Customer Churn

* **Name**: Hemant Patidar
* **Batch**: PGPDSBA Online Sep\_A 2021
* **Date**: 04/09/2022

Contents

[**Introduction** 5](#_Toc113223902)

[**Business opportunities and need of the study/project** 5](#_Toc113223903)

[**EDA and Business Implication** 5](#_Toc113223904)

[**Data Overview** 5](#_Toc113223905)

[**Univariate Analysis** 6](#_Toc113223906)

[**Churn** 6](#_Toc113223907)

[**City Tier** 7](#_Toc113223908)

[**Service Score** 7](#_Toc113223909)

[**CC Agent Score** 8](#_Toc113223910)

[**Complain last 12 months** 8](#_Toc113223911)

[**CC Contacted last 12 month** 9](#_Toc113223912)

[**Tenure** 9](#_Toc113223913)

[**Payment** 9](#_Toc113223914)

[**Gender** 10](#_Toc113223915)

[**Account User Count** 10](#_Toc113223916)

[**Account Segment** 11](#_Toc113223917)

[**Marital** **Status** 11](#_Toc113223918)

[**Revenue Per Month** 11](#_Toc113223919)

[**Revenue Growth YoY** 12](#_Toc113223920)

[**Coupon used in last 12 month** 12](#_Toc113223921)

[**Day Since CC Connect** 12](#_Toc113223922)

[**Cashback in last 12 month** 13](#_Toc113223923)

[**Login Device** 13](#_Toc113223924)

[**Bi-Variate/Multi-Variate Analysis** 13](#_Toc113223925)

[**Correlation Plot** 13](#_Toc113223926)

[**Pair Plot** 14](#_Toc113223927)

[**City Tier vs. Churn & Service Score vs. Churn** 15](#_Toc113223928)

[**CC Agent Score vs. Churn & Complain in Last 12 month vs. Churn** 16](#_Toc113223929)

[**Gender vs. Churn & Payment vs. Churn** 16](#_Toc113223930)

[**Account per user vs. Churn & Account segment vs. Churn** 17](#_Toc113223931)

[**Insights Derived from EDA** 17](#_Toc113223932)

[**Data Cleaning and Pre-processing** 18](#_Toc113223933)

[**Missing values % and Outlier Proportion** 18](#_Toc113223934)

[**Imputation** 18](#_Toc113223935)

[**Variables Transformation** 19](#_Toc113223936)

[**Variables Removal or Addition** 19](#_Toc113223937)

[**Model building** 19](#_Toc113223938)

[**Churn proportion in data** 20](#_Toc113223939)

[**Classifier Models’ Building** 20](#_Toc113223940)

[**Logistic Regression** 20](#_Toc113223941)

[**Linear Discriminant Analysis** 21](#_Toc113223942)

[**Naïve Bayes** 22](#_Toc113223943)

[**K Nearest Neighbours (kNN)** 23](#_Toc113223944)

[**Random Forest** 24](#_Toc113223945)

[**Artificial Neural Network** 26](#_Toc113223946)

[**Positive Class prediction, ROC – AUC Score & Interpretation** 27](#_Toc113223947)

[**Model Tuning** 28](#_Toc113223948)

[**Hyper Parameter Tuning** 28](#_Toc113223949)

[**RF After SMOTE** 29](#_Toc113223950)

[**AdaBoosting** 30](#_Toc113223951)

[**Gradient Boost** 32](#_Toc113223952)

[**XG Boost** 33](#_Toc113223953)

[**Model validation** 34](#_Toc113223954)

[**Optimum Model** 35](#_Toc113223955)

[**Final interpretation / recommendation** 35](#_Toc113223956)

[**Insights** 35](#_Toc113223957)

[**Recommendations** 35](#_Toc113223958)

**List of Figures**

[Figure I - Churn 8](#_Toc113223959)

[Figure II - City Tier 8](#_Toc113223960)

[Figure III - Service Score 8](#_Toc113223961)

[Figure IV - CC Agent Score 9](#_Toc113223962)

[Figure V - Complaint In Lat 12 Months Data Plot 9](#_Toc113223963)

[Figure VI - CC Contacted Last 12 Months 10](#_Toc113223964)

[Figure VII - Tenure 10](#_Toc113223965)

[Figure VIII - Payment 10](#_Toc113223966)

[Figure IX - Gender 11](#_Toc113223967)

[Figure X - Account User Count 11](#_Toc113223968)

[Figure XI - Account Segment 12](#_Toc113223969)

