Customer Churn Analysis Report

(E-Commerce)

*Submitted by:*

*Rahul Kumar*

*Under guidance of: Abhay Poddar*

*Batch – PG-DSBA Sep19*

Contents

[**Final Report Part – I (EDA & Pre-processing)** 4](#_Toc54519091)

[1 Introduction 4](#_Toc54519092)

[1.1 Problem Statement 4](#_Toc54519093)

[1.2 Need of the Project 4](#_Toc54519094)

[1.3 Understanding business/social opportunity 4](#_Toc54519095)

[2 Data Report 5](#_Toc54519096)

[2.1 Understanding how data was collected in terms of time, frequency and methodology 5](#_Toc54519097)

[2.2 Visual inspection of data 5](#_Toc54519098)

[2.3 Understanding attributes 5](#_Toc54519099)

[3 EDA (Exploratory Data Analysis) & Business Implications 6](#_Toc54519100)

[3.1 Non-Visual Analysis of data 6](#_Toc54519101)

[3.2 Univariate analysis 6](#_Toc54519102)

[3.3 Bivariate Analysis 13](#_Toc54519103)

[4 Data Cleaning and re-processing 16](#_Toc54519104)

[4.1 Removal of unwanted variable 16](#_Toc54519105)

[4.2 Outlier Treatment 16](#_Toc54519106)

[4.3 Variable Transformation 17](#_Toc54519107)

[4.4 Addition of new variable (Feature Engineering) 18](#_Toc54519108)

[5 Business Insights from EDA 18](#_Toc54519109)

[5.1 Is the data unbalanced? If so, what can be done? Please explain in the context of the business 18](#_Toc54519110)

[5.2 Clustering Business Insights: 19](#_Toc54519111)

[5.3 Other business insights: 19](#_Toc54519112)

[6 Appendix (PN – I) 21](#_Toc54519113)

[**Final Report Part– II (Model building, Model validation and Recommendation)** 22](#_Toc54519114)

[1 Model building and interpretation. 22](#_Toc54519115)

[1.1 Which models to choose and why? & Building Base Models 23](#_Toc54519116)

[1.1.1 Linear Discriminant Analysis (LDA) 23](#_Toc54519117)

[1.1.2 Logistic Regression 25](#_Toc54519118)

[1.1.3 KNN 27](#_Toc54519119)

[1.1.4 Naïve Bayes 30](#_Toc54519120)

[1.1.5 SVM 33](#_Toc54519121)

[2 Model Tuning and Optimization 35](#_Toc54519122)

[2.1 Ensembling Technique 35](#_Toc54519123)

[2.1.1 Bagging (Random Forest) 36](#_Toc54519124)

[2.1.2 Boosting (XGBoost) 37](#_Toc54519125)

[2.2 Hyperparameter Tuning 40](#_Toc54519126)

[2.2.1 Tuning LDA model 40](#_Toc54519127)

[2.2.2 Tuning Logistic model 41](#_Toc54519128)

[2.2.3 Tuning KNN model 43](#_Toc54519129)

[2.2.4 Tuning Naïve Bayes model 44](#_Toc54519130)

[2.2.5 Tuning SVM model 46](#_Toc54519131)

[2.3 SMOTE 47](#_Toc54519132)

[2.3.1 Logistic model with SMOTE 47](#_Toc54519133)

[2.3.2 LDA model with SMOTE 48](#_Toc54519134)

[2.3.3 KNN model with SMOTE 50](#_Toc54519135)

[2.3.4 Naïve Bayes model with SMOTE 51](#_Toc54519136)

[2.3.5 SVM model with SMOTE 52](#_Toc54519137)

[3 Model Validations, Comparison & Selecting Final Model 53](#_Toc54519138)

[4 Business Insights and Recommendations 57](#_Toc54519139)

[4.1 Model Recommendation 57](#_Toc54519140)

[4.2 Business Recommendations: 57](#_Toc54519141)

[4.2.1 Business Recommendations for Churning customers 57](#_Toc54519142)

[4.2.2 Business Recommendations for Non-churning customers 58](#_Toc54519143)

[5 Appendix (PN-II) 59](#_Toc54519144)

# **Final Report Part – I (EDA & Pre-processing)**

# 1 Introduction

A wide customer base is the root foundation for any profit-making business. Importance of retaining customer as well as prior information of potential churn of the customers are the key to plan business strategies and being ahead of the curve in dynamic market.

## 1.1 Problem Statement

A leading online retail (E commerce) company has a business requirement to plan strategy for offering promos. For a data-driven approach to implement the business plan for promos, there is a requirement by the company for a detailed churn analysis of the customers.

## 1.2 Need of the Project

For this Project, we are required to provide prediction for customer churn, using the different provided feature related to the customers, performing pre-requisite analysis and processes, building a well optimized Machine Learning models. As a part of Project Notes – 1, we will particularly keep our focus on the following :

* Data:
* Understanding how data was collected in terms of time, frequency and methodology
* Visual inspection of data
* Understanding of attributes
* **Exploratory data analysis:**
* Univariate analysis
* Bivariate analysis
* Removal of unwanted variables
* Missing Value treatment
* Outlier treatment
* Variable transformation
* Addition of new variables
* **Business insights from EDA**
* Is the data unbalanced? If so, what can be done? Please explain in the context of the business
* Any business insights using clustering
* Any other business insights

## 1.3 Understanding business/social opportunity

Understanding customer churn is very important factor for Online business. Indeed, knowing potential churn of customer base gives an edge over the competitors.

This can be a golden opportunity to take the business to next level by re-forming business strategy on a more granular level based on churn flag of the customer.

Moreover, offering special discounts for customers who will potentially churn in near future is a business loss, this issue can also be addressed by Churn Predictions.

# 2 Data Report

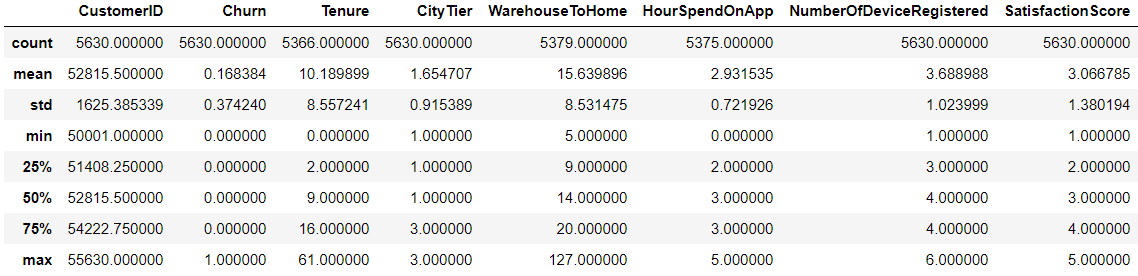
## 2.1 Understanding how data was collected in terms of time, frequency and methodology

The data for the project is sourced from a leading E Commerce retail company, provided for a specific period of time for customers of the company.

## 2.2 Visual inspection of data

There are 20 features and 5630 records in the given data set.

Below are the descriptive details:



Insights:

* The data needs scaling, as different features different mean, median and max. values
* Record count of features are also different, indicating the missing values in the data
* There are Outliers in the data, indicated by mean and median values in the above details. For instance, we have mean tenure of 10 whereas having max. value for tenure as 61, indicating outliers in the data.

## 2.3 Understanding attributes

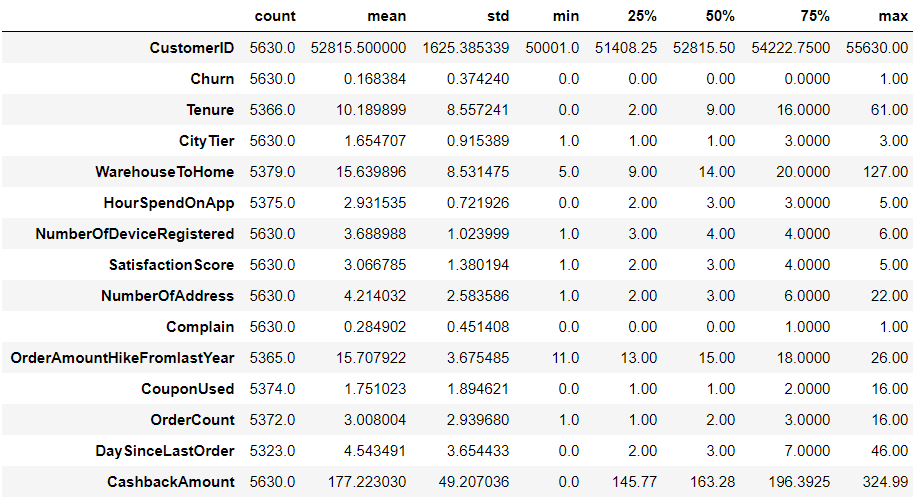
Below is the info for the variables used in the data:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| CustomerID | Unique customer ID |
| Churn | Churn Flag |
| Tenure | Tenure of customer in organization |
| PreferredLoginDevice | Preferred login device of customer |
| CityTier | City tier |
| WarehouseToHome | Distance in between warehouse to home of customer |
| PreferredPaymentMode | Preferred payment method of customer |
| Gender | Gender of customer |
| HourSpendOnApp | Number of hours spend on mobile application or website |
| NumberOfDeviceRegistered | Total number of deceives is registered on particular customer |
| PreferedOrderCat | Preferred order category of customer in last month |
| SatisfactionScore | Satisfactory score of customer on service |
| MaritalStatus | Marital status of customer |
| NumberOfAddress | Total number of added added on particular customer |
| Complain | Any complaint has been raised in last month |
| OrderAmountHikeFromlastYear | Percentage increases in order from last year |
| CouponUsed | Total number of coupon has been used in last month |
| OrderCount | Total number of orders has been places in last month |
| DaySinceLastOrder | Day Since last order by customer |
| CashbackAmount | Average cashback in last month |

* Renaming of the variables is not required in this case (There seems no space issue in the variable name).

