Customer Churn Prediction- Final report

Name: Adithya Manivannan

PGP Data Science & Business Analytics

Table of Contents

[Project Notes -1 Customer Churn Prediction 0](file:///C:\Users\Adithya%20Manivannan\Downloads\CC_EDTH_02_Customer+Churn_02\ProjectNotes_1_Adithya_Manivannan.docx#_Toc113112895)

[List of Figures 1](#_Toc113112896)

[List of Tables 2](#_Toc113112897)

[**Customer Churn Prediction Project:** 2](#_Toc113112898)

[**Introduction** 2](#_Toc113112899)

[**Data Report** 3](#_Toc113112900)

[**Exploratory data analysis & Business Insights** 5](#_Toc113112901)

[**Data Cleaning and Pre-processing** 16](#_Toc113112902)

[**Model building and interpretation.** 19](#_Toc113112903)

[**Models** 19](#_Toc113112904)

[1. Logistic Regression 19](#_Toc113112905)

[2. Linear Discriminant Analysis 20](#_Toc113112906)

[3. Naïve Bayes Model 21](#_Toc113112907)

[4. KNN Model 22](#_Toc113112908)

[5. Decision Tree Model 23](#_Toc113112909)

[6. Random Forest 23](#_Toc113112910)

[**Business insights** 28](#_Toc113112911)

[**Business Recommendations:** 29](#_Toc113112912)

# List of Figures

Fig. 1. - Marital Status Distribution 6

Fig. 2. - Payment Method Distribution 7

Fig. 3. - Gender Distribution 7

Fig. 4. - Account Segment Distribution 8

Fig. 5. - Complain Raised Distribution 8

Fig. 6. - Churn Rate Distribution 9

Fig. 7. - Number of customers by tenure 9

Fig. 8. - Number of customers by Service Score 9

Fig. 9. - Number of customers by Account User Count 9

Fig. 10 - Number of customers by Agent Score 9

Fig. 11 - Number of customers by cashback 9

Fig. 12 - Churn by Account Segment 13

Fig. 13 - Churn by Login Device 13

Fig. 14 - Churn by City Tier 14

Fig. 15 - Churn by Service score 14

Fig. 16 - Churn by Agent Score 15

Fig. 17 - Pair plot 15

Fig. 18 - Correlation Heatmap Plot 16

Fig. 19 – Boxplot with outliers 17

Fig. 20 – Boxplot without outliers 18

Fig. 21. - Classification Matrix – Training Data– Logistic Regression 21

Fig. 22. - Classification Matrix – Test Data– Logistic Regression 21

Fig. 23. - AUC Curve – Train & Test – Logistic Regression 21

Fig. 24. - Classification Matrix – Training Data– Linear Discriminant Analysis 22

Fig. 25. - Classification Matrix – Test Data– Linear Discriminant Analysis 22

Fig. 26. - AUC Curve – Train & Test – Linear Discriminant Analysis 22

Fig. 27. - Classification Matrix – Training Data– Naïve Bayes Model 23

Fig. 28. - Classification Matrix – Test Data– Naïve Bayes Model 23

Fig. 29. - AUC Curve – Train & Test – Naïve Bayes Model 23

Fig. 30 - Classification Matrix – Training Data– KNN Model 24

Fig. 31 - Classification Matrix – Test Data– KNN Model 24

Fig. 32 - AUC Curve – Train & Test – KNN Model 24

Fig. 33 - Classification Matrix – Training Data– Random Forest Model 25

Fig. 34 - Classification Matrix – Test Data– Random Forest Model 25

Fig. 35 - AUC Curve – Train & Test – Random Forest Model 25

Fig. 36 – Feature Importance 30

# List of Tables

Table 1. – Information on Dataset 4

Table 3. – Description of Dataset 5

Table 4. – Updated Information on Dataset 5

Table 5. – Description of Dataset 6

Table 6. – Model Comparison table 29

# **Customer Churn Prediction Project:**

# **Introduction**

**Problem Statement**

The problem statement provided is that an E Commerce company is facing a lot of competition in the current market and it has become a challenge to retain the existing customers in the current situation. Hence, the company wants to develop a model through which they can do churn prediction of the accounts and provide segmented offers to the potential churners. In this company, account churn is a major thing because 1 account can have multiple customers. hence by losing one account the company might be losing more than one customer. We have been assigned to develop a churn prediction model for this company and provide business recommendations on the campaign.

**Need of the study/project:**

The need of the study is to understand who are the customers who are churning and what are the factors that influence the same. This can be done by classification of customers into different segments and as we know, in a competitive world like ours, each company faces stiff competition from other companies in the industry and are constantly trying to poach customers from each other. As customers, even they have a lot of different options and hence different factors influence in customers churning from a company and adopting another. It is noted that the Customer Acquisition Costs (CAC) is usually way higher than Customer Retention Costs and hence if a company is able to predict the customer churn, it will help them invest more into retention than acquisition. This will eventually increase the company’s profitability in the long term.

**Understanding business/social opportunity**

As mentioned above, churn affects an organization in many ways and hence being able to predict churn and work towards reducing it offers a great business opportunity for companies to be more profitable and maintain better relationships with their customers. Some of the effects of churn are:

* It reduces the profitability
* Poor brand image
* Reduces the market share
* Affects prospects of future business
* Increased customer acquisition costs lead to lower profit margins

Keeping the above in mind, we can see that if we are able to classify customers into different segments and predict the ones who are on the verge of churning, this will help the company save a lot

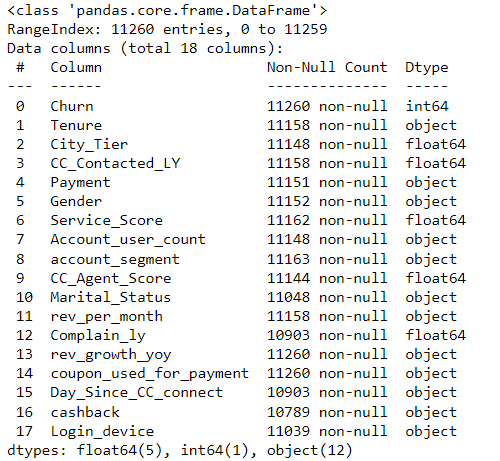
# **Data Report**

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

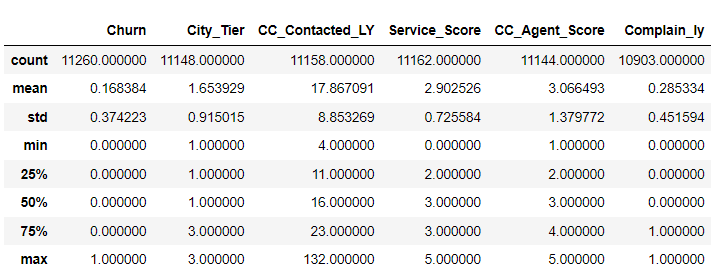
We believe that at a point of time, the data of all the customers in an organization is being collected along with other factors such as demographics, their account details and whether or not they have churned. Since the days since customer contacted support is there as a metric, we can confirm our assumption that the data is observed on a date. Most of the data available is obtained from the customers account details and the usage of the same.

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

The data set has 11260 rows and 19 columns by default.



The above information of the dataset shows that by default there are 5 float variables, 1 int variable and 12 object variables. We can also see that except Churn and “Coupon used for payment”, every other variable has null values.

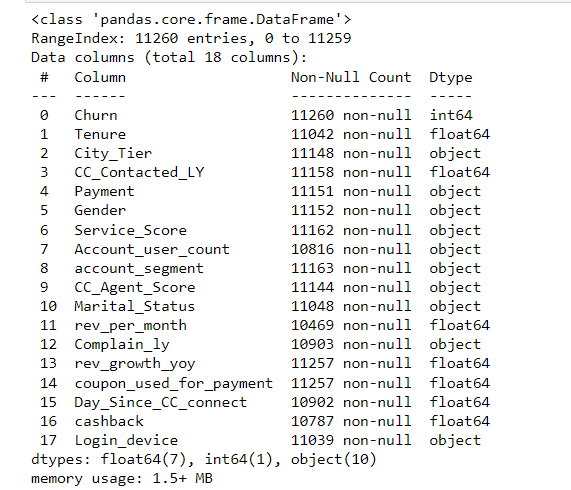


Looking at the description of the data, we can see that some of these variables have been incorrectly marked as float or int while they should be an object datatype while some of those classified as object should in fact be a continuous variable.

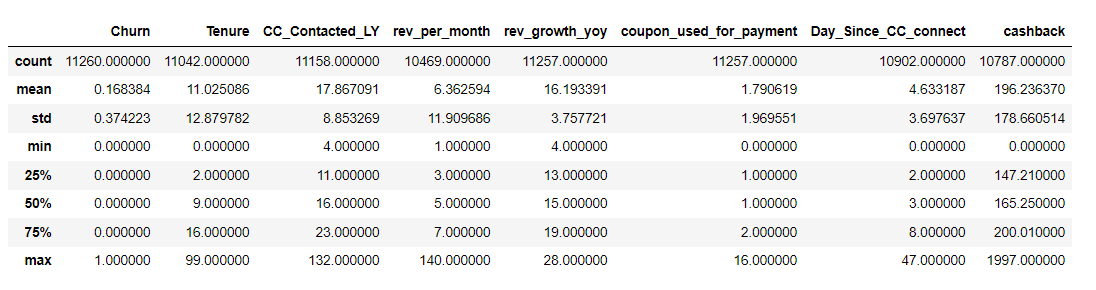
Hence, we changed the datatype and assign the correct ones.

