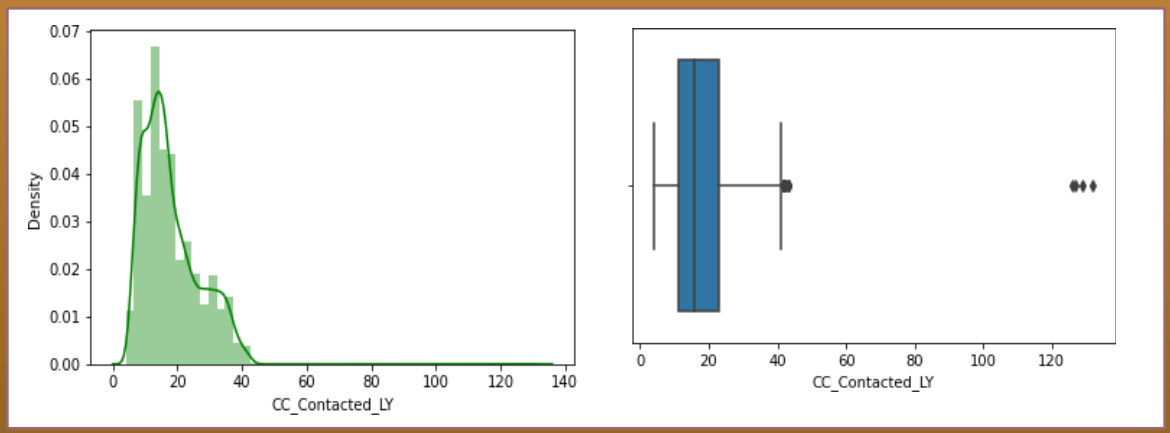
Boxplot for Complain_LY



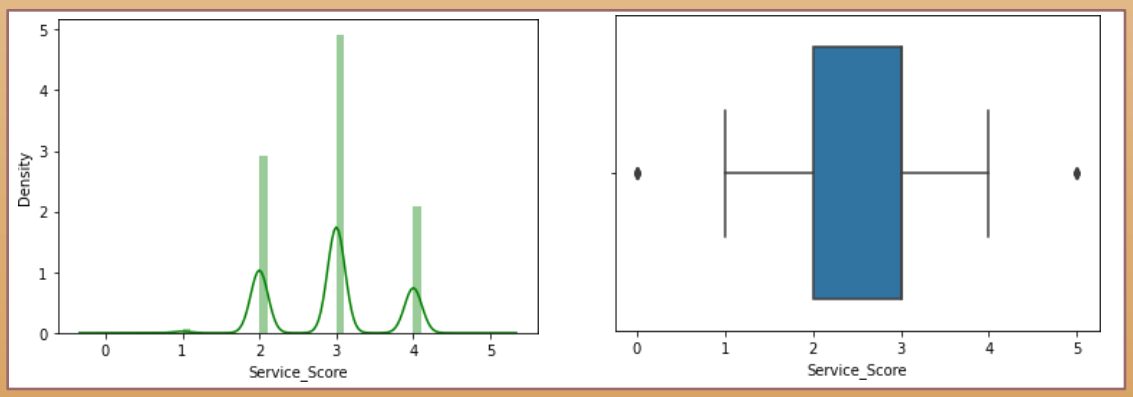
Distribution plot of Churn **Boxplot of Churn**