# 3 EDA (Exploratory Data Analysis) & Business Implications

## 3.1 Non-Visual Analysis of data



**Insights:**

* **There are missing values in the data**, as max. record count is 5630 but features like ‘Tenure’, ‘Warehouse to Home’, ‘HoursSpendOnApp’ and few others have count of less than 5630.
* **Mean, median and max. values of all different for all features** and varying in range, thus data is not scaled and **scaling is required.**
* There is much difference between mean and median values for few continuous features. For example, ‘Day since last order’ has mean of 4.54 and median of 3.0, thus indicating it is right skewed. Thus, **data is skewed**
* Skewness in data indicates **potential presence of Outliers** in the data

## 3.2 Univariate analysis

**Categorical features**

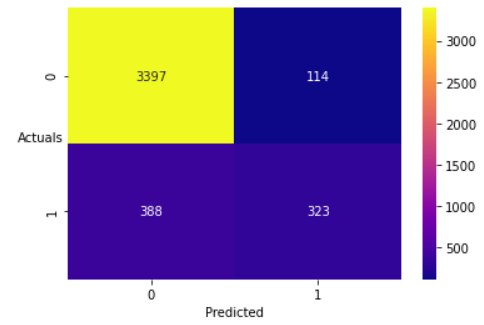
Below is the distribution of all the categorical (including Ordinal Categorical) features:

Dependent Feature (Churn)



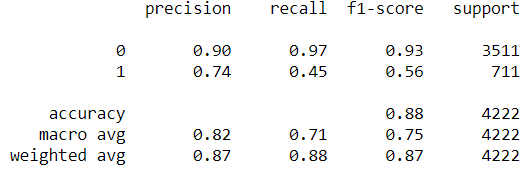
Insight:

There are 948 customer that churn out of all the customers. Over 4682 customers will be retained.



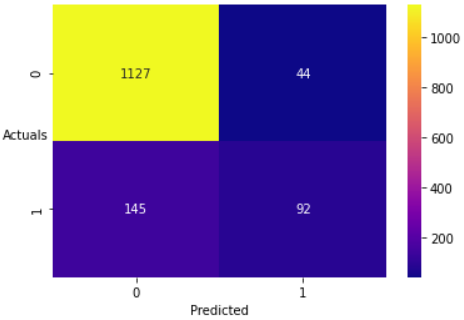
Insights:

* Most of the customer belongs to City Tier 1.
* Second highest count of customers is in City Tier 3
* Lowest customers are there in City Tier 2



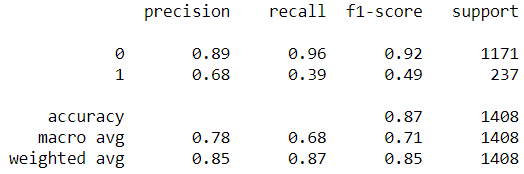
Insights:

* Most of the customers are Male and there are only 2246 females.
* This data is very essential to decide what type of products can be made more available on the online store. For example: While restocking the products in warehouse, more focus can be given to products used by male. Again, Order count also needs to be taken under consideration while deciding that.



Insights:

* Most number of Customers are married.
* This again is a crucial insight, as this indicated the approx. age group, which can be around 28 years and above.
* This age criteria can be utilised in deciding products to offer.



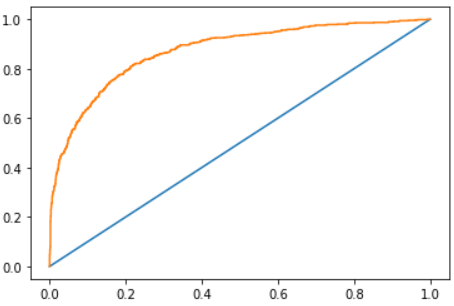
Insights:

* Mobile Phone is the most preferred category, followed by Laptop & Accessary, Fashion, Grocery and others.
* This indicates Electronics products are most preferred among the customers, followed by cloths and fashion products.
* These can be focussed to plan discount strategy.



Insights:

* Most preferred login device is Mobile, followed by Computer.
* More focus can be given to enhance user experience over the app.



Insights:

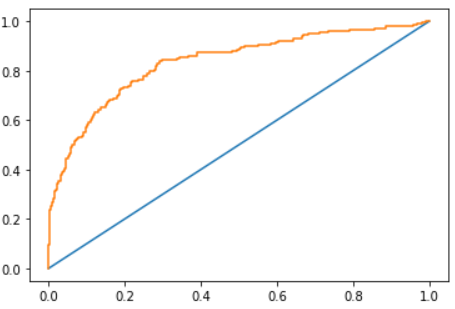
* Most preferred payment method is Debit Card, followed be Credit Card, E wallet, COD, UPI, CC.
* We can provide cashback on using most preferred payment method, i.e. Debit Card and Credit Card, this will potentially increase order count.



Insights:

* Most of the customers give a rating of 3 out of 5, followed by rating 1, rating 5, rating 4 and rating 2.
* This insight brings out that quit a few customers gave rating 1.
* Thus, focus can be made on better customer care and feedback and further implementing required plan.

**Continuous Features**



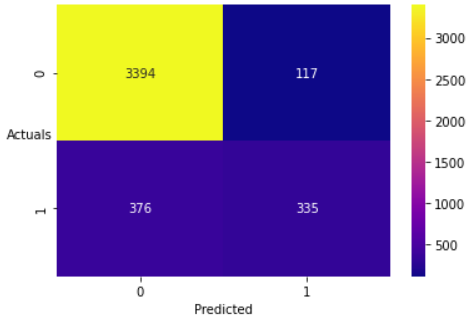
Insights:

* Tenure distribution is right skewed.
* Highly right skewed distribution indicates potential outliers.



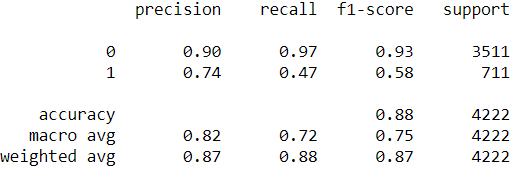
Insights:

* Most of the customers spent 3 hours on app, followed by 2hrs and 4 hrs.
* The distribution indicated multimodal distribution.



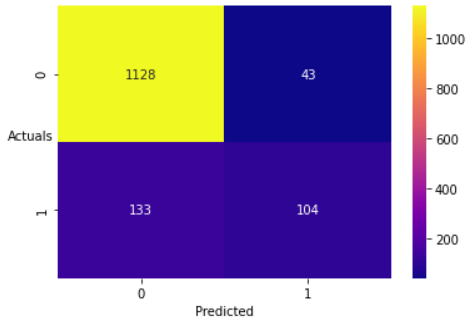
Insights:

* The distribution is multimodal left skewed.
* Most observed number of devices registered are 4, followed by 3, 5, 2, 1 and 6.
* This indicates how many devices or potentially different person are logging in with same user id.



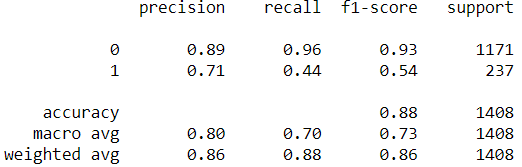
Insights:

* The distribution is right skewed, with maximum number of address used for order delivery is 2.
* This indicates potential orders by the person registered for other person as a gift (like family member and friends).
* This also indicates indirect satisfaction and trust of the customer.



Insights:

* The distribution indicates the most of the customers have raised zero complain.
* Focus can be made on reducing the complain counts.



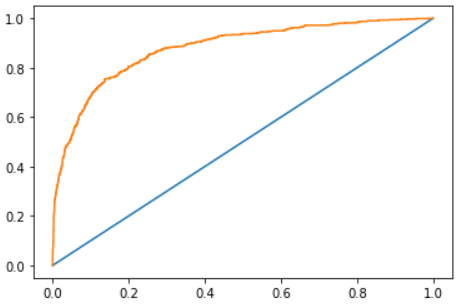
Insights:

* The distribution is right skewed, indicating potential outliers.
* The mode of the order hike is approx. 13%.



Insights:

* The distribution is right skewed, indicating potential outliers.
* The mode of number of coupon last month used is approx. 1



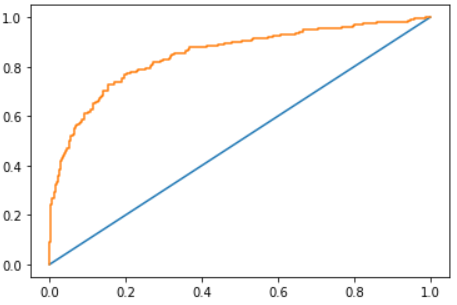
Insights:

* The distribution is right skewed, indicating potential outliers.
* The mode of the order count in last month is approx. 2.



Insights:

* The distribution is right skewed, indicating potential outliers.
* The mode of the order hike is approx. 3
* The lower the value better is the recency value for the customer.



Insights:

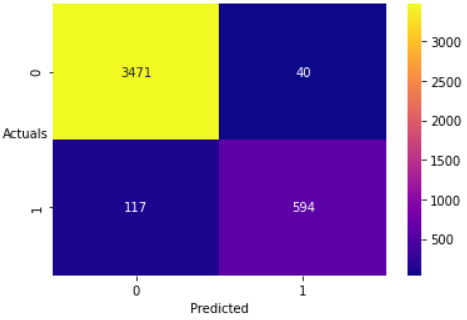
* The distribution is right skewed, indicating potential outliers.
* The mode of the Cash back amount is approx. 160.

## 3.3 Bivariate Analysis

**Correlation:**

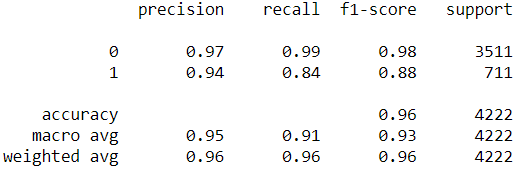
Insights

* Most of the variable are not highly correlated.
* Coupon used and Order count are highly correlated with correlation score of 0.75, which is considerably high. This may cause high multicollinearity.
* To address multicollinearity, we drop ‘CouponUsed’.