[Figure XII - Marital Status 12](#_Toc113223970)

[Figure XIII - Revenue Per Month 12](#_Toc113223971)

[Figure XIV - Revenue Growth YoY 13](#_Toc113223972)

[Figure XV - Coupon Used in Last 12 Month 13](#_Toc113223973)

[Figure XVI - Days Since CC Connect 13](#_Toc113223974)

[Figure XVII - Cashback in Last 12 Month 14](#_Toc113223975)

[Figure XVIII - Login Device 14](#_Toc113223976)

[Figure XIX - Heat Map (Categorical) 15](#_Toc113223977)

[Figure XX - Heat Map (Numerical) 15](#_Toc113223978)

[Figure XXI - Pair Plot 16](#_Toc113223979)

[Figure XXII - City Tier vs. Churn & Service Score vs. Churn 16](#_Toc113223980)

[Figure XXIII - CC Agent Score vs. Churn & Complain in Last 12 month vs. Churn 17](#_Toc113223981)

[Figure XXIV - Gender vs. Churn & Payment vs. Churn 17](#_Toc113223982)

[Figure XXV (3.b) - Account/user vs. Churn & Account segment vs. Churn 18](#_Toc113223983)

[Figure XXVI - Marital Status vs. Churn & Login Device vs. Churn 18](#_Toc113223984)

[Figure XXVII - Outlier Analysis 19](#_Toc113223985)

[Figure XXVIII - Confusion Matrix of Logistic Regression 22](#_Toc113223986)

[Figure XXIX - Classification Report of Logistic Regression 22](#_Toc113223987)

[Figure XXX - Confusion Matrix of LDA Model 23](#_Toc113223988)

[Figure XXXI - Classification Report of LDA Model 23](#_Toc113223989)

[Figure XXXII - Confusion Matrix of Naive Bayes 24](#_Toc113223990)

[Figure XXXIII - Classification Report of Naive Bayes Model 24](#_Toc113223991)

[Figure XXXIV - Confusion Matrix of kNN Model 25](#_Toc113223992)

[Figure XXXV - Classification Reports of kNN Model 25](#_Toc113223993)

[Figure XXXVI - Features Importance of Random Forest (Plot) 26](#_Toc113223994)

[Figure XXXVII - Confusion Matrix of Random Forest 26](#_Toc113223995)

[Figure XXXVIII - Classification Report of Random Forest 27](#_Toc113223996)

[Figure XXXIX - Confusion Matrix of ANN Model 27](#_Toc113223997)

[Figure XL - Classification Report of ANN Model 28](#_Toc113223998)

[Figure XLI - ROC-AUC Plot All Models 29](#_Toc113223999)

[Figure XLII - Accuracy & ROC-AUC Score All Models 29](#_Toc113224000)

[Figure XLIII - Confusion Matrix of RF Tuned Model 30](#_Toc113224001)

[Figure XLIV - Classification Report of RF Tuned Model 30](#_Toc113224002)

[Figure XLV - Confusion Matrix after SMOTE 31](#_Toc113224003)

[Figure XLVI - Classification Report after SMOTE 31](#_Toc113224004)

[Figure XLVII - Feature Importance after AdaBoost 32](#_Toc113224005)

[Figure XLVIII - Confusion Matrix of AdaBoost 32](#_Toc113224006)

[Figure XLIX - Classification Report of AdaBoost 33](#_Toc113224007)

[Figure L - Confusion Matrix of Gradient Boost 33](#_Toc113224008)

[Figure LI - Classification Report after Gradient Boost 34](#_Toc113224009)

[Figure LII - Feature Importance after XGBoost 34](#_Toc113224010)

[Figure LIII - Confusion Matrix of XGBoost 35](#_Toc113224011)

[Figure LIV - Classification Report of XGBoost 35](#_Toc113224012)

**List of Tables**

[Table 1 - Descriptive Stats 9](#_Toc113224013)

[Table 2 - Unique values per Feature 9](#_Toc113224014)

[Table 3 - Missing Values and Outlier Percentage 22](#_Toc113224015)

[Table 4 - Feature Importance of Logistic Regression 24](#_Toc113224016)

[Table 5 - Features Importance of Linear Discreminent Analysis 26](#_Toc113224017)

[Table 6 - Features Importance of Random Forest 29](#_Toc113224018)

[Table 7 - Positive Class Prediction of All Models 31](#_Toc113224019)

[Table 8 - Models Comparison 39](#_Toc113224020)

# **Introduction**

We have an E Commerce company provider 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.