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

We type cast the datatype correctly and classify them as below.



As we can see now, the Dtype of some of the variables have been updated and the data makes more sense now.



Looking at the description now, we can say that the mean tenure of an account is 11.02 while the median is 9.00. This could suggest the existence of outliers in the column.

Similarly, the average revenue per month from an account is 6.3 while the average revenue growth year on year is around 16%.

We can note that the average coupons used for payment in the last year is 1.79 and the median is 1 suggesting that not a lot of coupons are being used by customers for making payment. Also, the average cashback per account is around 196 and the average days since customer connected with customer care is 4.6.

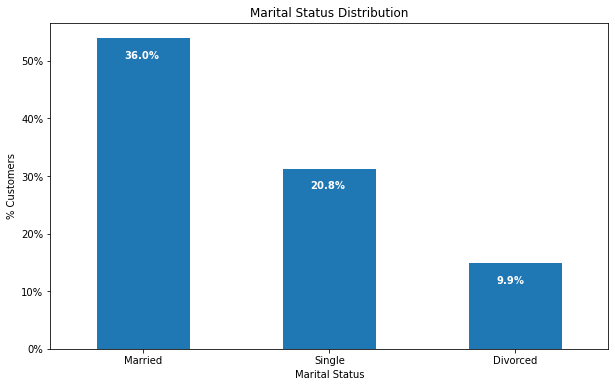
We also remove the column “Account ID” since it is a unique value to each row and hence won’t add any value in the model building.

# **Exploratory data analysis & Business Insights**

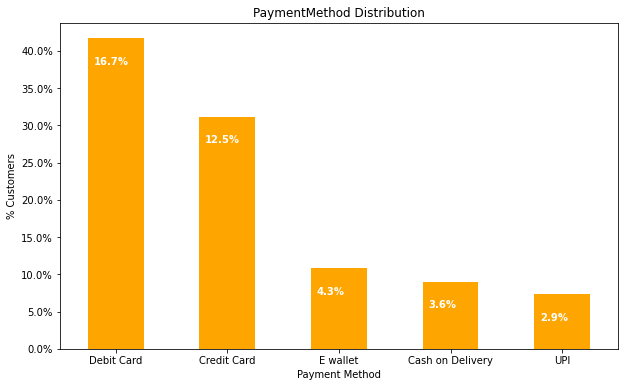
## 

**Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)**

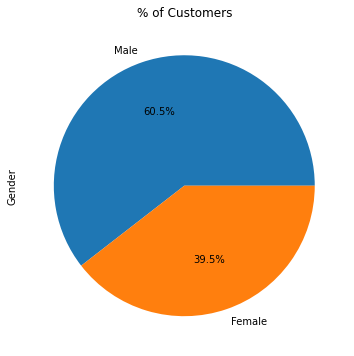
Distribution of data for categorical variables:



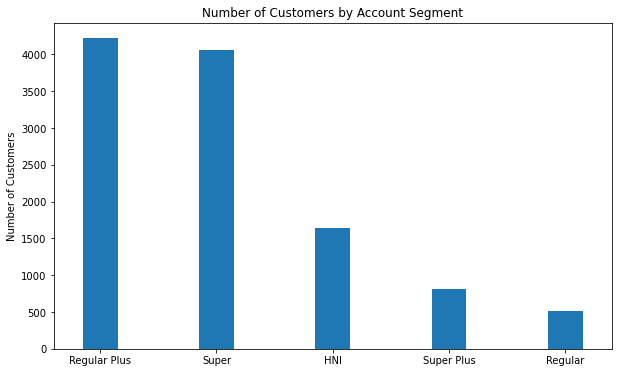
The above figure shows the distribution of data across marital status. We can see that 36% of our customers are married while 20.8% are single and the remaining 10% are divorced.



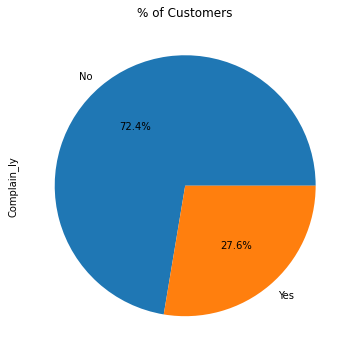
We can see that Debit Card is the most popular payment method with 16.7% of customers opting for this and closely followed by Credit Card with 12.5%. The other methods such as E wallet, COD and UPI are relatively smaller.



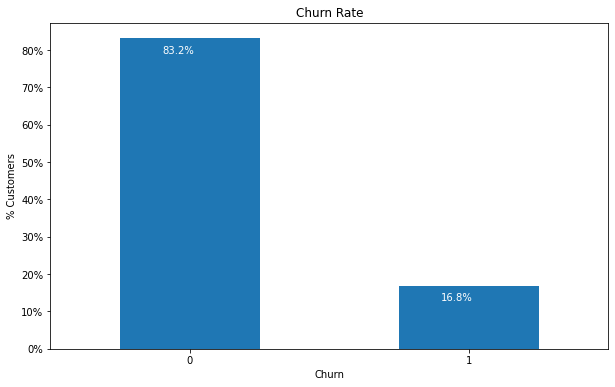
We can see that 60.5% of the customers base is Male while 39.5% of them are Female.



It can be observed that “Regular Plus” is the segment of customers that has the highest number closely followed by Super. The HNI has an average count while the “Super Plus” and “Regular” have relatively lower count of the same.

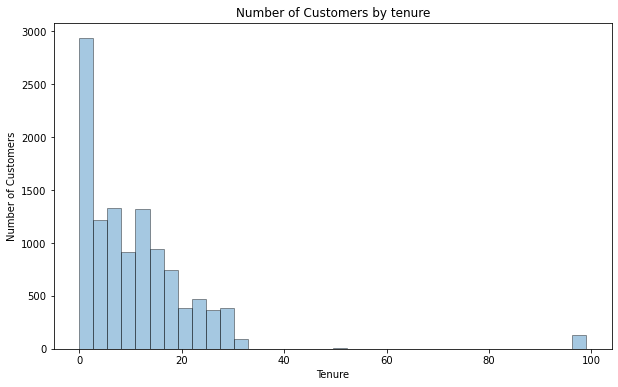


We can see that more than 72% of customers have not raised a complaint in the last 12 months while 27.6% of them have.

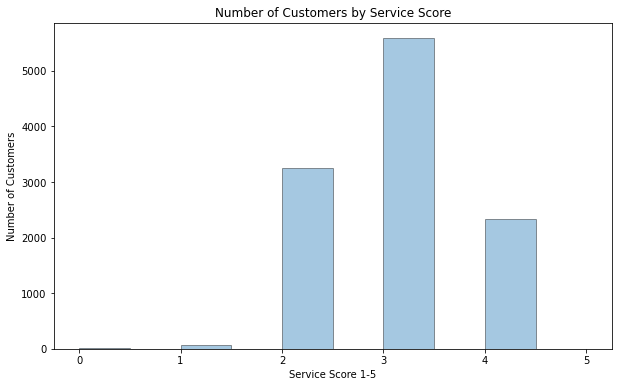


We can see that this is an unbalanced dataset with more than 83% of accounts in the dataset have not churned while only 16.8% have churned. Hence, the dataset would be biased and we would have to act upon it as we proceed to get a good model.

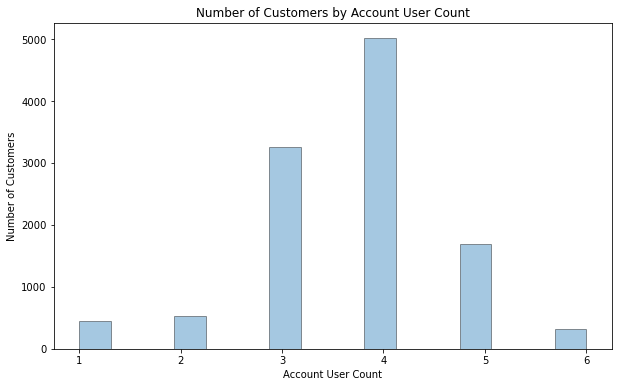
Distribution and spread for continuous variables:



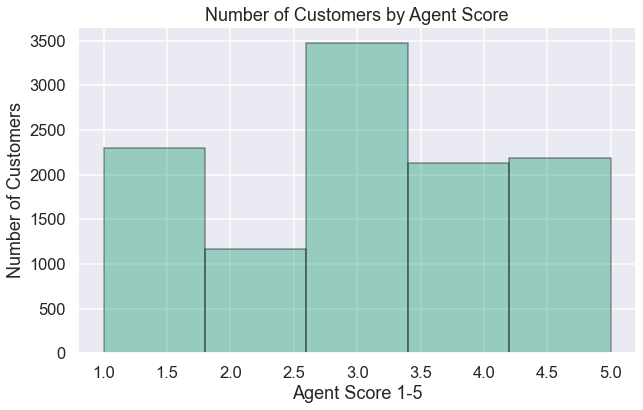
We can see the spread of customers based on the tenure. It can be observed that most of our customers are relatively newer and that the graph decreases as the tenure increase. We can also see a customer with 99 years of tenure. This is an outlier and we will be treating it.