Distribution plot of City Tier **Boxplot of City Tier**

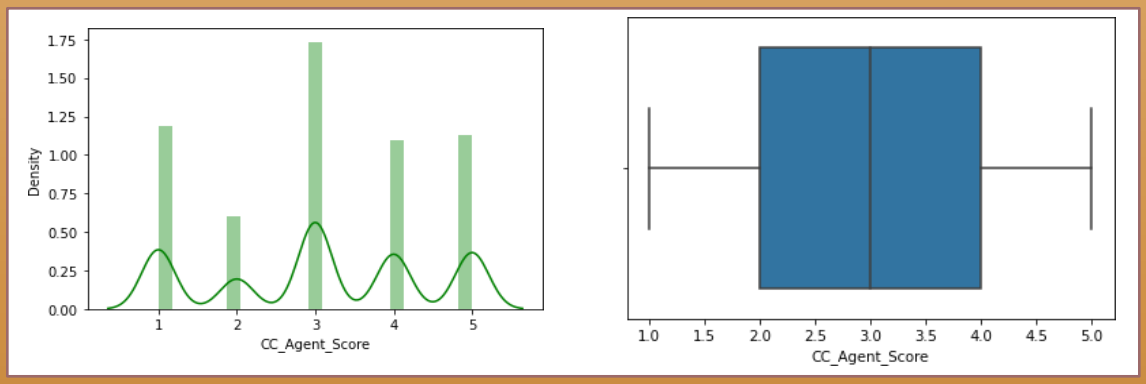


Distribution plot of CC_Contacted_LY **Boxplot of CC_Contacted_LY**



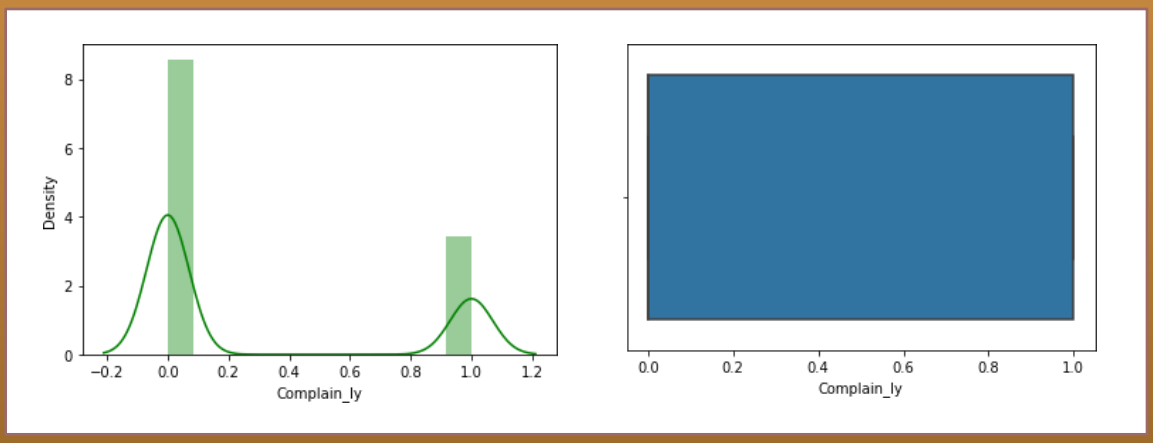
Distribution Plot of Service score

Boxplot of Service score



Distribution Plot of CC_Agent_Score

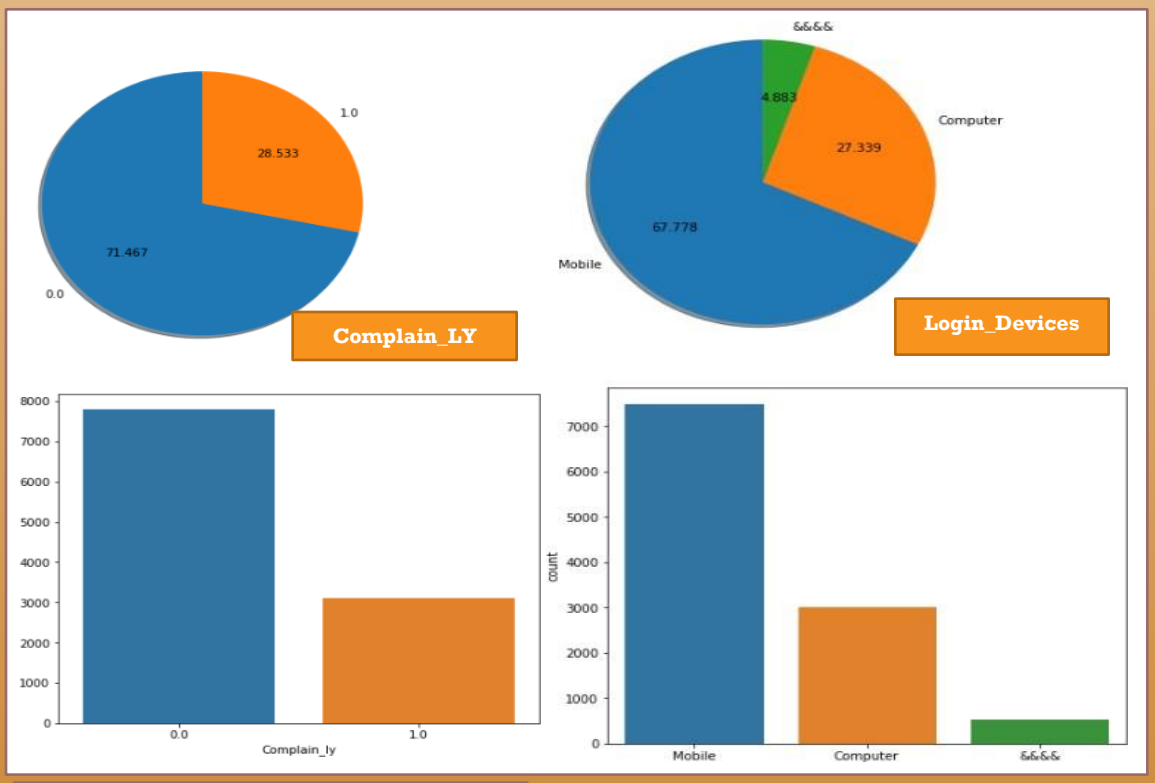
Boxplot of CC_Agent_Score

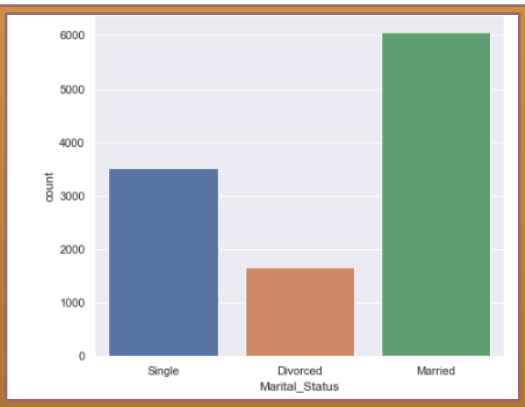


Distribution plot of Complain_ly