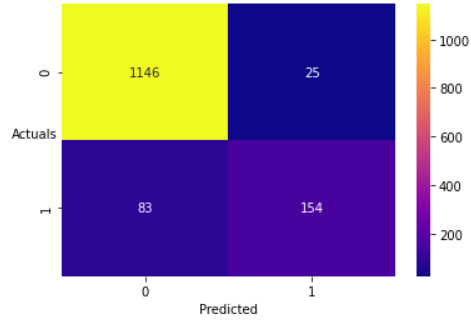
Insights:

* Highest sum of Order amount hike is for City Tier 1 customers, followed by City Tier 3 and 2.



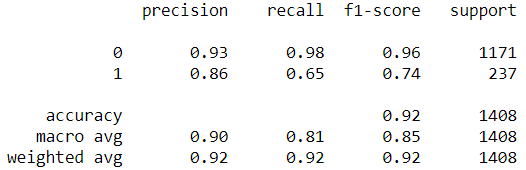
Insights:

* Highest sum of order count is for Laptop & Accessary, which is 35.09%
* Second highest order count sum is for Mobile phones 28.10%



Insights:

* City Tier 1 is ranked 1 for order recency, followed by City Tier 2 and City Tier 3.



Insights:

* Mobile phone order is best in recency and ranked 1, followed by Laptop & Accessary with ranked 2.
* Others category of products are worst in recency for order.

# 4 Data Cleaning and re-processing

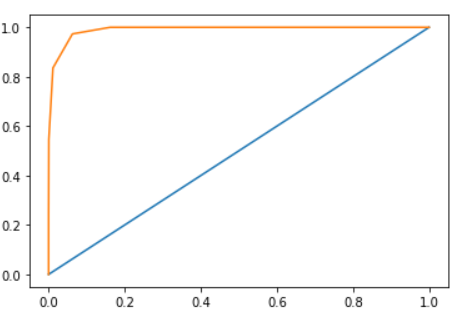
## 4.1 Removal of unwanted variable

* ‘CustomerID’ has been removed from the data. This variable is of no significance and will act as a noise to the Machine Learning models.

## 4.2 Outlier Treatment

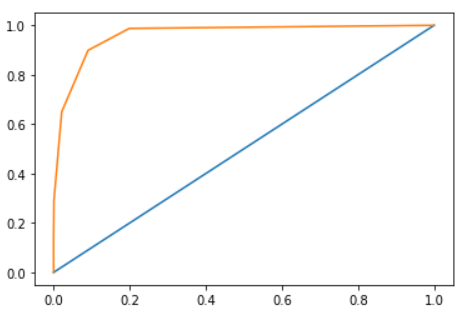
Below are the boxplots to identify presence of outliers using IQR:





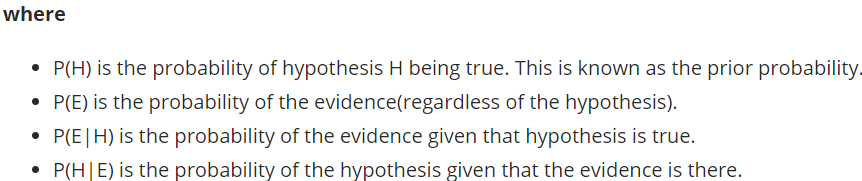
* Above boxplots indicates the presence of outliers.
* We will perform quantile-based outlier capping, because there seems to be quite a few outlier so removing outlier will not be a suitable option
* Below are boxplots post capping the outliers:





## 4.3 Variable Transformation

* There are variable with ‘object’ as data type. These string values are not permissible by the Machine Learning models. Thus, we must transform or encode the string values.
* We have encoded the object type features using Label encoding as technique. Using One-Hot encoding would have created too many variables.
* Below is the detail of encoded variables:



## 4.4 Addition of new variable (Feature Engineering)

* We have formulated new variable - ‘Recency’ using variable ‘DaySinceLastOrder’.
* We have formulated the variable as per below:
* Recency = 100 – Days since last order
* 100 is chosen here as constant number and as a numeric value greater than max. of ‘DaySinceLastOrder’ (which is 14.5)

# 5 Business Insights from EDA

## 5.1 Is the data unbalanced? If so, what can be done? Please explain in the context of the business

* Yes, the data is imbalanced as the target variable has 83.17 data point with Churn flag 0 and 16.84% of data point has Churn flag as 1.
* Thus, there is imbalanced in the proportion of the classes of the target feature.
* This may result in Machine Learning model getting to learn less patterns in the data with Churn flag 1 as the amount of data available for the same is very less compared to other target class.
* We have technique of SMOTE to address the class imbalanced.

## 5.2 Clustering Business Insights:

* There are 3 clusters formed using Hierarchical clustering, which is preferred due to small data set.
* Below are the details for 3 clusters:



* Different promotional offers can be made based on clusters shown above.
* Cluster 1 has highest tenure but low recency, thus they can be provided premium membership as a old customer of the company.
* Order count is also highest for cluster 1 but a bit low for cluster 2 and cluster 3, thus a specific offer can be provided based on individual customer order history. This will improve the order counts for cluster 2 and cluster 3
* There can be 3 membership provided in the following patter:

NOTE : [There will be three membership, Golden, Silver and bronze with Golden being highest class of membership and Bronze being lowest]

* **Golden Membership**: This can be provided for cluster 1 customers, as they stayed with company for longest, with highest order count.
* **Silver Membership:** This can be provided to cluster 2 customers as they second highest in order count and stayed most long with company after cluster 1 customers
* **Bronze Membership**: This can be offered to Cluster 3 customers to get them more attracted to the company and get the order count high.

## 5.3 Other business insights:

Below are the business insights:

* Most of the customer belongs to City Tier 1.
* Second highest count of customers is in City Tier 3
* Lowest customers are there in City Tier 2
* Thus, more focus needs to be done for City Tier 2, more advertising need to be implemented in City Tier 2.

Based on City Tier:

* Most number of Customers are married.
* This again is a crucial insight, as this indicated the approx. age group, which can be around 28 years and above.
* This age criteria can be utilised in deciding products to offer. For example, household products and electronics products can be focussed

Based on product preference:

* Mobile Phone is the most preferred category, followed by Laptop & Accessary, Fashion, Grocery and others.
* This indicates Electronics products are most preferred among the customers, followed by cloths and fashion products.
* These can be focussed to selected promotional offer for these products.

Based on preferred login device:

* Most preferred login device is Mobile, followed by Computer.
* More focus can be given to enhance end user experience over the app.

Based on preferred payment method:

* Most preferred payment method is Debit Card, followed by Credit Card, E wallet, COD, UPI, CC.
* We can provide cashback on a specific company’s Debit Card or Credit Card; this will potentially increase order count.

Based on customer ratings:

* Most of the customers give a rating of 3 out of 5, followed by rating 1, rating 5, rating 4 and rating 2.
* This insight brings out that quit a few customers gave rating 1.
* Thus, focus can be made on better customer care and support.
* A good survey among the customer would provide a useful feedback.

Based on Order Count for product categories:

* Highest sum of order count is for Laptop & Accessary, which is 35.09%
* Second highest order count sum is for Mobile phones 28.10%.
* This info shows mostly customers prefer the Online store for electronics product. Thus the E Commerce should expand the customer relationship for other product areas like Groceries etc.

Based on Recency of order:

* City Tier 1 is ranked 1 for order recency, followed by City Tier 2 and City Tier 3.

# 6 Appendix (PN – I)

* Tool Used: Python, Tableau
* Data Dictionary:



* Tableau Visualization:

**Error! Not a valid link.**

* Python Code:



# **Final Report Part– II (Model building, Model validation and Recommendation)**

# **1 Model building and interpretation.**

Getting a best optimized ML model is an iterative process, which involves building and testing various algorithms and fine tuning them to optimize it to its best. Therefore, have built different classification models, performed Hyperparameter tuning, involve Ensemble technique to check for improved performance metrics and finally concluding the best model for the given case. Below is the workflow:

NOTE: \*Please note that the model comparison based on performance metrics is discussed in ‘Model Comparison and Final Model selected’ topics (section: 3 of PN-II)

## 1.1 Which models to choose and why? & Building Base Models

**Which model to choose and Why?**

* We have used LDA, Logistic, KNN, Naïve Bayes, SVM, Random forest and XGboost as classification models

**Building base models:**

* Firstly, we have split the data in ratio 75:25 for training and testing respectively.
* Scaling was performed on data via Standard Scalar, as the features are on different scale.
* Below are the various base models built

### Linear Discriminant Analysis (LDA)

LDA is a Supervised Learning algorithm which is capable for classification problems even with more than two level of classes in the target variable. Two criteria are used by LDA to create a new axis:

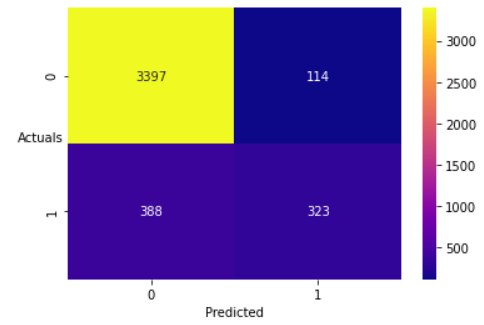
1. Maximize the distance between means of the two classes.
2. Minimize the variation within each class.