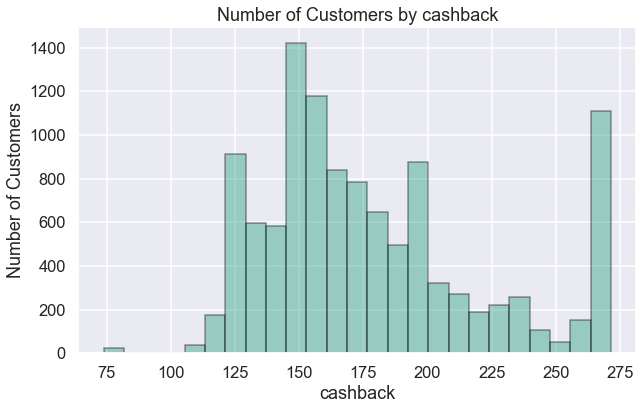
It is alarming to note that the number of 5 service scores is almost negligible while most of the customers have given a score of 3 which suggests that we have an average service according to our customers. This is something we need to analyze further and work on.



On an average, we can see that most of our accounts have 4 users while accounts with 3 and 5 users are pretty common as well.

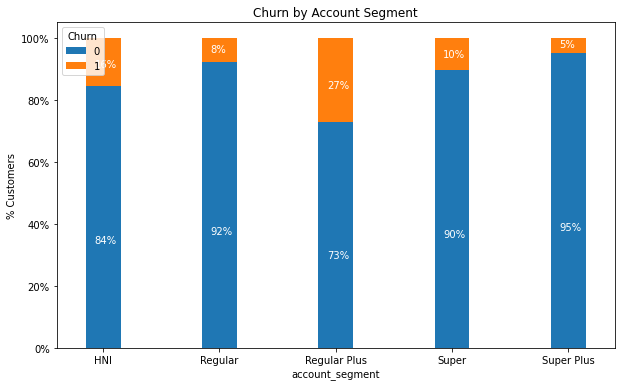


We can see that most common score obtained by our agents is 3 followed by 1. This shows that a lot of our customers are not satisfied with the agents service in resolving their issue.

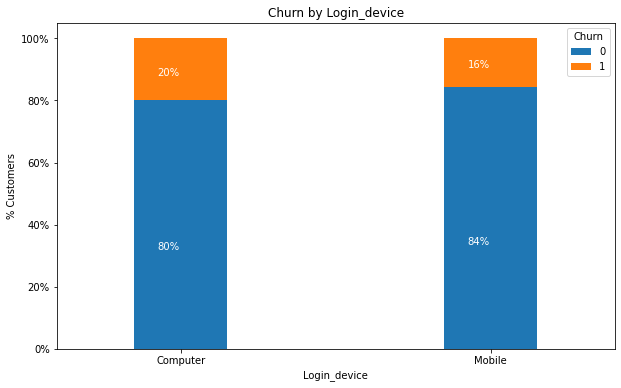


We can see the distribution of customers by the cashback obtained in the last 12 months. There are quite a few customers who have obtained a large cashback while excluding them, most of the remaining follow a bell-shaped curve showing a distribution as expected.

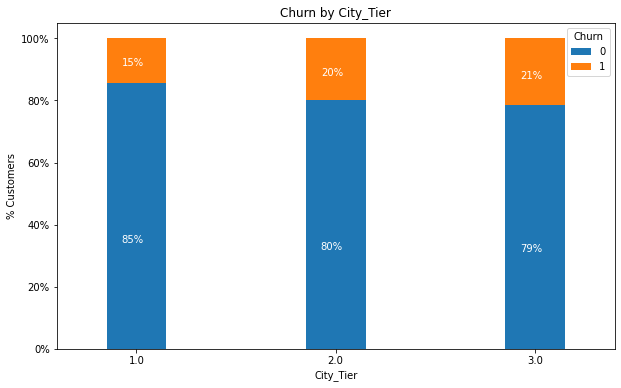
**Bivariate analysis (relationship between different variables, correlations)**



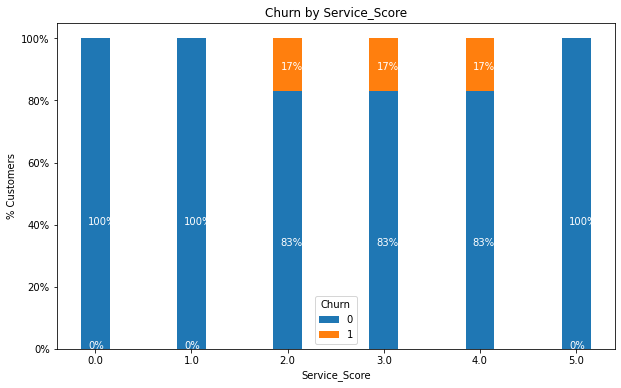
We see a distribution of Account segments and churn based on it. We can see that within each segment of customers, 27% of Regular plus segment churn. This shows that this is a segment we need to focus on to understand more about what is going wrong. On the other end of the spectrum, Super Plus segment has a very low churn rate and this is a segment we can learn from.



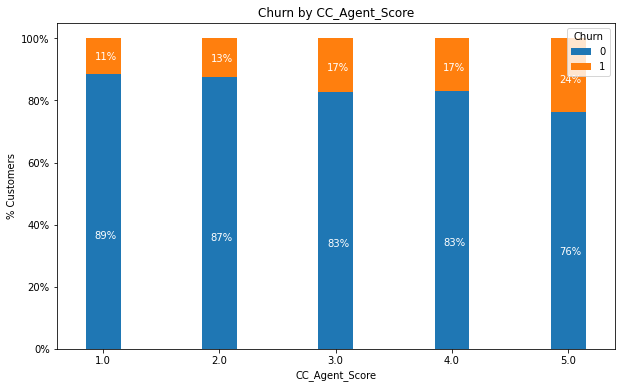
There is not much difference between the login\_device used and churn. This does not seem like a variable that adds much value.



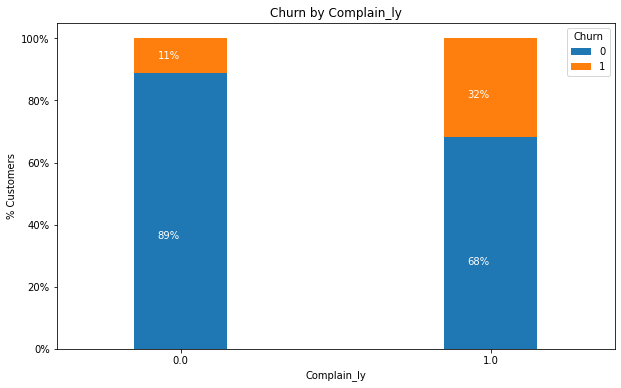
It can be noted that Tier 1 cities have a lower churn rate compared to tier 2 and 3. Tier 3 as expected has the higher churn rate. This could be due to tier 2 and 3 being more cost sensitive and less brand loyal.



It needs to be noted that only customers who have provided a service score of 2,3&4 have churned. Surprising to note that customers who have provided a 1 service score have not churned.

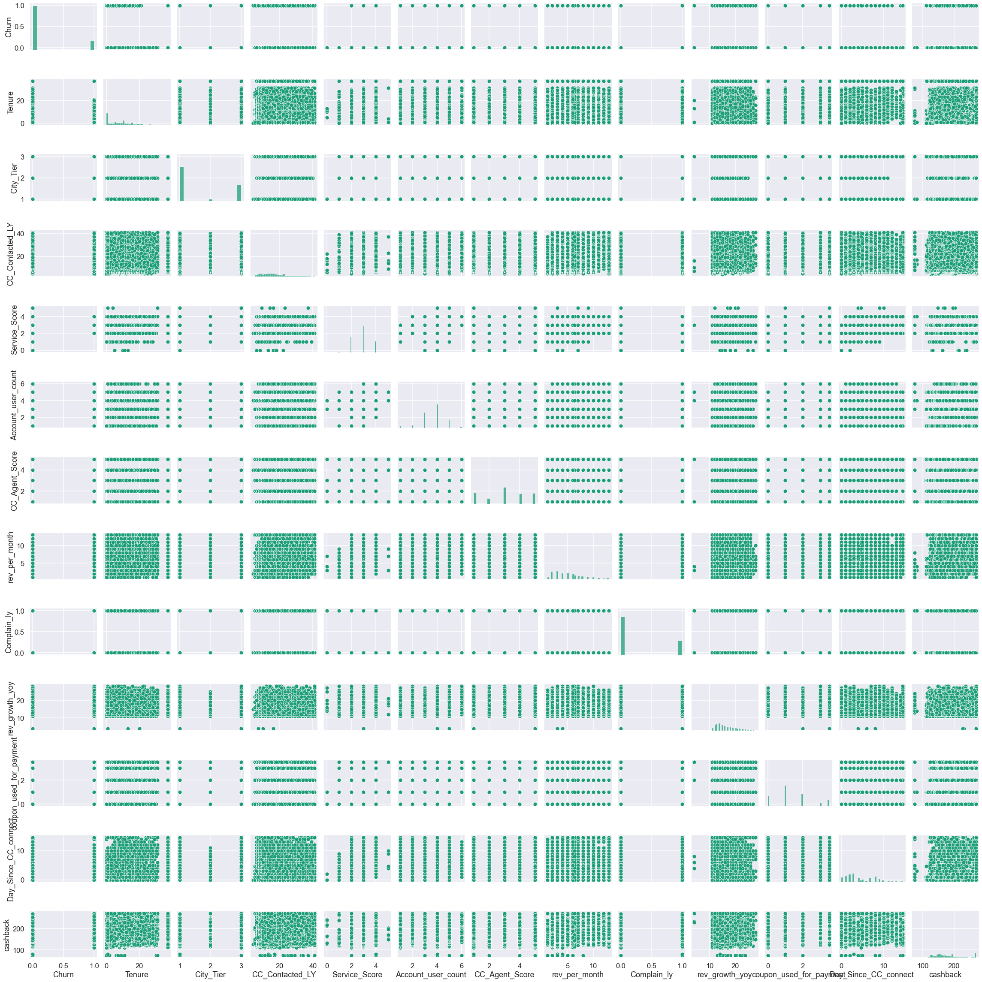


Another strange thing to note is that of customers who have provided an agent score of 5, 24% have churned while of those who have provided 1, only 11% have churned. We would expect it to be the other way and since it is not, it has to be noted and analyzed more.