## **Business opportunities and need of the study/project**

Over the years, E commerce companies have grown significantly in terms of usage and purchases of items. Every E commerce company has primary goal to keep their customer base strong.

**Opportunities** –

Customers have built a habit of purchasing item with ease, and they quickly order from most used E commerce platform by them instead of finding the same product on other platforms for small profits.

E commerce companies have been steadily eating up the worldwide retail market sales, so company with lesser churn rate would grow fast.

**Need of Study –**

In order to prevent customers/account churn, we need to develop a churn prediction model that prevents customers from leaving the company.

Our study will cover key patterns and observations which makes a customer leave the company. And we should be able to make recommendations helping company ahead of their competitor.

# **EDA and Business Implication**

## **Data Overview**

* We have 11,260 rows (customers) and 19 columns (features)
* There are no duplicate data captured

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Features** | **count** | **unique** | **top** | **freq** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **AccountID** | 11260 |  |  |  | 25629.5 | 3250.62635 | 20000 | 22814.75 | 25629.5 | 28444.25 | 31259 |
| **Churn** | 11260 |  |  |  | 0.168384 | 0.374223 | 0 | 0 | 0 | 0 | 1 |
| **Tenure** | 11158 | 38 | 1 | 1351 |  |  |  |  |  |  |  |
| **City\_Tier** | 11148 |  |  |  | 1.653929 | 0.915015 | 1 | 1 | 1 | 3 | 3 |
| **CC\_Contacted\_LY** | 11158 |  |  |  | 17.867091 | 8.853269 | 4 | 11 | 16 | 23 | 132 |
| **Payment** | 11151 | 5 | Debit Card | 4587 |  |  |  |  |  |  |  |
| **Gender** | 11152 | 4 | Male | 6328 |  |  |  |  |  |  |  |
| **Service\_Score** | 11162 |  |  |  | 2.902526 | 0.725584 | 0 | 2 | 3 | 3 | 5 |
| **Account\_user\_count** | 11148 | 7 | 4 | 4569 |  |  |  |  |  |  |  |
| **account\_segment** | 11163 | 7 | Super | 4062 |  |  |  |  |  |  |  |
| **CC\_Agent\_Score** | 11144 |  |  |  | 3.066493 | 1.379772 | 1 | 2 | 3 | 4 | 5 |
| **Marital\_Status** | 11048 | 3 | Married | 5860 |  |  |  |  |  |  |  |
| **rev\_per\_month** | 11158 | 59 | 3 | 1746 |  |  |  |  |  |  |  |
| **Complain\_ly** | 10903 |  |  |  | 0.285334 | 0.451594 | 0 | 0 | 0 | 1 | 1 |
| **rev\_growth\_yoy** | 11260 | 20 | 14 | 1524 |  |  |  |  |  |  |  |
| **coupon\_used\_for\_payment** | 11260 | 20 | 1 | 4373 |  |  |  |  |  |  |  |
| **Day\_Since\_CC\_connect** | 10903 | 24 | 3 | 1816 |  |  |  |  |  |  |  |
| **cashback** | 10789 | 5693 | 155.62 | 10 |  |  |  |  |  |  |  |
| **Login\_device** | 11039 | 3 | Mobile | 7482 |  |  |  |  |  |  |  |

Table 1 - Descriptive Stats

|  |  |
| --- | --- |
| **Features** | **Distinct value count** |
| Churn | 2 |
| City\_Tier | 3 |
| CC\_Contacted\_L12m | 44 |
| Service\_Score | 6 |
| CC\_Agent\_Score | 5 |
| Complain\_l12m | 2 |
| Tenure | 38 |
| Payment | 5 |
| Gender | 4 |
| Account\_user\_count | 7 |
| account\_segment | 7 |
| Marital\_Status | 3 |
| rev\_per\_month | 59 |
| rev\_growth\_yoy | 20 |
| coupon\_used\_l12m | 20 |
| Day\_Since\_CC\_connect | 24 |
| cashback\_l12m | 5693 |
| Login\_device | 3 |

Table 2 - Unique values per Feature

## **Univariate Analysis**

### **Churn**

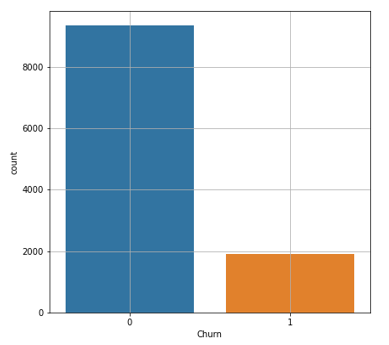
****

Figure I - Churn

Churn is our target variable which we will be predicting for existing customers through our model, and we have 9,364 (83.16% of total records) account not churned but 1,896 (16.83%) churned.