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Train data:

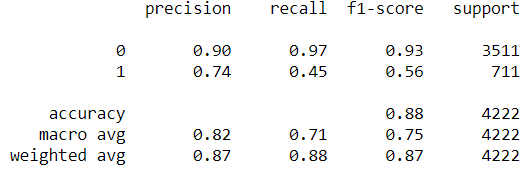


TN = 3397

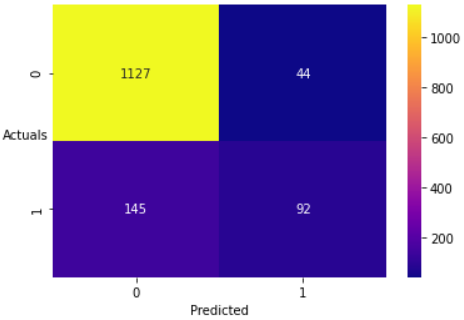
TP = 323

FP = 114

FN = 388



* Confusion matrix and Classification report for Test data:

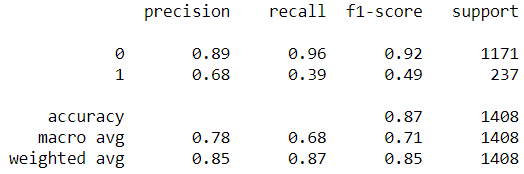


TN = 1127

TP = 92

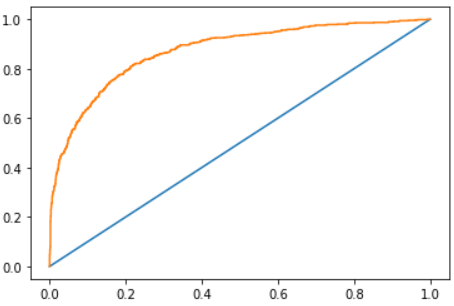
FP = 44

FN = 145



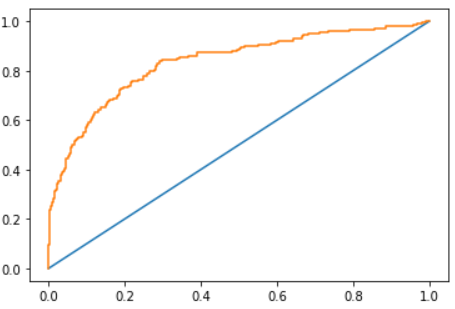
* AUC and ROC curve for Train data:





* AUC and ROC curve for Test data:





### Logistic Regression

Logistic regression is one of the most popular Machine learning algorithms that comes under Supervised Learning techniques.

It can be used for Classification as well as for Regression problems, but mainly used for Classification problems. Logistic regression is based on the concept of **Maximum Likelihood estimation**. According to this estimation, the observed data should be most probable.

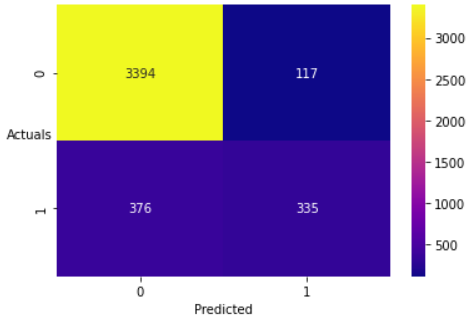
In logistic regression, we pass the weighted sum of inputs through an activation function that can map values in between 0 and 1. Such activation function is known as **sigmoid function** and the curve obtained is called as sigmoid curve or S-curve.

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Train data:

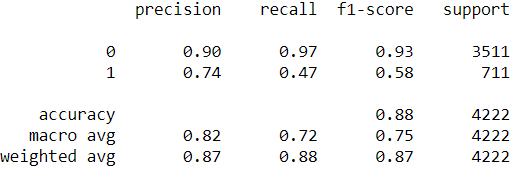


TN = 3394

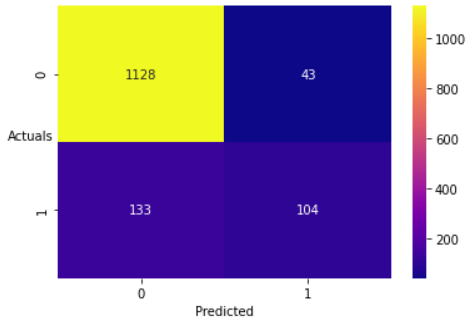
TP = 335

FP = 117

FN = 376



* Confusion matrix and Classification report for Test data:

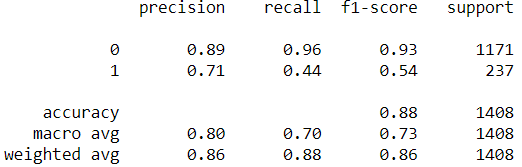


TN = 1128

TP = 104

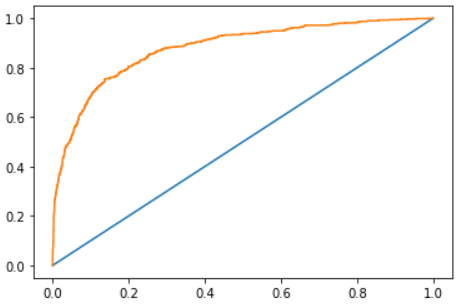
FP = 43

FN = 133



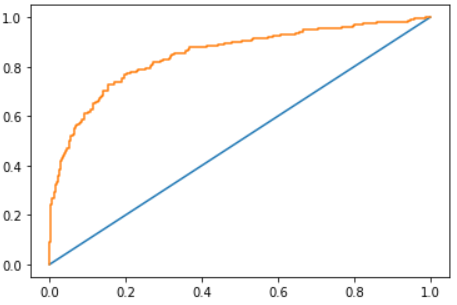
* AUC and ROC curve for Train data:





* AUC and ROC curve for Test data:





### KNN

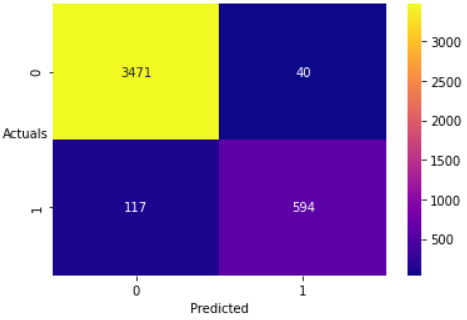
 k-nearest neighbours algorithm (k-NN) is a Non-Parametric method, and a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. KNN works on the concept that things exist in proximity. In other words, similar things are near to each other. For this algorithm, we can only have 2 classes in target (0 and 1).

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Train data:

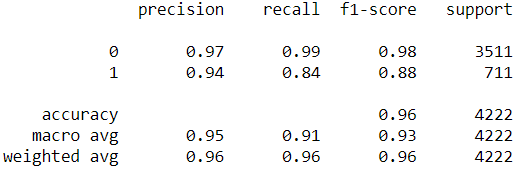


TN = 3471

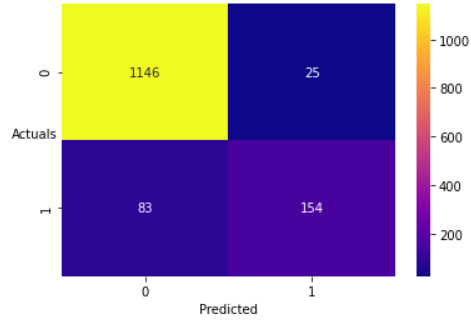
TP = 594

FP = 40

FN = 117



* Confusion matrix and Classification report for Test data:

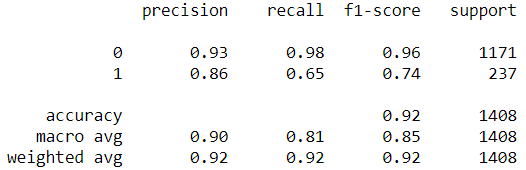


TN = 1146

TP = 154

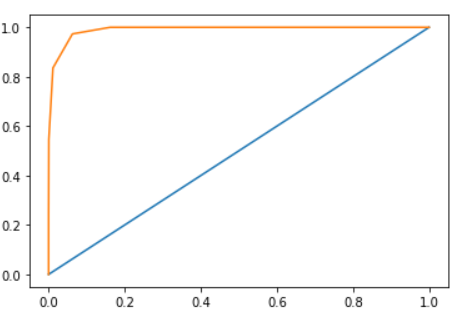
FP = 25

FN = 83



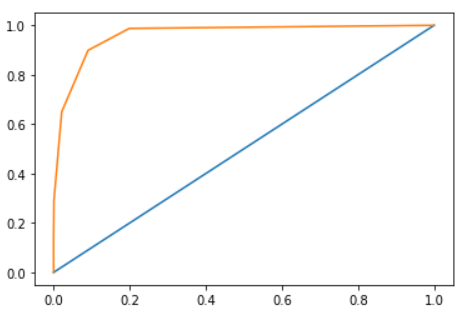
* AUC and ROC curve for Train data:





* AUC and ROC curve for Test data:



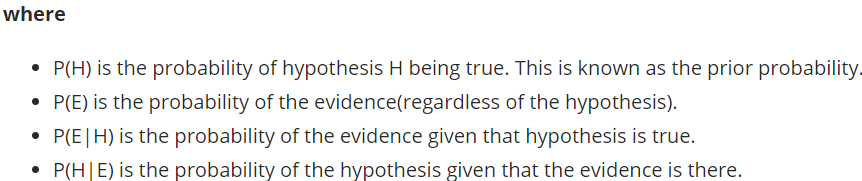


### Naïve Bayes

Naïve bayes is a classification algorithm that works on the basis on Bayes theorem of conditional probability.