Boxplot of Complain_ly

Exploratory Analysis of Categorical variables



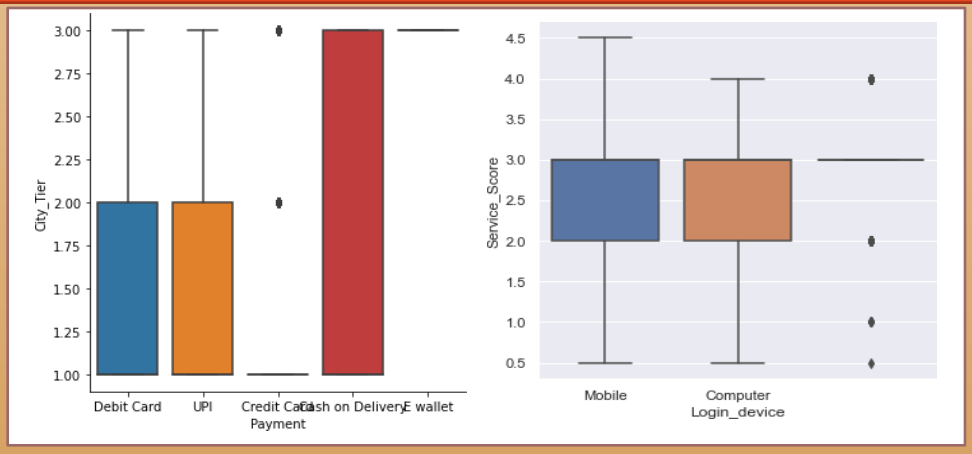


Marital Status

Facts to be noted:

1. 28% people have complained on the service in last year.
2. 68% of Customers use Mobile phones for login and only 27% of customers use desktop for login.
3. Married Customers constitute a larger proportion of the population (55%) and rest of them are single (32%) and Divorced (13%)