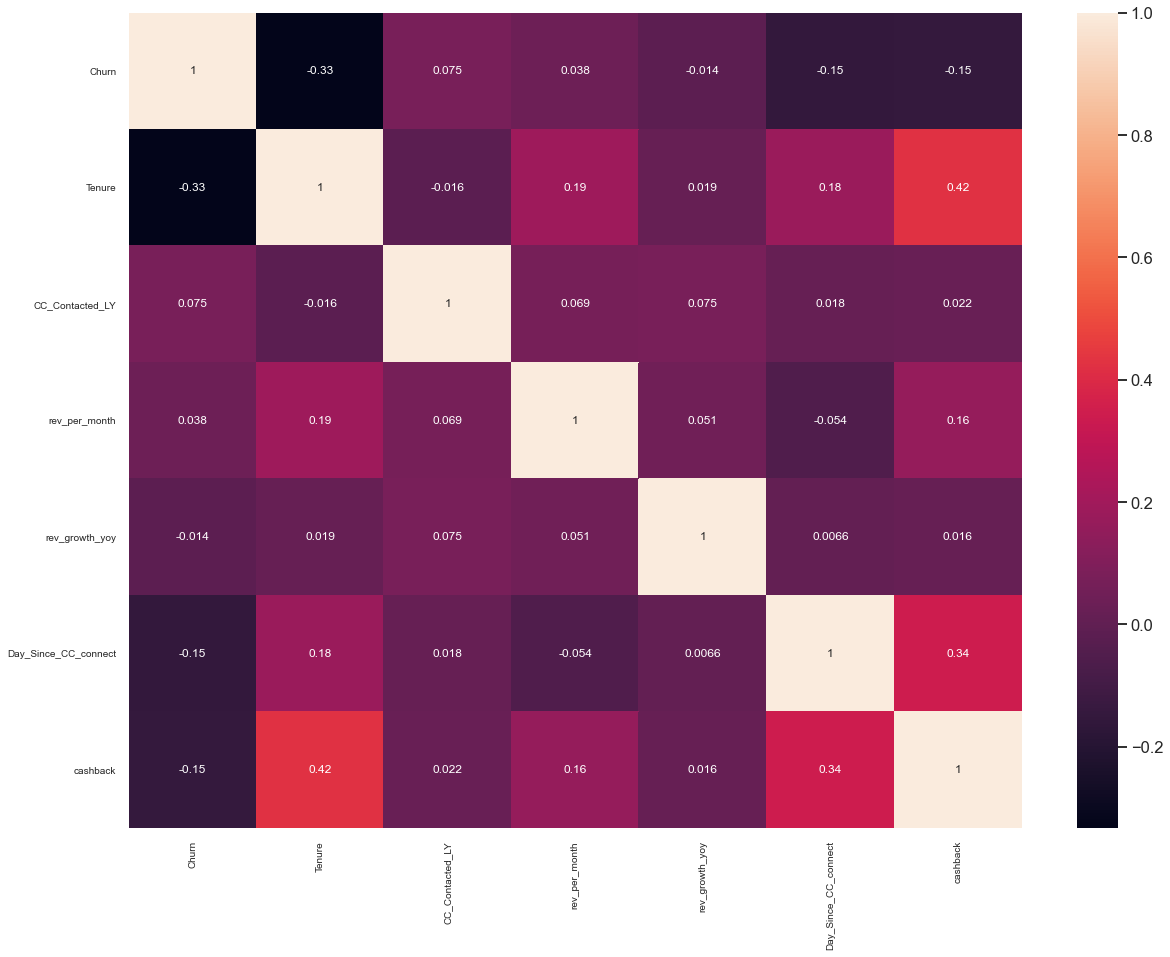
As expected, there is a higher percentage of churn from customers who have complained in the last 12 months rather than those who have not.

Below, we have a pair plot to look for any linear relations between variables and to look for patterns.

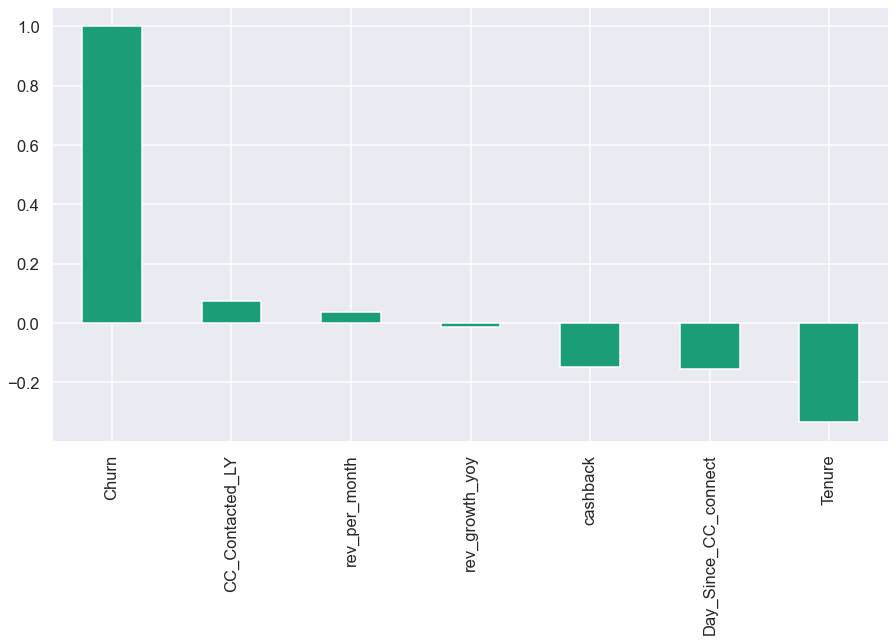


It can be noted that there is hardly any strong linear relationship between 2 variables based on the above plot.

We can confirm this with the help of a heatmap.



The heatmap shows that there is not much of positive correlation between any two variables. But we can see a strong negative correlation between a few of the variables. This is something we need to analyze further.



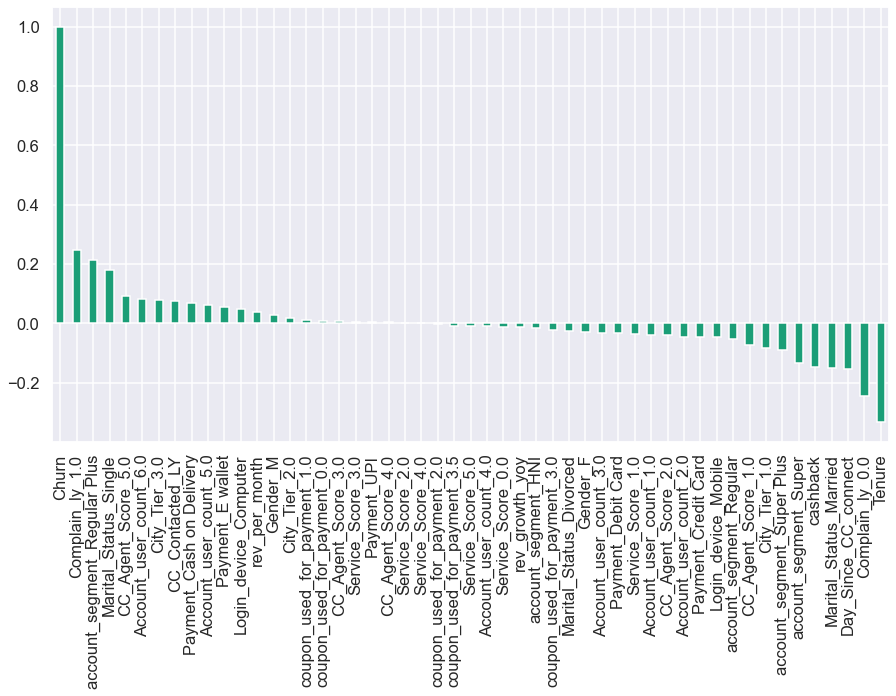
The above image also shows how much each variable correlates to “Churn”. As noted above, there is high negative correlation between Churn and Tenure. The other correlations are negligible as they are not strong.