As per Bayes theorem:

P(H|E) = P(E|H) \* P(H)/ P(E)

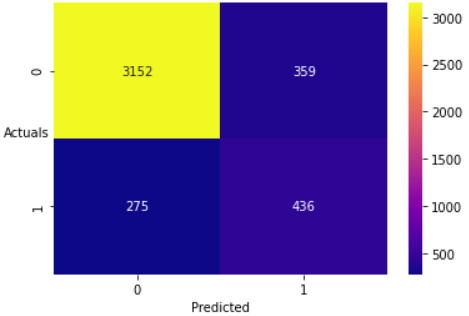


**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Train data:

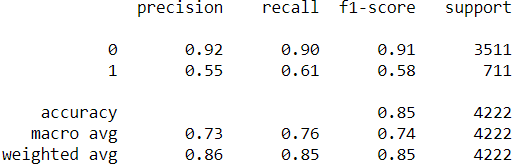


TN = 3152

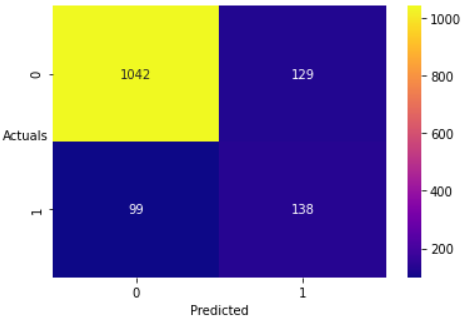
TP = 436

FP = 359

FN = 275



* Confusion matrix and Classification report for Test data:

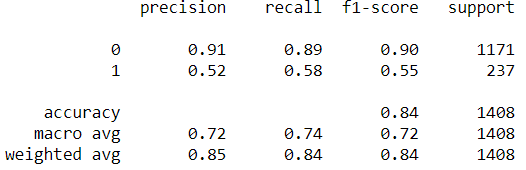


TN = 1042

TP = 138

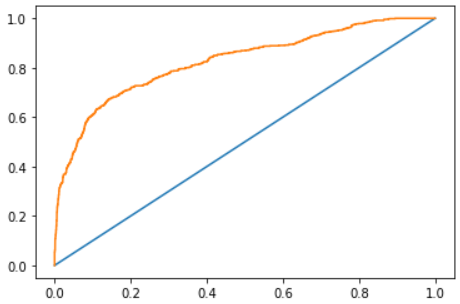
FP = 129

FN = 99



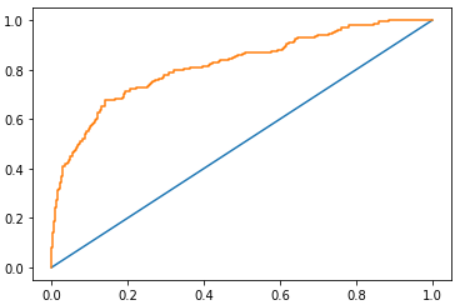
* AUC and ROC curve for Train data:





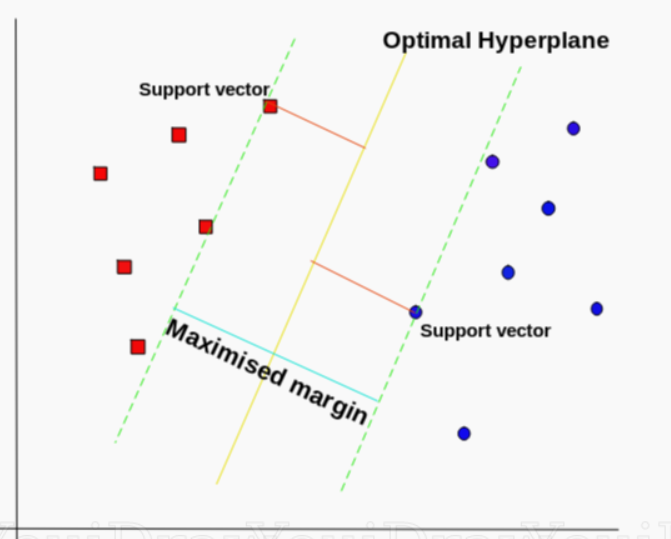
* AUC and ROC curve for Test data:





### 1.1.5 SVM

SVM is a powerful classification algorithm and the idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. According to the SVM algorithm we find the points closest to the line from both the classes. These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called the margin. Our goal is to maximize the margin. The hyperplane for which the margin is maximum is the optimal hyperplane.

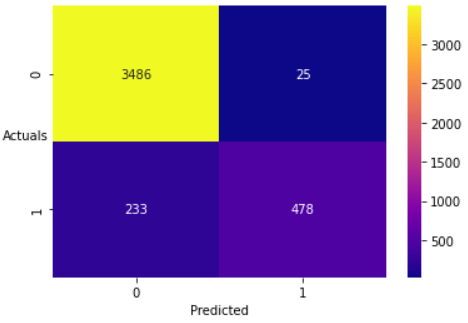


**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Train data:

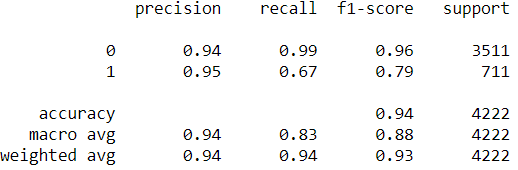


TN = 3486

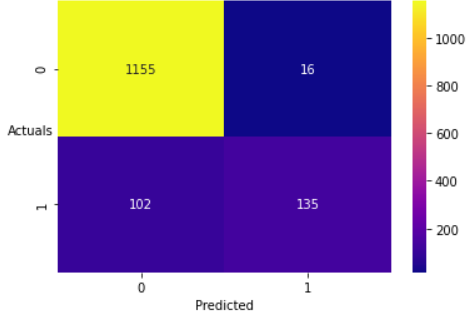
TP = 478

FP = 25

FN = 233



* Confusion matrix and Classification report for Test data:

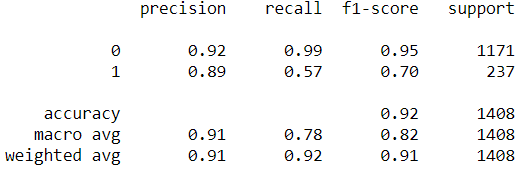


TN = 1155

TP = 135

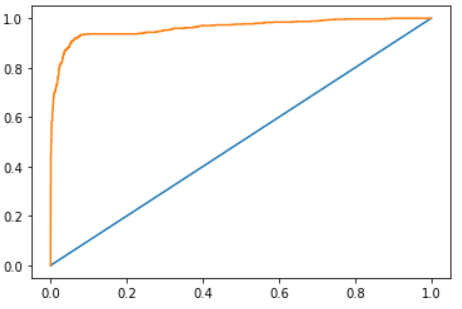
FP = 16

FN = 102



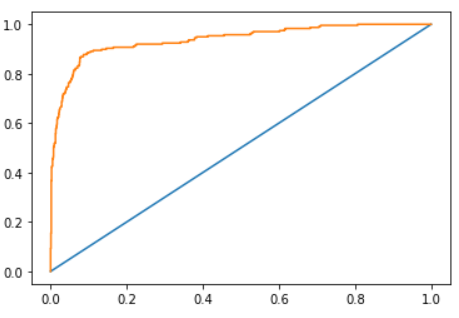
* AUC and ROC curve for Train data:





* AUC and ROC curve for Test data:





# Model Tuning and Optimization

We will perform Model tuning to enhance performance metrics of the base model. Below technique or process will be applied as Model tuning:

* Ensembling technique (Bagging and Boosting)
* Hyper-parameter tuning
* SMOTE (for balancing target classes)

## Ensembling Technique

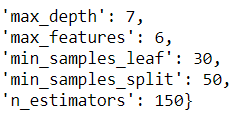
### Bagging (Random Forest)

Bagging or Bootstrap aggregating is an ensemble technique.

* Building Base model of Random Forest, we have the following model accuracies:



* Above detail clearly indicates the case of Overfitting as Train data completely fit the model with 100% accuracy.
* Thus, this base model cannot be considered. We will try pruning the trees via hyper-parameter tuning its attributes.
* Below are the best parameters after tuning the hyper-parameter:

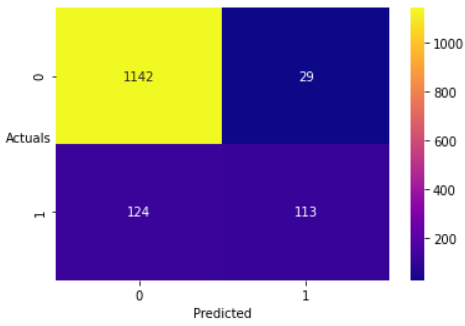


**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

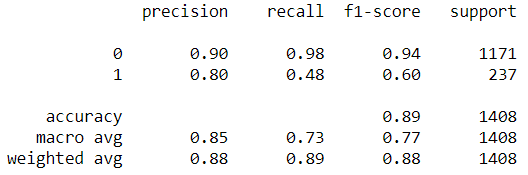


TN = 1142

TP = 113

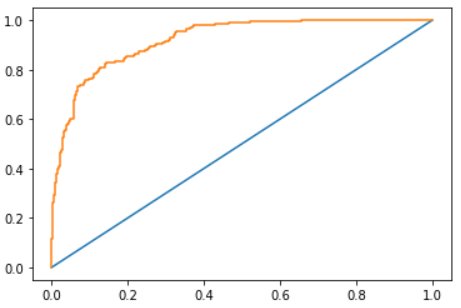
FP = 29

FN = 124



* AUC and ROC curve for Test data:





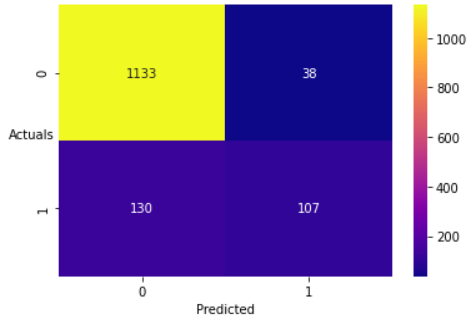
### Boosting (XGBoost)

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

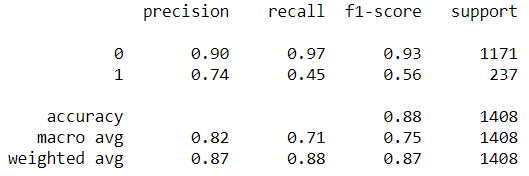


TN = 1133

TP = 107

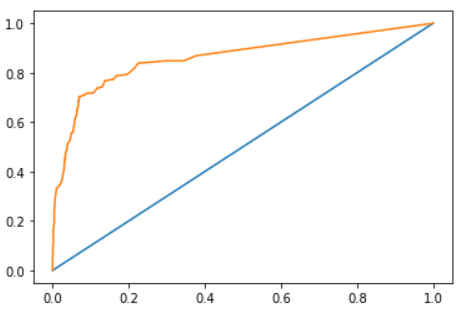
FP = 38

FN = 130



* AUC and ROC curve for Test data:





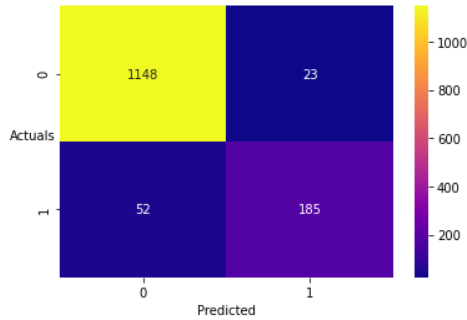
Hyper-parameter Tuning XGBoost model:

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

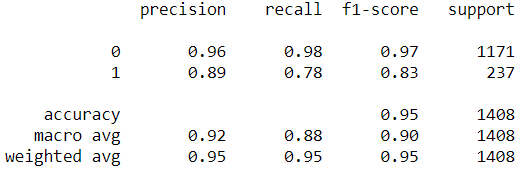


TN = 1148

TP = 185

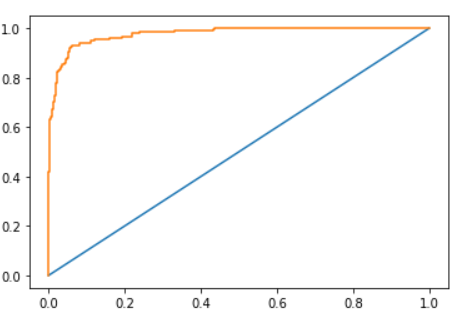
FP = 23

FN = 52



* AUC and ROC curve for Test data:





## 2.2 Hyperparameter Tuning

Hyper-parameter tunning refers to the process of fine-tuning different attributes of a model using ‘Grid Search’ for better performance. Below is hyper-parameter tuning done for different models:

### 2.2.1 Tuning LDA model

Here, we have tuned two parameters of LDA model: ‘Solver’ and ‘tol’ (tolerance rate).

We have svd, lsqr and eigen as a different solver in case of LDA model.

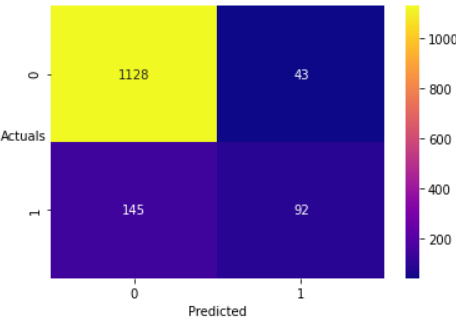
Best parameter comes out to be ‘lsqr’ as solver and tolerance rate of 1e-05.

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

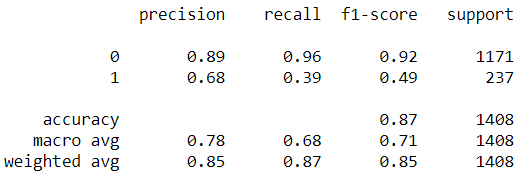


TN = 1128

TP = 92

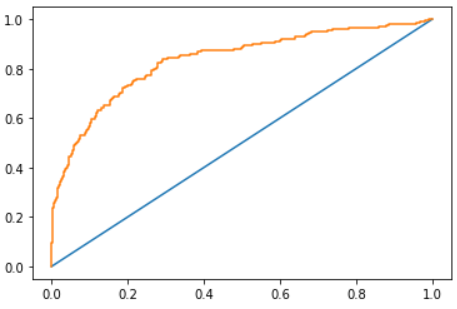
FP = 43

FN = 145



* AUC and ROC curve for Test data:





### 2.2.2 Tuning Logistic model

Here, we have tuned three parameters of Logistic model: ‘Solver’, ‘C’ and ‘tol’ (tolerance rate).

We have lbfgs, sag, saga and newton-cg as a different solver in case of Logistic model. The trade-off parameter of logistic regression that determines the strength of the regularization is called C, and higher values of C correspond to less regularization (where we can specify the regularization function).

Best parameter comes out to be:

Solver = ‘sag’

C = ‘1.0’

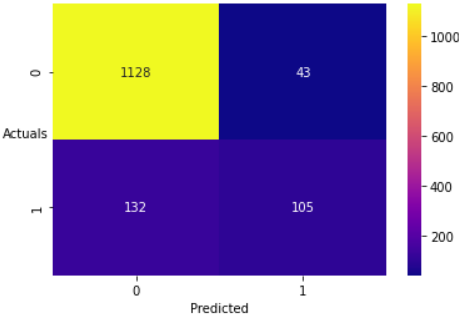
Tol = ‘0.01’

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

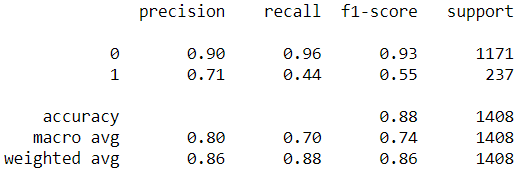


TN = 1128

TP = 105

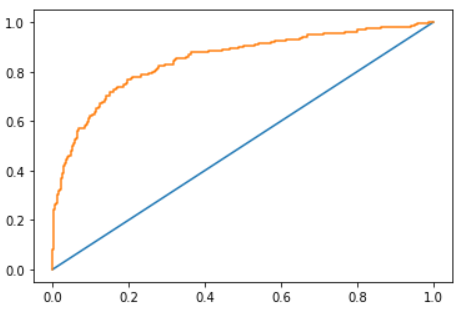
FP = 43

FN = 132



* AUC and ROC curve for Test data:





### 2.2.3 Tuning KNN model

Here, we have tuned two parameters of KNN model: ‘algorithm’ and ‘p’ (tolerance rate).

We have auto, ball-tree, kd-tree and brute as algorithm and p is Power parameter for the Minkowski metric. When p = 1, this is equivalent to using manhattan\_distance (l1), and euclidean\_distance (l2) for p = 2

Best parameter comes out to be:

algorithm = ‘auto’

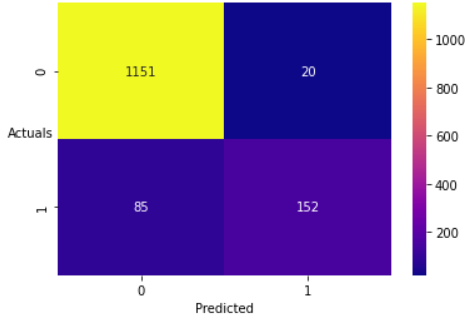
p = 1

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

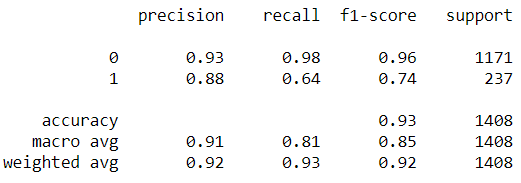


TN = 1151

TP = 152

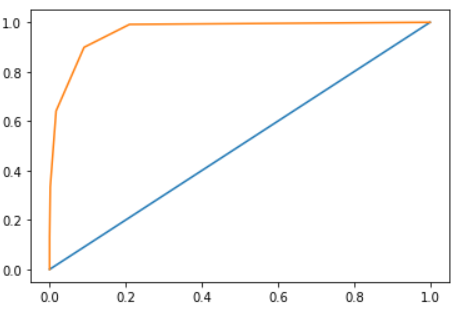
FP = 20

FN = 85



* AUC and ROC curve for Test data:





### 2.2.4 Tuning Naïve Bayes model

Here, we have tuned one parameter of Naïve Bayes model: ‘var\_smoothing’.

Portion of the largest variance of all features that is added to variances for calculation stability.

Best parameter comes out to be:

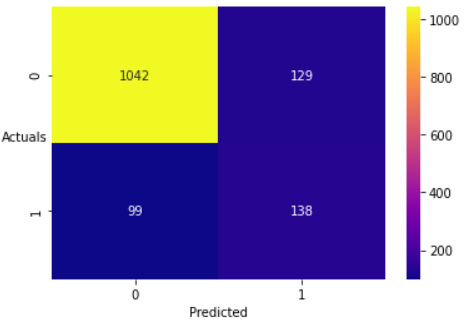
var\_smoothing= 1e-16, where 1e-9 is the default value.

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

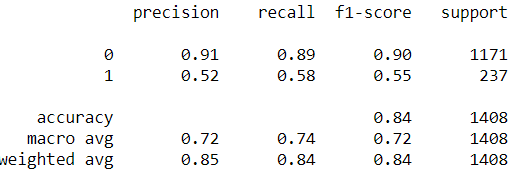


TN = 1042

TP = 138

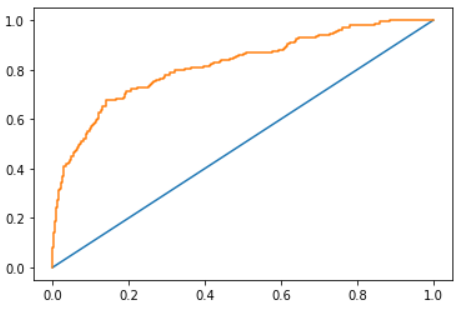
FP = 129

FN = 99



* AUC and ROC curve for Test data:





### 2.2.5 Tuning SVM model

Here, we have tuned two parameters of SVM model: ‘kernel’ and ‘tol’ (tolerance rate).

We have linear, poly, rbf and sigmoid as kernel type to be used.

Below is best parameters:

kernel = ‘poly’

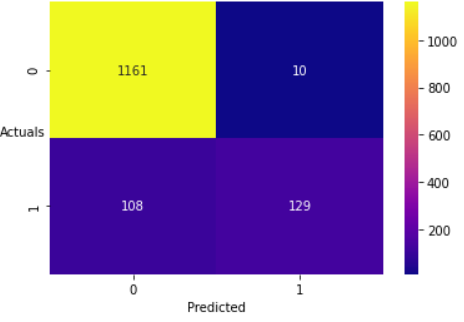
tol = 1e-05, where default value is 1e-03.