Bivariate Analysis of the data



Payment Vs City Tier

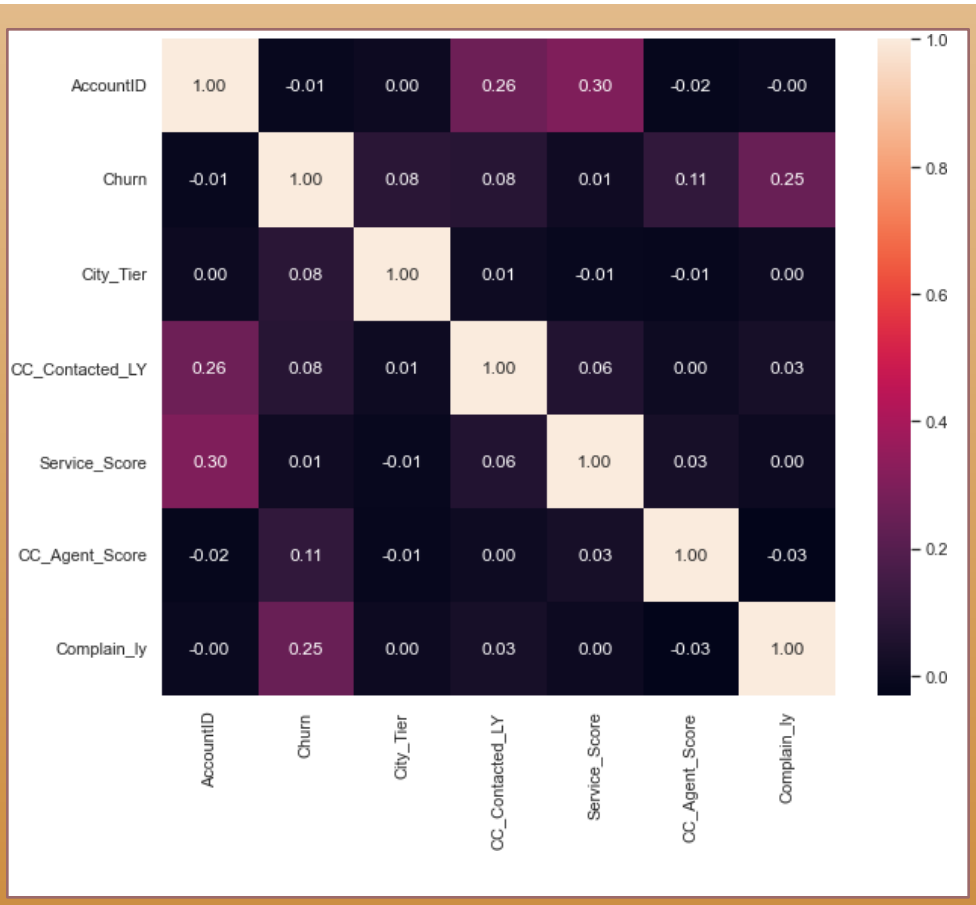
Login devices Vs Service Score

Here we see that the Customers residing in tier 1 cities are opting for digital payments and the customers residing over lower tier cities are opting for Cash payments.

There is no much difference on the service score marked by Desktop and Mobile users

Multivariate Analysis of data

Pair Plot



Heat Map

Even though we do not see a strong trend between the variables in the pairplot, We could see some positive correlation between (Complain_LY vs Churn) and (CC_agent_score vs Churn)

Data Cleaning and Pre-processing

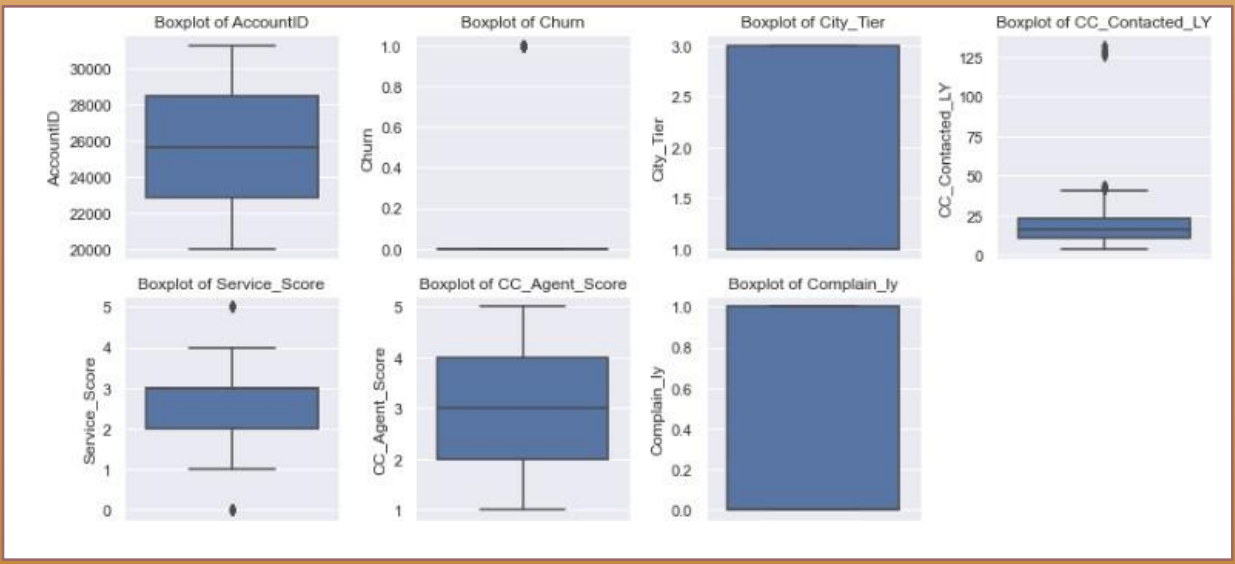
Removal of unwanted variables

While we work the ML model, we would have to firstly classify the variables into Dependent and Independent variables, we need only the independent variables for our model. In the current dataset, Account_id would be the dependent dataset that we do not require for the model preparation. Hence, we would get rid of the column while we go ahead with the model preparation.

Treatment for Missing values

Through the Exploratory data analysis we understood that there are 14 columns that has Missing values in the dataset,We should be filling this missing values with some values that would help the model to avoid skipping the records. We could fill the values using mean, median and mode method depending on the need of the dataset. In the present dataset we are choosing the Mode method to fill the missing values. Firstly we would be filling all the Missing values with text "NaN". Then replacing it with mode values. Basically Mode would replace the balnk fileds with the values that has high number occurrence in the respective columns.

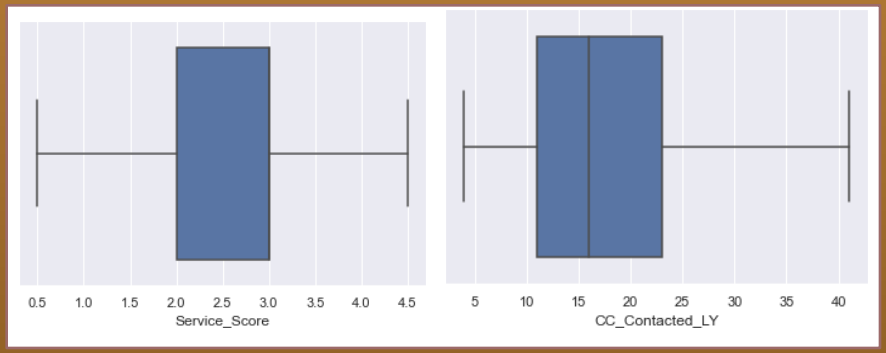
Outlier Treatment



Outlier treatment stands necessary for variables being finalized for our model. Outlier is nothing but a deformality or a impossible scenario in a respective case. Hence this needs to be stabilized.