### **City Tier**

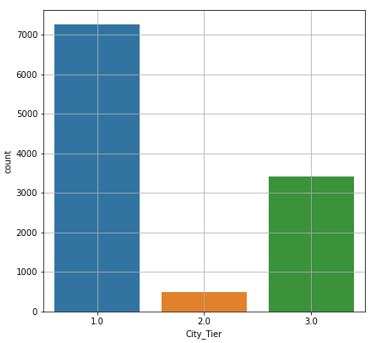
****

Figure II - City Tier

Tier 1 city has more account compared to tier 2 or 3 cities (65.15%), and tier 3 city has 3,405 (30.54%) accounts in collected data.

### **Service Score**

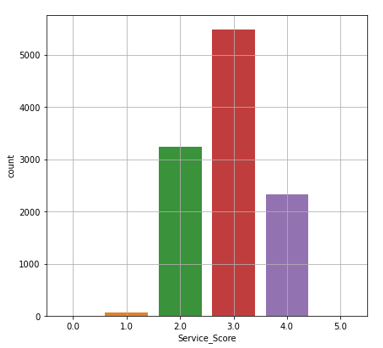


Figure III - Service Score

Almost 50% customers have given a rating of 3 on service score, and 99.99% customers gave service score between 2-4.

5 customers rated service high (satisfied with company’s service), and 85 rated low 0-1 (not satisfied).

### **CC Agent Score**

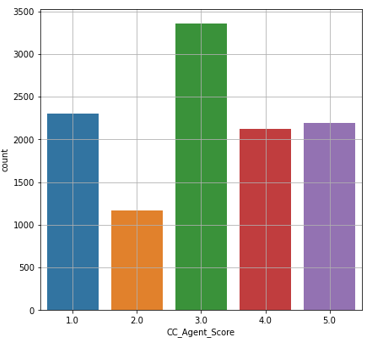


Figure IV - CC Agent Score

Almost 31% customers are not satisfied with customer care agents, given a rating of 1 or 2. And 40% customers given high satisfaction score 4 to 5.

### **Complain last 12 months**

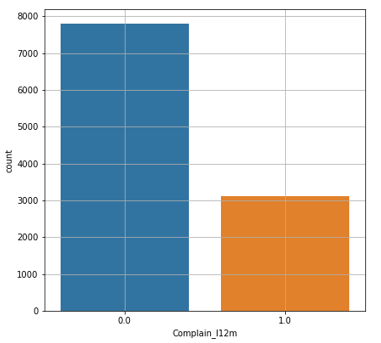
****

Figure V - Complaint In Lat 12 Months Data Plot

28.5% customers raised complain in last 12 months.

### **CC Contacted last 12 month**

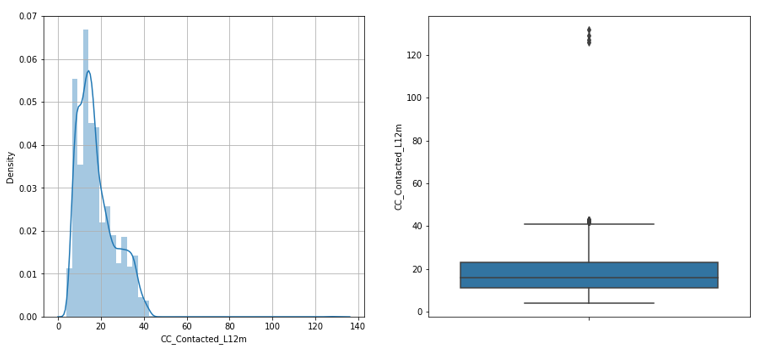
****

Figure VI - CC Contacted Last 12 Months

In the collected data, we have some outlier accounts contacted more than 100 times in last 12 months.