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

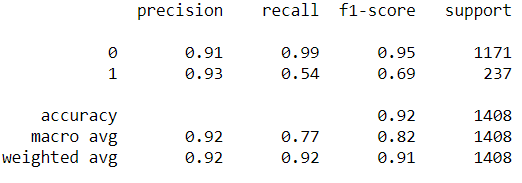


TN = 1161

TP = 129

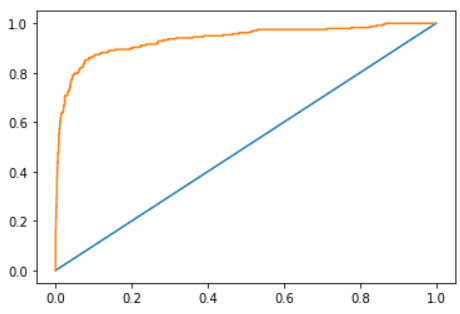
FP = 10

FN = 108



* AUC and ROC curve for Test data:





## SMOTE

SMOTE is an oversampling technique used to balance target classes in the data set. This becomes crucial in case we have a large difference between proportion of classes in target, as this may lead to models learning less of minority target class.

We apply SMOTE only on Train data set.

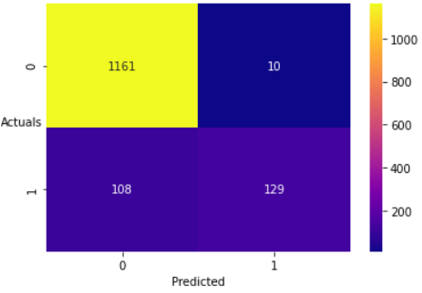
### Logistic model with SMOTE

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

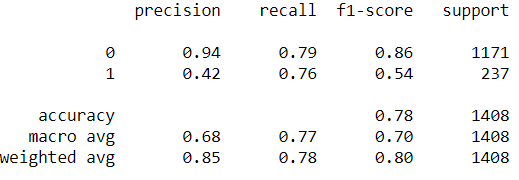


TN = 1161

TP = 129

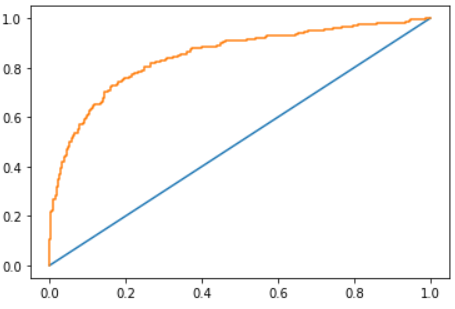
FP = 10

FN = 108



* AUC and ROC curve for Test data:





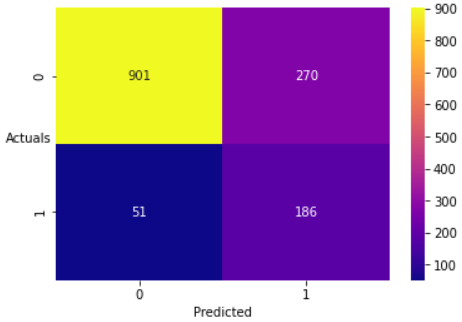
### LDA model with SMOTE

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

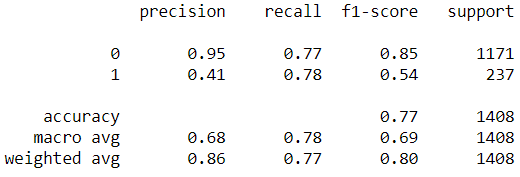


TN = 901

TP = 186

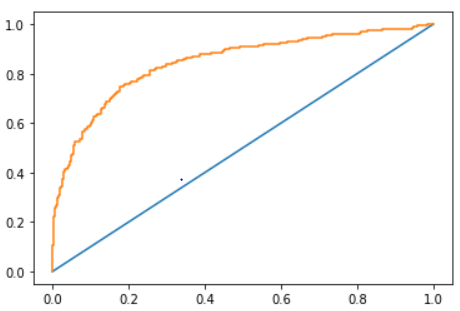
FP = 270

FN = 51



* AUC and ROC curve for Test data:





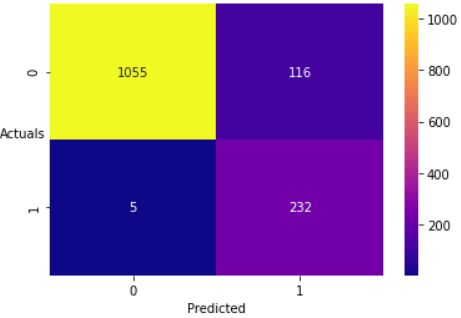
### KNN model with SMOTE

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

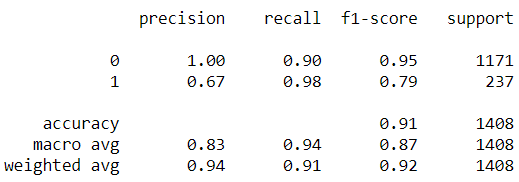


TN = 1055

TP = 232

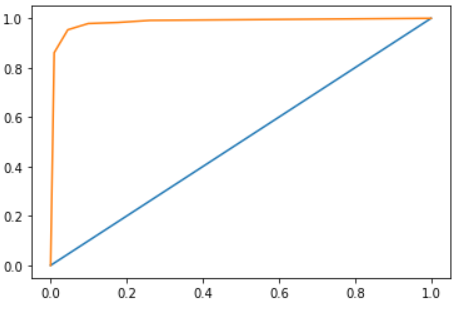
FP = 116

FN = 5



* AUC and ROC curve for Test data:





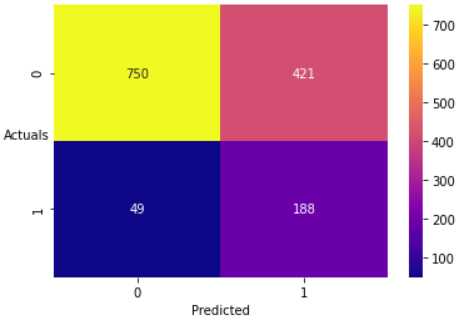
### Naïve Bayes model with SMOTE

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

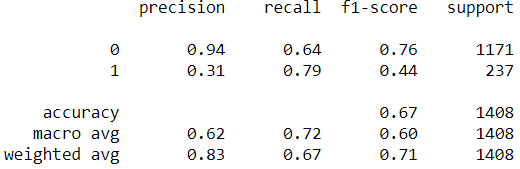


TN = 750

TP = 188

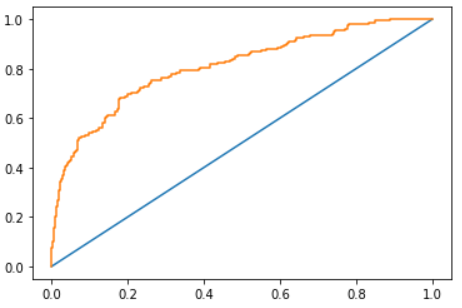
FP = 421

FN = 49



* AUC and ROC curve for Test data:





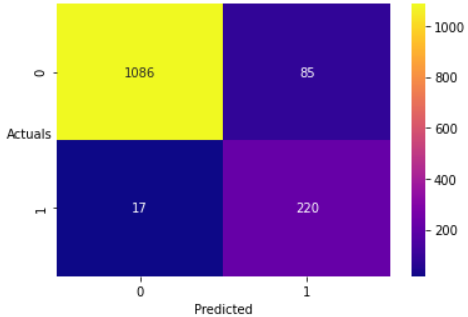
### SVM model with SMOTE

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Test data:

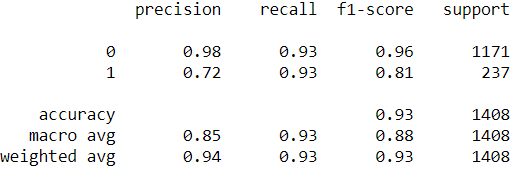


TN = 1086

TP = 220

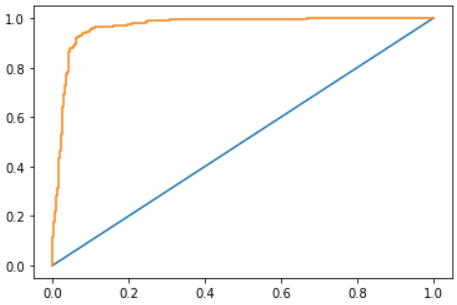
FP = 85

FN = 17



* AUC and ROC curve for Test data:





# Model Validations, Comparison & Selecting Final Model

We must analyse and decide the appropriate performance metrics of model, according to the given business case. Below is short explanation for the same:

1. The key objective here is to accurately identify the potential churning and retaining of the customers.
2. There are two cases for misclassification:

First case is predicting actual churn (1) as a non-churn (0), which is a scenario of False Negative (FN)

Second case is predicting actual non-churn (0) as a churn (1), which is scenario of False Positive (FP)