We would be using the IQR (Inter Quartile Range) method for treating the outliers CC_Contacted_LY and Service_Score are the two variables which would need to undergo the outlier treatment

IQR (Inter Quartile Range) Inter Quartile Range approach to finding the outliers is the most commonly used and most trusted approach used in the research field.



Post Treatment view

Variable Transformation

The column Marital Status has three value Single, Married and Divorced. But divorced are also single. So Divorced value is converted into Single.

The column Gender having four value Male, Female, M, and F. But M are also Male and F are Female. So, M and F value are converted into Male and Female respectively.

Addition of new variables

There is no additional feature/variable introduced here. We would go ahead with the given set of variables

Applying SMOTE technique on the dataset

During the Univariate analysis of the dataset, we have understood that the data is imbalanced i.e. 83% of the sample belong to unchurned customers and only 17% belongs to churned customers. Due to which the model would not be able to understand maximum possibilities for the occurrence of our target variable. That is when SMOTE comes into picture to avoid the situation.

SMOTE stands for Synthetic Minority Over-sampling Technique. This method creates synthetic samples of your data, so rather than taking copies of observations, SMOTE uses a distance measure to create synthetic samples of data points that would not be far from your data points. We used this in our churn analysis to balance the data.

As you can see that the accuracy is quite low, and as it's an imbalanced dataset, we shouldn't consider Accuracy as our metric to measure the model, as Accuracy is cursed in imbalanced datasets. Hence, we need to check the recall, precision & f1 score for the minority class, and it's quite evident that the precision, recall & f1 score is too low for Class 1, i.e. churned customers. Hence, moving ahead to call SMOTEENN (UpSampling + ENN)

One Hot Encoding to treat categorical data parameters

Most Machine Learning algorithms cannot work with categorical data and needs to be converted into numerical data. Sometimes in datasets, we encounter columns that contain categorical features (string values)

All the categorical Data set applied with one-hot coding

Data split: Split the data into train and test (70:30)

1) Copy all the predictor variables into an X data frame. Since 'vote' is a dependent variable drop it . Copy the 'vote' column alone into the y data frame. This is the dependent variable

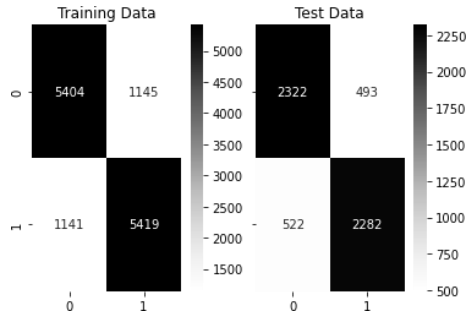
2) Split X and y into training and test set in 70:30 ratio

Model Building and Interpretation

Models applied on the dataset

Logistic Regression

Logistic regression is a statistical model used for binary classification that calculates the probability of an event occurring based on input features. It employs the logistic function to map input values to a range between 0 and 1, making it suitable for modeling probabilities and making binary decisions.



Confusion Matrix

Logistic regression Classfication report
Classification Report of the training data:

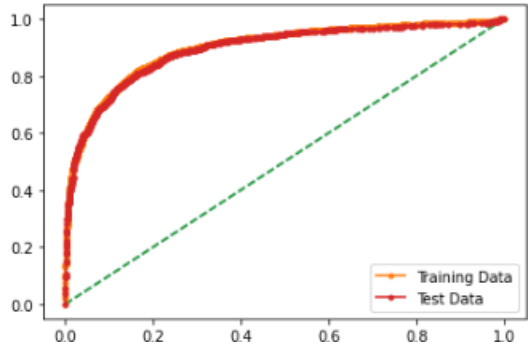
	precision	recall	f1-score	support
0	0.83	0.83	0.83	6549
1	0.83	0.83	0.83	6560
accuracy			0.83	13109
macro avg	0.83	0.83	0.83	13109
weighted avg	0.83	0.83	0.83	13109

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.82	0.82	0.82	2815
1	0.82	0.81	0.82	2804
accuracy			0.82	5619
macro avg	0.82	0.82	0.82	5619
weighted avg	0.82	0.82	0.82	5619