**Business insights from EDA**

We have cleaned and process the given data to perform EDA so far and to check for relationships between variables before we go to the model building. As of the observations made now, we can see that one of the major factors influencing churn is whether a customer has complained in the last 12 months or not.

Based on the high correlation to churn, customers who have complained in the last 12 months, belong to the “Regular Plus” segment and are Single have a higher probability of churning compared to other groups. So, we need to focus on these groups to provide them with better customer service, offer discounts and do all that we can to retain them.

Similarly, those with a longer tenure, have not complained in the last 12 months and are married are some of our most loyal customers and we need to continue keeping them happy.



# **Data Cleaning and Pre-processing**

**Removal of unwanted variables (if applicable)**

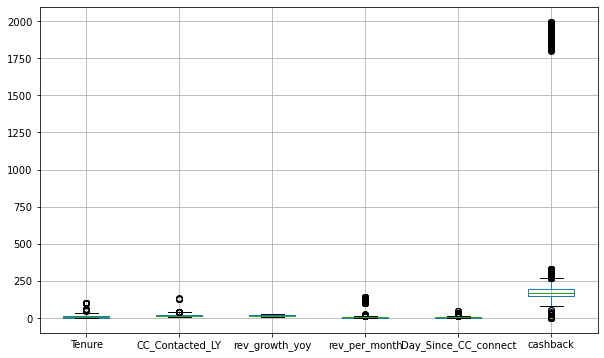
We have removed the column “Account ID” since it is a unique value to each row and hence won’t add any value in the model building.

We also checked each variable to see if they have bad data i.e. data that is not usable or junk such as “$”,”%”,”&” etc. We replaced all of them with null values and will treat them as the same.

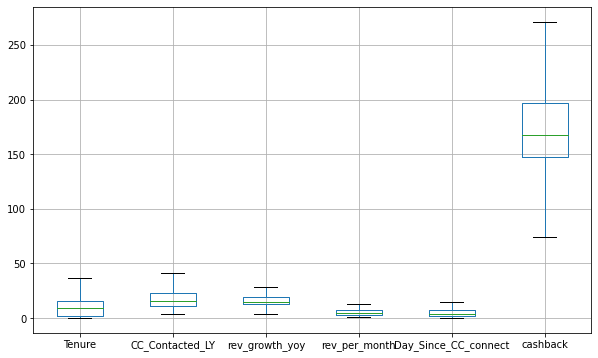
**Missing Value treatment (if applicable)**As mentioned above, we replaced junk data as null values and treated them together. We treat them by replacing the null values with the mean of the column for continuous variables and by the mode of the column for object class variables. By this way, we do not affect the variable in a big way and it retains its statistical properties.

**Outlier treatment (if required)**

Upon treating the null values, we next check for outliers in the data set. It was observed that most of the continuous variables had outliers.



Since outliers will impact the model, we need to get rid of them. Removing the outliers is not an option since this will massively reduce the number of data we have. Hence, we decide to treat them by capping them within the Inter quartile range. So, if a data is more than the upper range, the value will be replaced by the upper range and similarly, if a data is below the lower range, it would be set to the lower range. After the outlier treatment, the boxplot can be seen below.



**Variable transformation (if applicable)**

We observed that a lot of the variable’s data types had been set incorrectly and hence based on our understanding, they had been transformed to the right data type. We perform dummy encoding to transform the categorical variables into dummy variables. This increased the total number of columns to 51 including the dummy columns that were created.

**Addition of new variables (if required)**

We have not added new variables into the dataset yet as we do not se the need for the same. While we start the model building process, if we see a need for it, we can come back to modify this.

**Unbalanced Data**

We observe that we have an unbalanced data in hand since the number of customers who have churned in the dataset is relatively lower compared to those who have not churned. If we do not solve this before building the model, the model would be biased and rate a higher number of customers as not churning despite them actually being on the verge of churn. Hence, we would have to use one of different techniques such as oversampling or under sampling resampling technique or go for another option such as using SMOTE to generate synthetic samples for minority class.

By doing so, we can reduce the bias of the model and generate a model that makes predictions accurately.

# **Model building and interpretation.**

With the given data with factors such as demographics and the account details, we build models to predict the customers who would churn. Since this is a classification problem, we use models such as Logistic Regression, KNN, LDA etc and we will then use different ensemble model techniques to improve the performance metrices.

**Balancing of the dataset, standardization and encoding**

The dataset we have is unbalanced and hence we perform **SMOTE** to synthetically increase the minority class of dataset.

We then divide the dataset into numerical and categorical data. Here, we convert the categorical data using get\_dummies function and for the numerical data, we use the min-max function to standardize the data to avoid bias.

The dataset is split into train and set data in the 70:30 ratio.

# **Models**

## Logistic Regression

 A logistic regression model predicts a dependent [‘churn’] data variable by analyzing the relationship between one or more existing independent variables.

We use the sklearn.linear library to use the Logistic Regression function and train the model and test it.

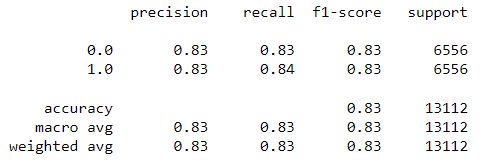
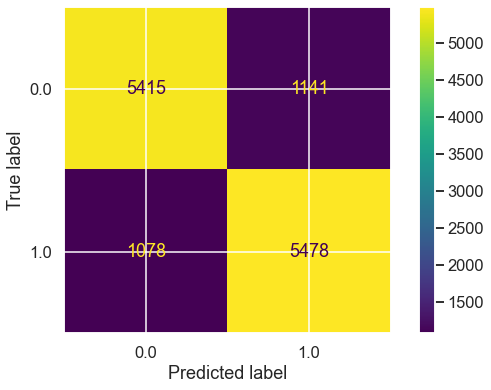
We then use the GridSearchCV method to obtain the best parameters and then perform the logistic regression.

{'penalty': 'none', 'solver': 'sag', 'tol': 0.0001}

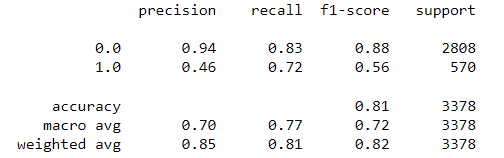
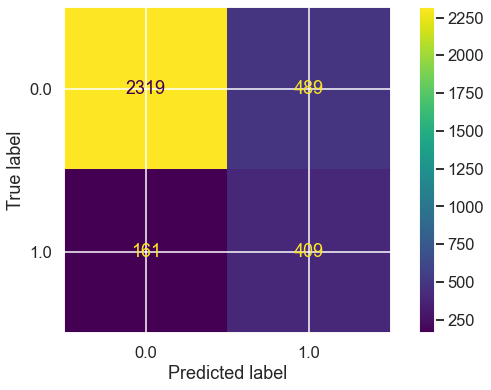
LogisticRegression(max\_iter=10000, n\_jobs=2, penalty='none', solver='sag')

These have been selected as the best params by the program and hence we go ahead with the same.

**Classification Matrix on the training data:**

**Classification Matrix on the test data:**

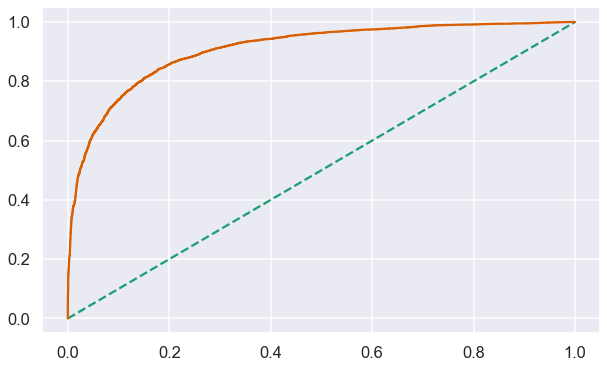
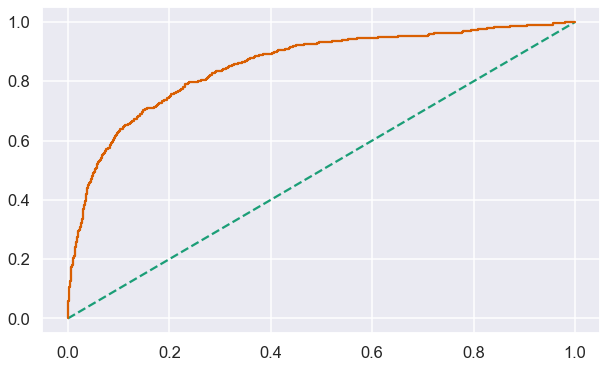
Although the accuracy scores are good for the test and train set, the recall score has a difference of more than 10 and hence the model is overfit.