1. The scenario of FN is more cost heavy on the company because misclassifying actual churning customer as non-churning, will lead to company spending time and money on a churning customer (in form of offers, discounts and man time).
2. With above discussion, its crucial the ML model has low FN with a descent accuracy percentage.
3. Recall = TP/ (TP + FN), indicating if we have a lower FN (which is required here), recall metric have a high value.
4. Therefore, we should focus at two performance metrics primarily, Recall (the value should be high) and accuracy should be good to ensure overall accurate classification without being overfit on train data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ML\_MODEL** | **Training\_Data\_Accuracy** | **Testing\_Data\_Accuracy** | **Recall (Test Data)** | **Precision (Test Data)** | **AUC (Test Data)** |
| LDA | **0.881099005** | **0.865767045** | **0.39** | **0.68** | **0.833962101** |
| LDA Hyper Parameter Tuned | **0.880388441** | **0.866477273** | **0.39** | **0.68** | **0.83402696** |
| LDA with SMOTE | **0.809313586** | **0.772017045** | **0.78** | **0.41** | **0.846227574** |
| Logistic | **0.883230696** | **0.875** | **0.44** | **0.71** | **0.847787783** |
| Logistic Hyper Parameter Tuned | **0.884178115** | **0.875710227** | **0.44** | **0.71** | **0.847949929** |
| Logistic with SMOTE | **0.809171176** | **0.781960227** | **0.76** | **0.42** | **0.847640049** |
| KNN | **0.962813832** | **0.923295455** | **0.65** | **0.86** | **0.961614546** |
| KNN Hyper Parameter Tuned | **0.960445287** | **0.925426136** | **0.64** | **0.88** | **0.964028725** |
| KNN with SMOTE | **0.968669895** | **0.9140625** | **0.98** | **0.67** | **0.983841212** |
| Naïve Bayes | **0.849834202** | **0.838068182** | **0.58** | **0.52** | **0.818623053** |
| Naïve Bayes Hyper Parameter Tuned | **0.849834202** | **0.838068182** | **0.58** | **0.52** | **0.818623053** |
| Naïve Bayes with SMOTE | **0.727428083** | **0.666193182** | **0.79** | **0.31** | **0.809092449** |
| SVM | **0.938891521** | **0.916193182** | **0.57** | **0.89** | **0.937761011** |
| SVM Hyper Parameter Tuned | **0.941733775** | **0.916193182** | **0.54** | **0.93** | **0.935726974** |
| SVM with SMOTE | **0.962831102** | **0.927556818** | **0.93** | **0.72** | **0.967581533** |
| RF (Bagging) | **1** | **0.979403409** | **Not Required as model Overfits** | **Not Required as model Overfits** | **Not Required as model Overfits** |
| RF (Bagging) Hyper Parameter Tuned | **0.899573662** | **0.891335227** | **0.48** | **0.8** | **0.92193192** |
| XGBoost (Boosting) | **0.88394126** | **0.880681818** | **0.45** | **0.74** | **0.861018928** |
| XGBoost Hyper Parameter Tuned | **0.980577925** | **0.946732955** | **0.78** | **0.89** | **0.977277886** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Colour coding used:** | Excellent | Good | Neutral | Poor |
|  |  |  |  |  |
| **Model accuracy** | **NA** | **Good (> 90%)** | **Neutral (80% to 90%)** | **Poor (< 80%)** |
| **Recall & Precision score** | **Excellent (> 0.90)** | **Good (70% to 90%)** | **Neutral (60% to 70%)** | **Poor (< 60%)** |

***Final Model:***

* As per above metrics of various models, **SVM model with SMOTE applied to it seems to be a good model comparatively** and is **selected as final optimized model**.
* It (SVM with SMOTE) has following metric data:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ML\_MODEL** | **Training\_Data\_Accuracy** | **Testing\_Data\_Accuracy** | | **Recall (Test Data)** | **Precision (Test Data)** | | | **AUC (Test Data)** |
| SVM with SMOTE | **0.962831102** | | **0.927556818** | **0.93** | | **0.72** | **0.967581533** | |

* Recall value for test data is quite good with **93% recall score**. Recall value is number of positive class correctly identified or classified as positive.
* Also **test data accuracy is 92.75% with train data accuracy of 96.28%**
* Thus, **SVM with SMOTE applied is selected as final model**.

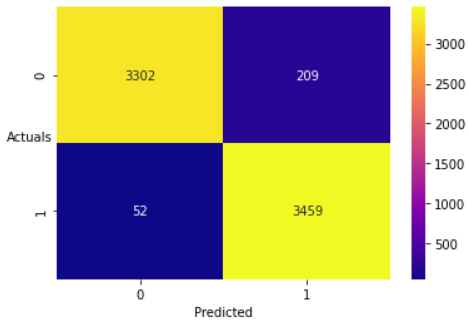
Below are some more performance metric for SVM model with SMOTE for test data:

**Performance Metrics**

* Model Accuracies:



* Confusion matrix and Classification report for Train data:

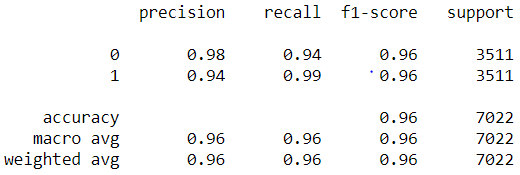


TN = 3302

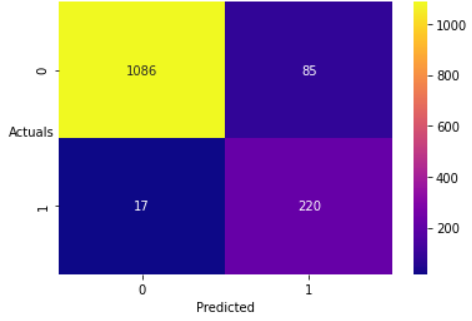
TP = 3459

FP = 209

FN = 52



* Confusion matrix and Classification report for Test data:

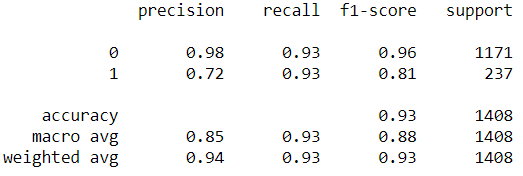


TN = 1086

TP = 220

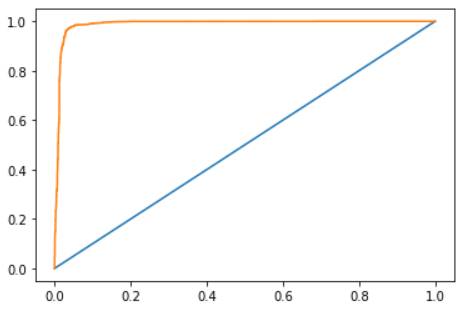
FP = 85

FN = 17



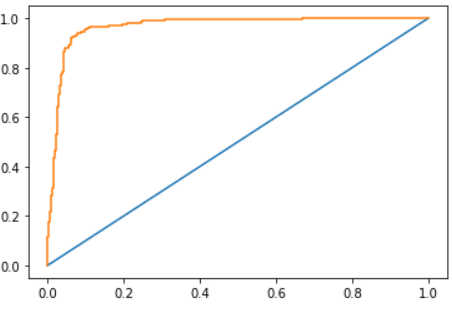
* AUC and ROC curve for Train data:





* AUC and ROC curve for Test data:





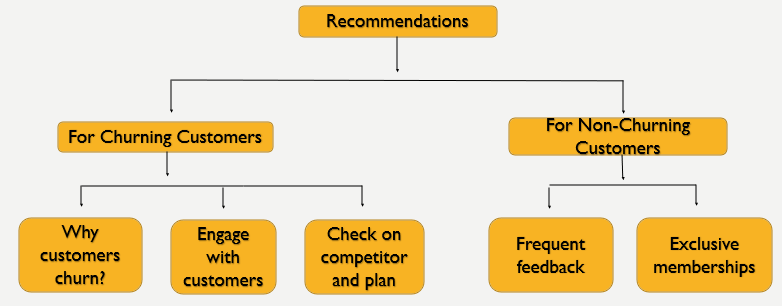
# Business Insights and Recommendations

## 4.1 Model Recommendation

SVM model with SMOTE applied can be selected as a ML model to proceed for any unseen data (test data) for predicting the churn or non-churn of the customers.

Thus, the selected model can be used in production environment for the given business case.

## 4.2 Business Recommendations:



As shown above, we have divided the business recommendation into two part:

1. Recommendations for Churning customers
2. Recommendations for Non-churning customers

### 4.2.1 Business Recommendations for Churning customers

Churning is unwanted process that cannot be stopped completely practically but surely can be aimed for reducing the churn count. Being able to take action for churning on time is itself a business benefit if done correctly.

* **Analyse why customer churns?**

An in-depth online survey for the customer aiming at analysing the pain-points of the customers with respect to the company services. This data can further be used to plan future strategy changes required.

* **Engage with customers**

Make more customer engagements through multiple communication channel like email, messages, social media and blogs, along with details and benefits of the products mentioned in the content.

* **Check on competitors and beneficial offers**

Making a check on competitors and their relative product price, offers and discounts, further planning for a discount offer on selective products will potentially reduce the customer churn. In fact, this is most used technique to bring down churn count.

### 4.2.2 Business Recommendations for Non-churning customers

Providing satisfactory service to a loyal non-churning customer ensures prolonged retention of the customer. We all agree to a statement that “Prevention is better than Cure”, Thus we should aim at providing persistently good service and customer care as preventive measure.

* **Frequent feedback**

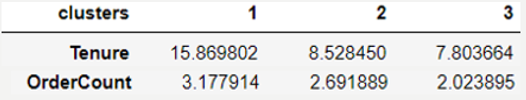
Feedbacks of non-churning customer for every order they receive and act promptly in case of any action required (like returning defective products etc.) is key to retain customers. Different channel including website and app, email and messages can be used to ask for a feedback.

* **Provide exclusive memberships**

Loyal customer can be provided with Platinum, Gold and Bronze membership based on the tenue and order counts, that offer some exclusive benefits and offers to the customers. Faster and timely product delivery, Access to a entertainment portal (if exist for the company) and early access to special discount days can be provided.

**Prescribed membership plan:**

Below is the Cluster details and exclusive membership plan: -



* **Gold Membership**: This can be provided for cluster 1 customers, as they stayed with company for longest, with highest order count.
* **Silver Membership:** This can be provided to cluster 2 customers as they are second highest in order count and stayed most long with company after cluster 1 customers.
* **Bronze Membership**: This can be offered to Cluster 3 customers to get them more attracted to the company and get the order count high.

# Appendix (PN-II)

* **Python notebook (codes):**



* **Model Comparison:**

Error! Not a valid link.