Classification Report

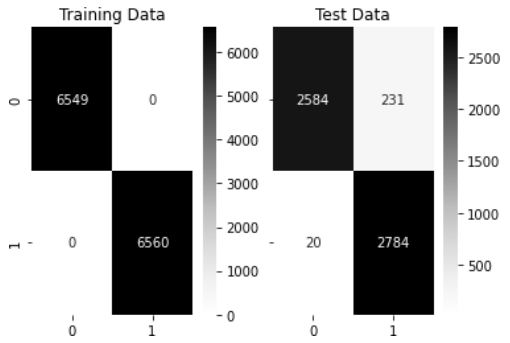
AUC and ROC FOR Logistic regression
AUC for the Training Data: 0.898
AUC for the Test Data: 0.894



ROC & AUC curve

KNN Model

K-Nearest Neighbors (KNN) is a simple supervised machine learning algorithm that classifies data points based on the majority class of their k nearest neighbors in the feature space. It's a non-parametric method that can be used for both classification and regression tasks.



Confusion Matrix

KNN Model Classfication report
Classification Report of the training data:

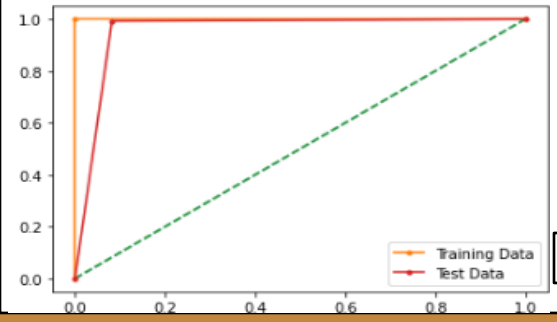
	precision	recall	f1-score	support
0	1.00	1.00	1.00	6549
1	1.00	1.00	1.00	6560
accuracy			1.00	13109
macro avg	1.00	1.00	1.00	13109
weighted avg	1.00	1.00	1.00	13109

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.99	0.92	0.95	2815
1	0.92	0.99	0.96	2804
accuracy			0.96	5619
macro avg	0.96	0.96	0.96	5619
weighted avg	0.96	0.96	0.96	5619

Classification Report

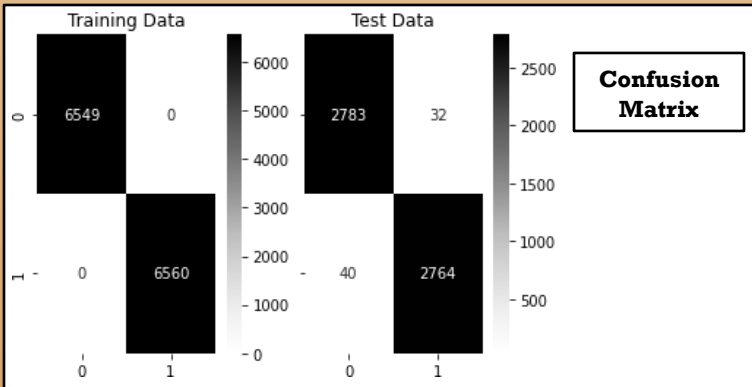
AUC and ROC FOR KNN Model
AUC for the Training Data: 1.000
AUC for the Test Data: 0.955



ROC and AUC curve

Random Forest

Random Forest is an ensemble machine learning model that combines multiple decision trees to make predictions. It improves accuracy and reduces overfitting by averaging the results of individual trees, making it a powerful tool for classification and regression tasks.



Confusion Matrix

Ensemble Random Forest Classifier Classfication report

Classification Report of the training data:

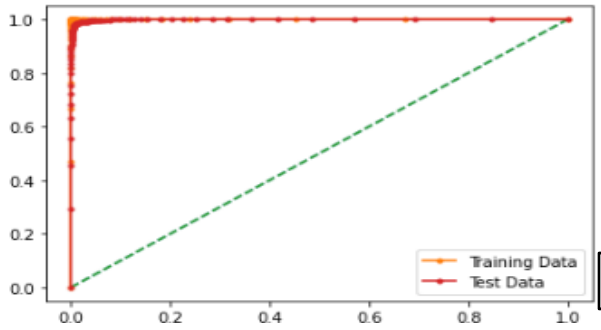
	precision	recall	f1-score	support
0	1.00	1.00	1.00	5089
1	1.00	1.00	1.00	6423
accuracy			1.00	11512
macro avg	1.00	1.00	1.00	11512
weighted avg	1.00	1.00	1.00	11512

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.98	0.96	0.97	2169
1	0.97	0.99	0.98	2765
accuracy			0.98	4934
macro avg	0.98	0.98	0.98	4934
weighted avg	0.98	0.98	0.98	4934