**AUC, ROC graph and score**

Train AUC score: 0.908

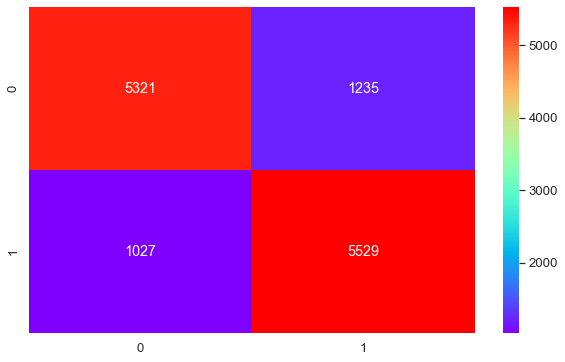
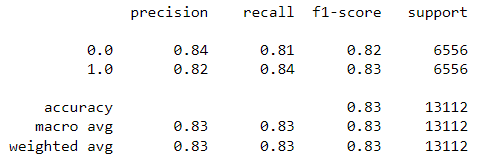
Test AUC score: 0.908

Train AUC curve Test AUC curve

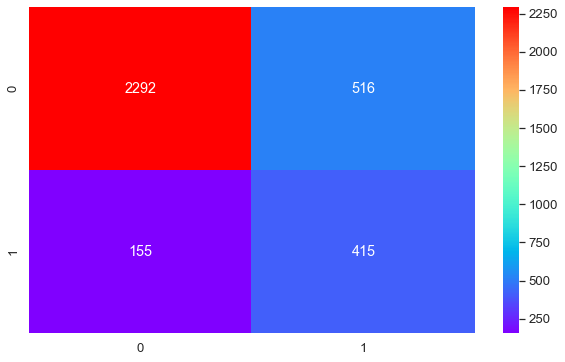
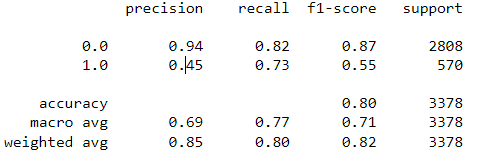
 

## Linear Discriminant Analysis

Train data set:



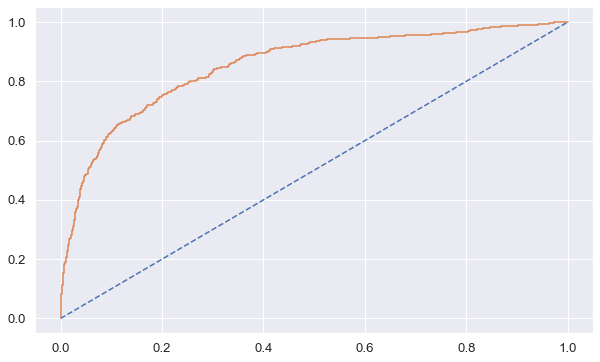
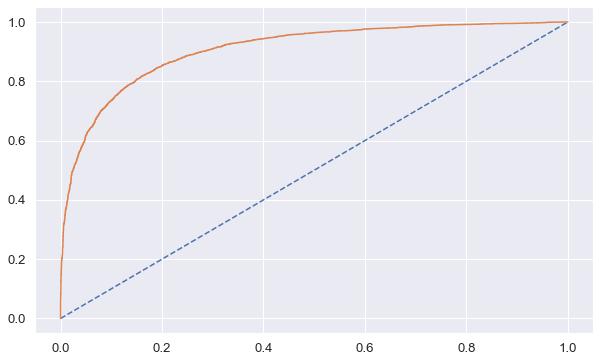
Test Data:



The difference in the performance matrices of the train and test data of more than 0.1 shows that the model is overfit.

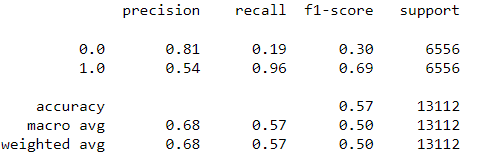
**AUC, ROC score and curve:**

Train AUC score: 0.907 Test AUC score: 0.907

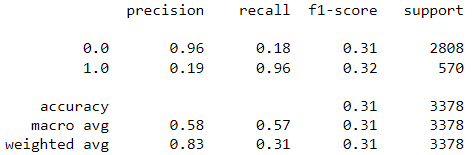
 

## Naïve Bayes Model

Train data set:

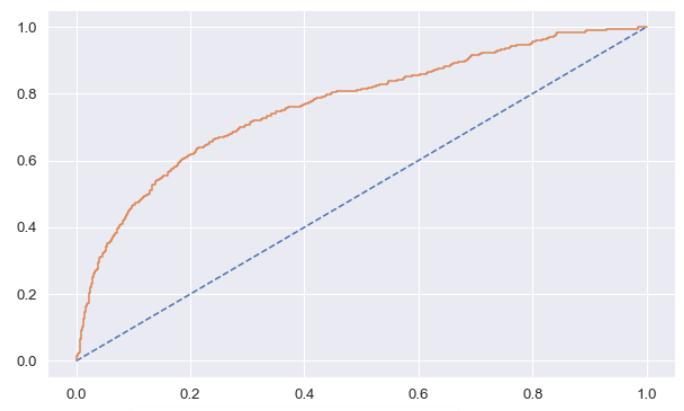
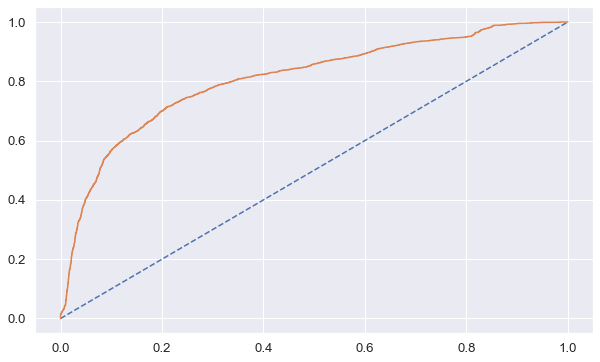
Test Data:

The model performs averagely in both the train and test models and hence we can skip it and try out other models with a better accuracy and recall.

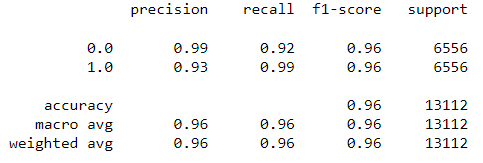
**AUC, ROC score and curve:**

Train AUC score: 0.998 Test AUC score: 0.998

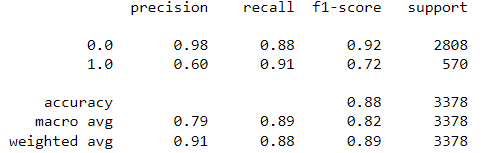
 

## KNN Model

Train data set:



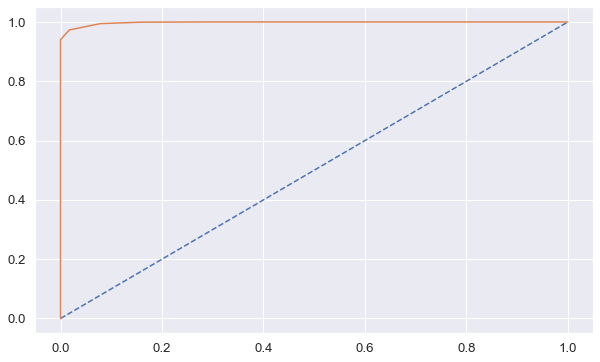
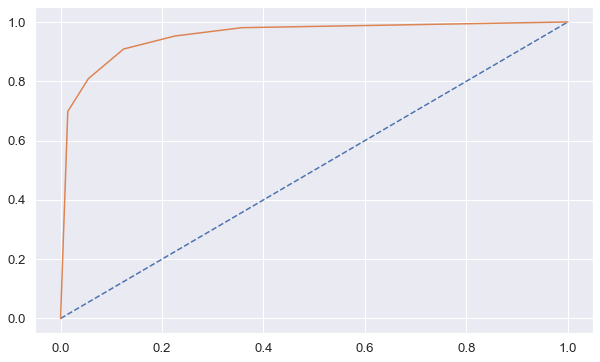
Test Data:



This is a model that has performed well in both the train and test sets.

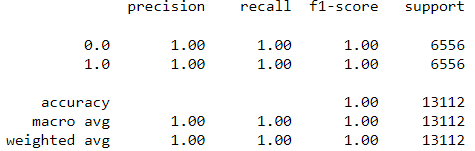
**AUC, ROC score and curve:**

Train AUC score: 0.998 Test AUC score: 0.998

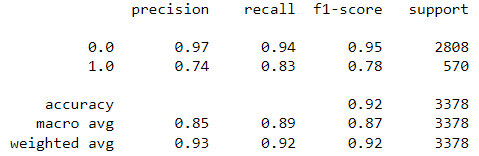


## Decision Tree Model

Train data set:

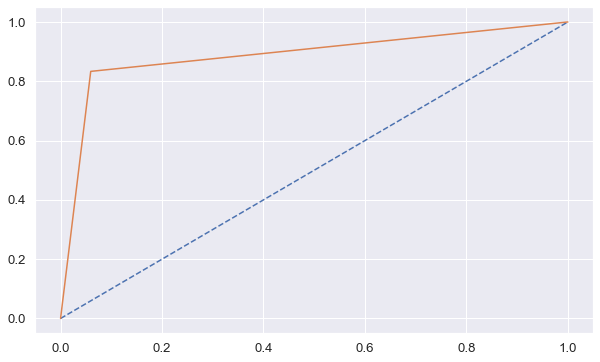
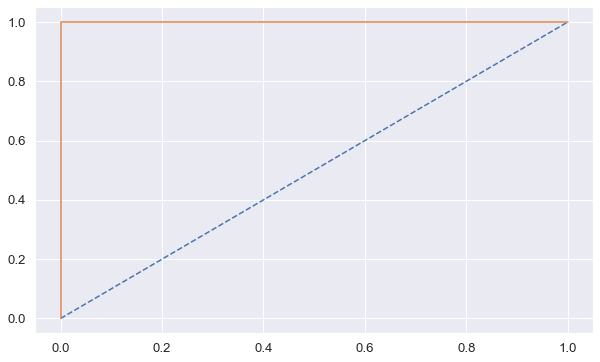
 

Test Data:

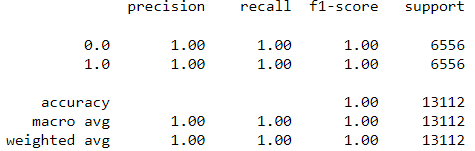
This is a model that has performed well in both the train and test sets.

**AUC, ROC score and curve:**

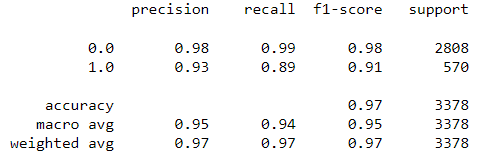
Train AUC score: 1.0 Test AUC score: 0.907 

## Random Forest

Train data set:

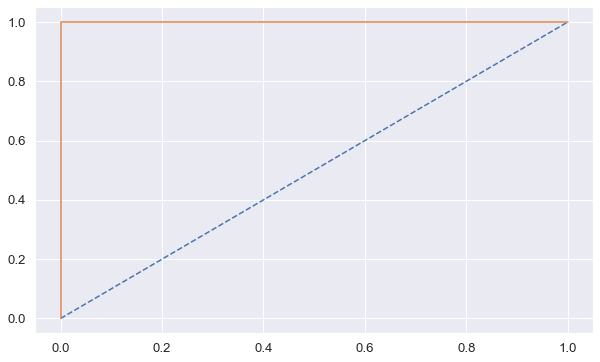
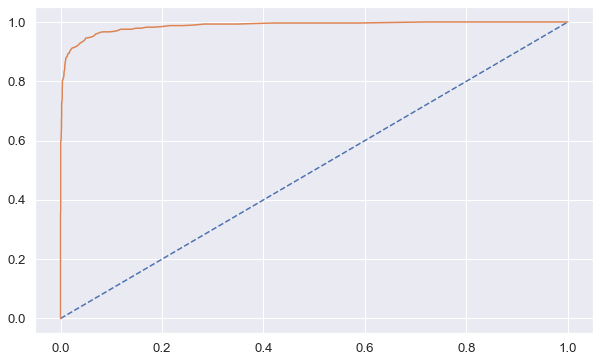
Test Data:

This is a model that has performed well in both the train and test sets.

**AUC, ROC score and curve:**

Train AUC score: 1.0 Test AUC score: 1.0

**2. Model Tuning**

Ensemble modeling is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications.

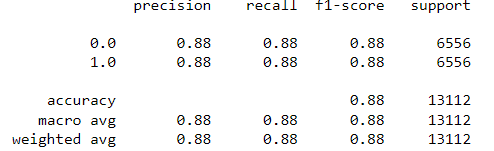
An ensemble can make better predictions and achieve better performance than any single contributing model.

We are gone use Bagging and boosting methods to tune the model and obtain the best performance.

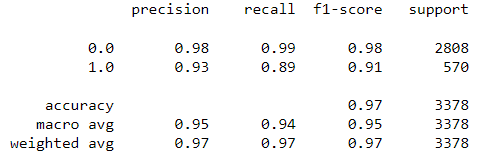
1. **ADA Boost**

We use the AdaBoostClassifier function from the sklearn.ensemble module to perform the same.

Train data set:

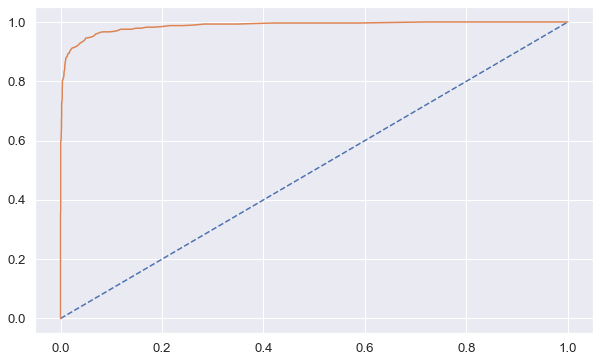
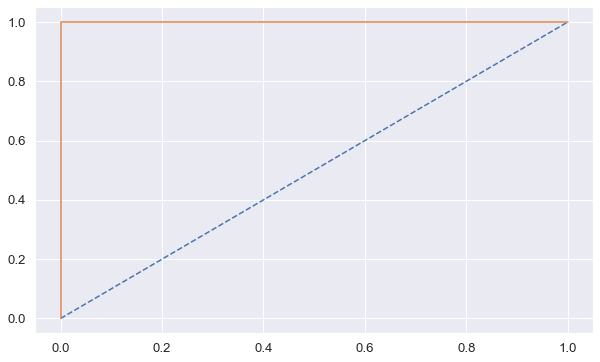
Test Data:

This is a model that has performed well in both the train and test sets.

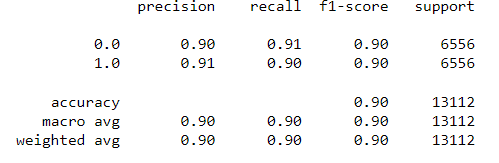
**AUC, ROC score and curve:**

Train AUC score: 1.0 Test AUC score: 1.0

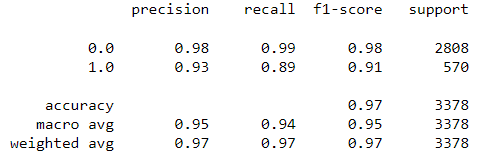


**Gradient Boosting**

Train data set:



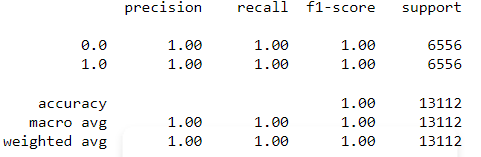
Test Data:



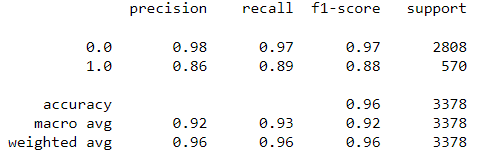
This is a model that has performed well in both the train and test sets.

**Bagging Model:**

Train data set:

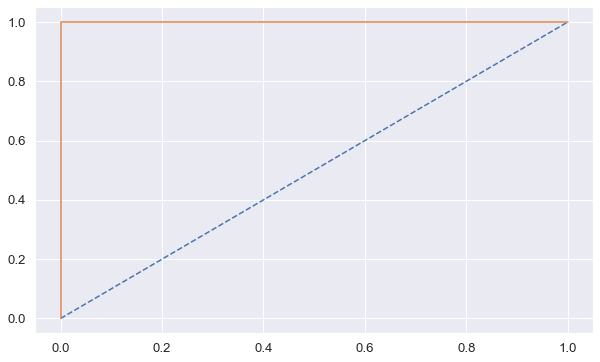
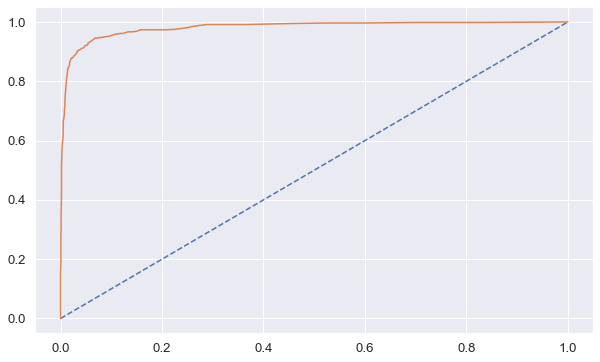
Test Data:



This is a model that has performed well in both the train and test sets.