Classification Report

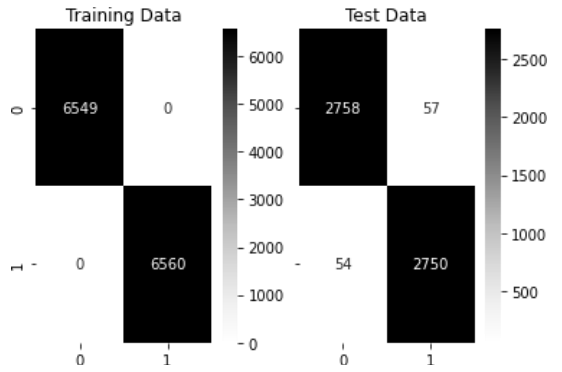
AUC and ROC FOR Ensemble Random Forest Classifier
AUC for the Training Data: 1.000
AUC for the Test Data: 0.999



ROC and AUC curve

Bagging Classifier

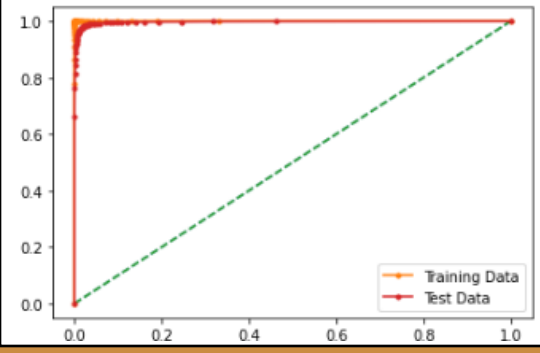
A Bagging Classifier is an ensemble machine learning model that combines the predictions of multiple base classifiers, typically decision trees, by training them on random subsets of the training data with replacement. This technique helps reduce variance and improve the model's overall accuracy and robustness.



Confusion Matrix

AUC and ROC FOR Ensemble Bagging Classifier
AUC for the Training Data: 1.000
AUC for the Test Data: 0.997

ROC and AUC curve



Classification Report

Ensemble Bagging Classifier Classfication report
Classification Report of the training data:

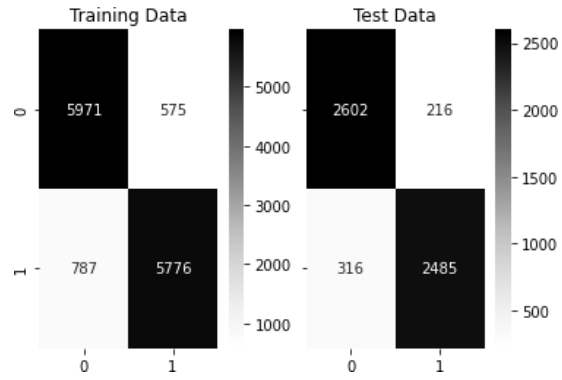
	precision	recall	f1-score	support
0	1.00	1.00	1.00	6549
1	1.00	1.00	1.00	6560
accuracy			1.00	13109
macro avg	1.00	1.00	1.00	13109
weighted avg	1.00	1.00	1.00	13109

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	2815
1	0.98	0.98	0.98	2804
accuracy			0.98	5619
macro avg	0.98	0.98	0.98	5619
weighted avg	0.98	0.98	0.98	5619

Gradient Boosting Model

Gradient Boosting is an ensemble machine learning technique that builds predictive models in a sequential manner by combining the strengths of multiple weak learners, typically decision trees. It minimizes errors by optimizing a loss function through gradient descent, resulting in a powerful and accurate predictive model.



Confusion Matrix

Classification Report

Ensemble Gradient Boosting Classfication report
Classification Report of the training data:

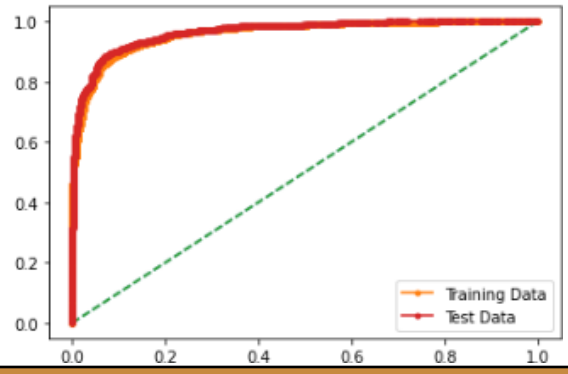
	precision	recall	f1-score	support
0	0.88	0.91	0.90	6546
1	0.91	0.88	0.89	6563
accuracy			0.90	13109
macro avg	0.90	0.90	0.90	13109
weighted avg	0.90	0.90	0.90	13109

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.89	0.92	0.91	2818
1	0.92	0.89	0.90	2801
accuracy			0.91	5619
macro avg	0.91	0.91	0.91	5619
weighted avg	0.91	0.91	0.91	5619

AUC and ROC FOR Ensemble Gradient Boosting
AUC for the Training Data: 0.961
AUC for the Test Data: 0.965

ROC and AUC curve



Model Tuning for better performance

Ensemble random forest:

best_params

{'max_depth': 6,'max_features': 2,'min_samples_leaf': 10,'min_samples_split': 50,'n_estimators': 150}