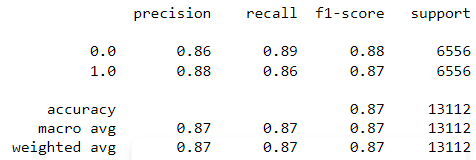
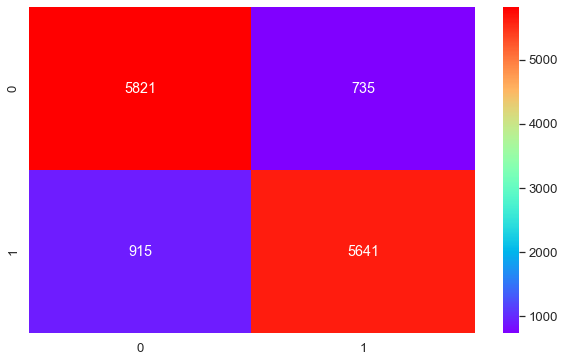
**AUC, ROC score and curve:**

Train AUC score: 1.0 Test AUC score: 1.0

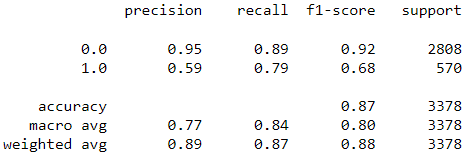
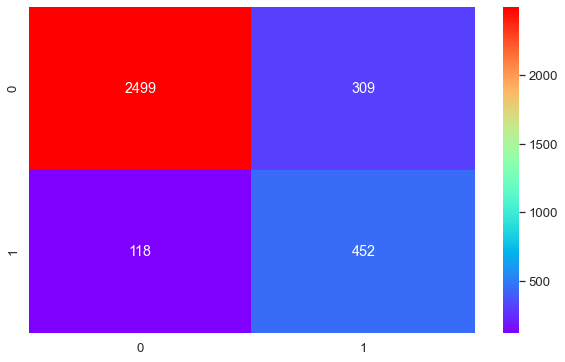
 

**XGB Classifier**

Train data set:

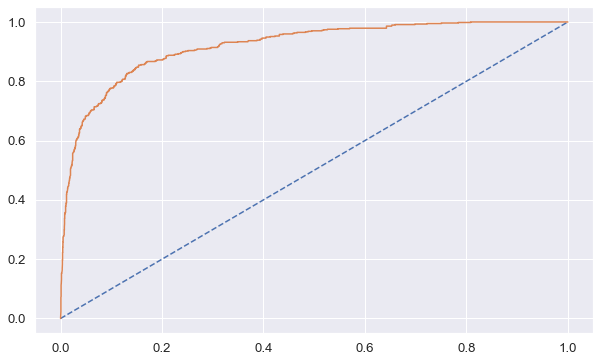
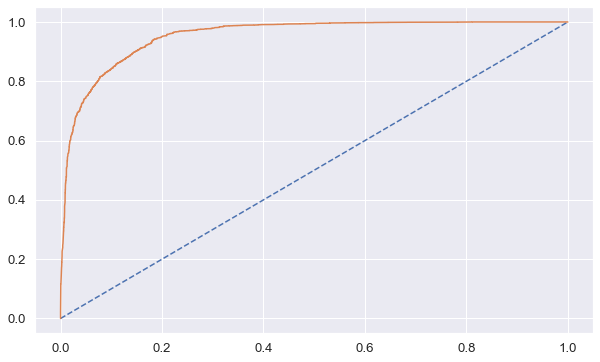
Test Data:

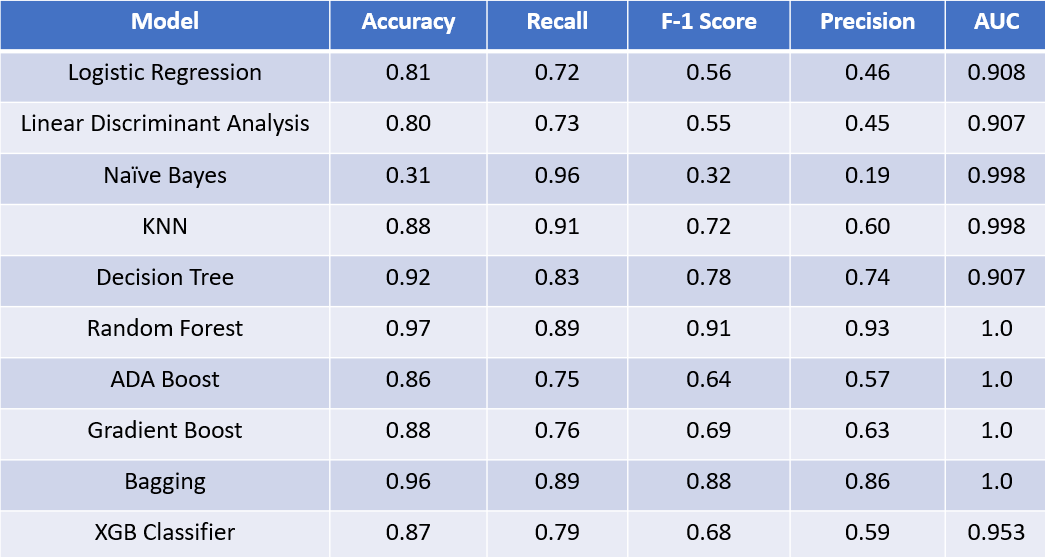
This is a model that has performed well in both the train and test sets.

**AUC, ROC score and curve:**

Train AUC score: 0.953 Test AUC score: 0.953



**Model Comparison Table:**



**Best Model and Model Validation:**

**The Random Forest model** is the one that has performed best in terms of accuracy, recall, and other scores in both the train and test models.

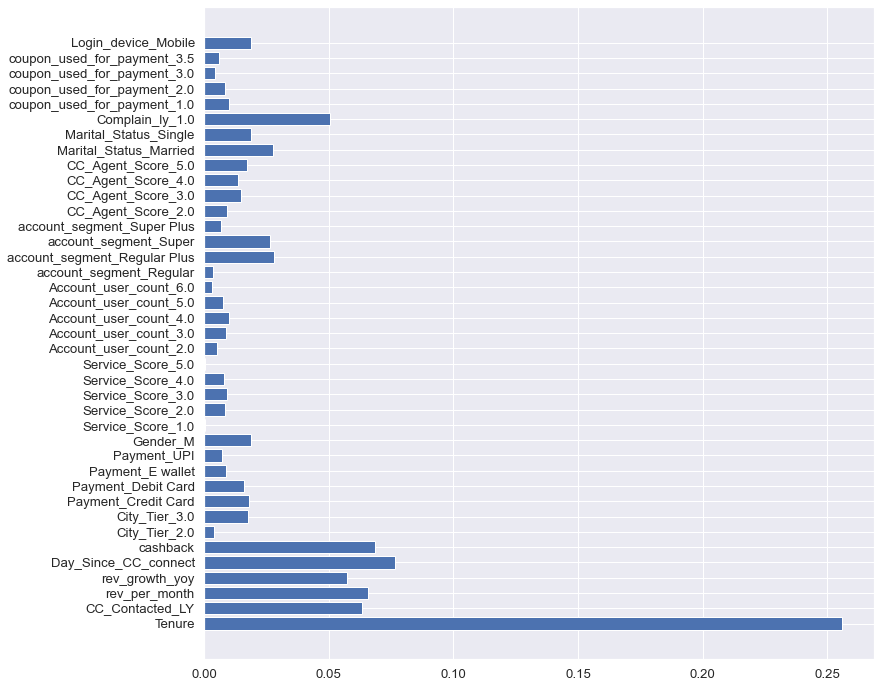
Since we are predicting whether a customer will churn or not, the most important parameters to consider are Recall and Accuracy.

The recall tells us of all the customers who churned, how many we were able to predict as churned correctly. In our case, the recall of 0.9 shows that our model is able to predict 90% of all churned customers correctly as churned. And with an accuracy of 0.97, overall, our model is able to correctly classify 97% of the customers on the basis of churn.

Upon selecting the RF model, we find out the best features i.e the features that contribute the most to the model. We find the Gini index for each feature to select the important ones.

Based on it, the best features:

* Tenure
* CC\_Contacted\_LY
* Day\_Since\_CC\_Connect
* Cashback
* Complain\_ly\_1.0
* Rev\_per\_month

****

## **Business insights**

Based on the Random Forest model and the analysis done, we can draw upon the following insights:

* Tenure is the most important feature that influences churn. The longer a customer has stayed with us, the less likely they will churn.
* Another feature is whether Customer care was contacted in the last year or not. If a customer has not reached out to CC in the last year, that probably means they are happy with the service and hence less likely that they will churn.
* Similarly, the longer it has been since customer care was contacted by a customer, the lesser probability of them churning.
* Cashback has an inverse correlation to churn. More the cashback that a customer has received, the lesser the probability of churn.
* There is a direct positive correlation between the revenue generated from the account per month and churn.
* Cash on delivery and e-wallet as the payment method has a higher probability of churn while customers using Credit cards are retained further.
* Based on the high correlation to churn, customers who have complained in the last 12 months, belong to the “Regular Plus” segment and are Single have a higher probability of churning compared to other groups. So, we need to focus on these groups to provide them with better customer service, offer discounts, and do all that we can to retain them.
* Accounts with a user count of 5 & 6 tend to churn more than those with 1 or 2 users.
* Customers who have not used a coupon for payment or used just 1 coupon churn more.

## **Business Recommendations:**

* Provide loyalty programs with better offers and higher discounts for customers who have been our customers for over 12 months.
* Give discounts ranging up to 25-40% if new users sign up for our annual membership with our Regular Plus, Super & Super Plus segments.
* Analyse the common problems raised by customers and resolve the root cause so that customers do not have to raise complaints or reach out to customer care frequently.
* Provide incentives for customers to use credit cards instead of COD or e-wallet in the form of cashback.
* Increase the rate of coupons being provided. Eg: Instead of providing a coupon for Rs.100, we can instead provide 5 coupons for Rs.20 each. The more coupons available will promote the usage of the same.
* Incentivize moving users from Regular & Regular plus segments to the Super & Super plus segments by giving a 25% discount if they upgrade and sign up for a 12-month membership along with a 6-month extended membership.
* We could alternatively also provide some of the features available in the super/super plus segments to the regular/regular plus segment on a trial basis for 1 month to promote them to upgrade.