Predicting with the best grid search
Training Data Class Prediction with a cut-off value of 0.5 Test Data

Class

Prediction with a cut-off value of 0.5

Bagging Classifier

best_params

{'n_estimators': 20}

Predicting with the best grid search

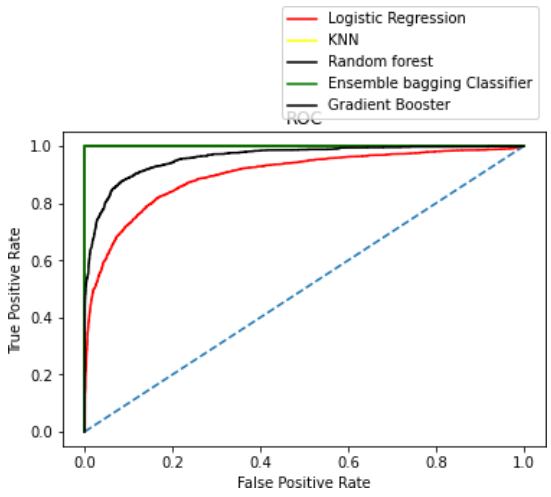
Training Data Class Prediction with a cut-off value of 0.5 Test Data

Class Prediction with a cut-off value of 0.5

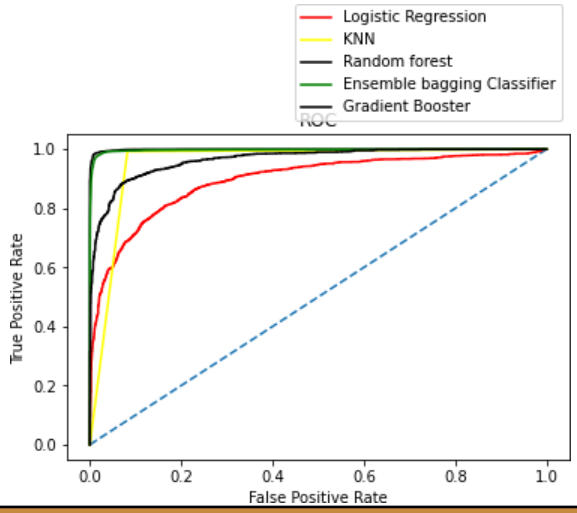
Model Validation

	Logistic reg Train	Logistic reg Test	KNN Train	KNN Test	Random Forest Train	Random Forest Test	Bagging cls. Train	Bagging cls. Test	Gradient Boosting Train	Gradient Boosting Test
Accuracy	0.83	0.82	1.0	0.96	1.0	0.99	1.0	0.98	0.90	0.91
AUC	0.90	0.89	1.0	0.96	1.0	1.00	1.0	1.00	0.96	0.97
Recall	0.83	0.82	1.0	0.92	1.0	0.99	1.0	0.98	1.00	0.97
Precision	0.83	0.82	1.0	0.99	1.0	0.99	1.0	0.98	1.00	0.98
F1 Score	0.83	0.82	1.0	0.95	1.0	0.99	1.0	0.98	1.00	0.97

Classification Report of all applied models



ROC and AUC for Train data



ROC and AUC for Test data

Accuracy and error rate is the de facto standard metrics for summarizing the performance of classification models. Classification accuracy fails on classification problems with a skewed class distribution because of the intuitions developed by practitioners on datasets with an equal class distribution.

Why use Precision and Recall in Machine Learning models?

This question is very common among all machine learning engineers and data researchers. The use of Precision and Recall varies according to the type of problem being solved.

oIf there is a requirement of classifying all positive as well as Negative samples as Positive, whether they are classified correctly or incorrectly, then use Precision.

oFurther, on the other end, if our goal is to detect only all positive samples, then use Recall. Here, we should not care how negative samples are correctly or incorrectly classified the samples.

After lot of feature engineering and tuning most of them appear overfitted.

So I would go with Logistic regression model. Having good precision and Recall value

Final Interpretation/ Recommendation

Insights

Tier 1 people use credit cards more and lower tier people depend on physical cash

Mobile users constitute the majority of customer.

The Service score of married people is better compared to others sections.

Super+ Account accounts have the least churn rate.

Married People have a low churn rate, While Single have more churn.

Married people are satisfied with company service

Recommendation

Provide cashback offers on digital payment and spread awareness among lower tier people on the offers available on making digital payment.

Keep the Mobile UX up to date so that mobile users do not drop out for poor CX. Also try to improve desktop experience for the same.

Provide additional Discount to the loyal customers on their recharge bill in order to keep them engaged with us.

Need to look into the Super+Account customer's compalaints as they constitute the major part of complaint.

Need to collect the feedback from the churned customer in case they have lodged 0 complains in the past.

Business may provide introductory offers to attract new customers and exclusive offers to existing new customers.