### **Tenure**

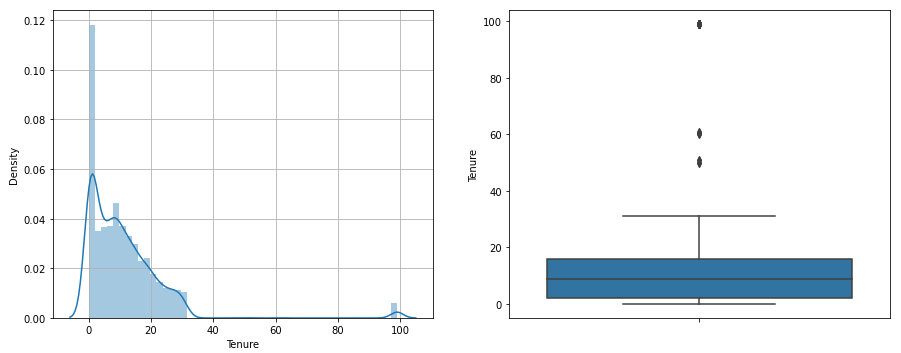


Figure VII - Tenure

In collected data, most of the customers are having low tenure (or recently joined). There are some outliers as well, where customers have stayed with company for more than 35 months.

### **Payment**

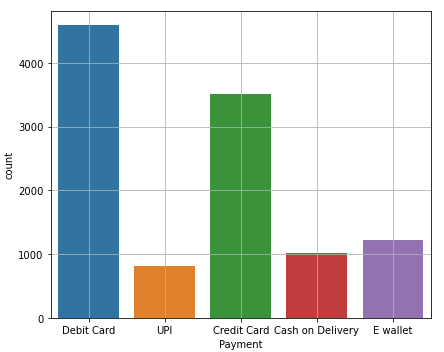


Figure VIII - Payment

Most of the customers (72%) are either paying by debit or with credit card. Few of them are paying on COD method.

### **Gender**

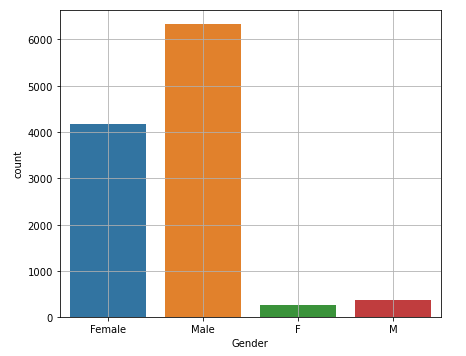


Figure IX - Gender

We will need some clean-up on gender values in collected data to show consistent values, F should be updated to Female and M to Male.

In the collected data, more customers are Male.

### **Account User Count**

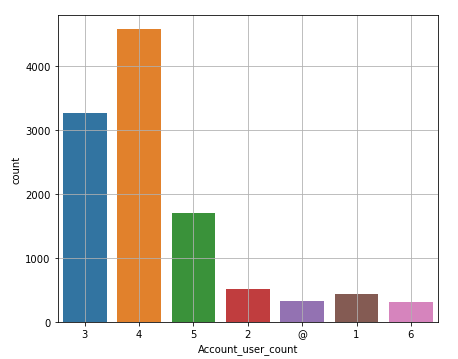


Figure X - Account User Count

Most of the accounts (85%) are shared among 3 to 5 customers. There is an invalid character in collected data “@”, we will clean-up prior to building prediction model.

### **Account Segment**

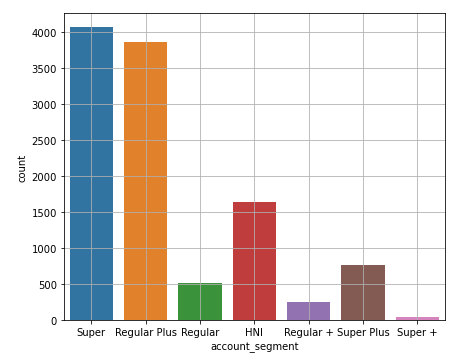


Figure XI - Account Segment

Based on spend, most of the customers (70%) are either Regular Plus or Super. In collected data, we again have inconsistent fields, “+” will be replaced with “Plus” to make them consistent.

### **Marital** **Status**

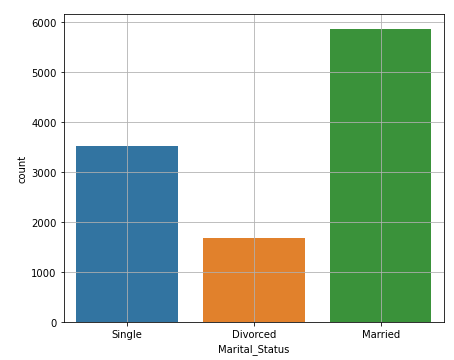


Figure XII - Marital Status

Among our company’s customers, many of them are married.

### **Revenue Per Month**

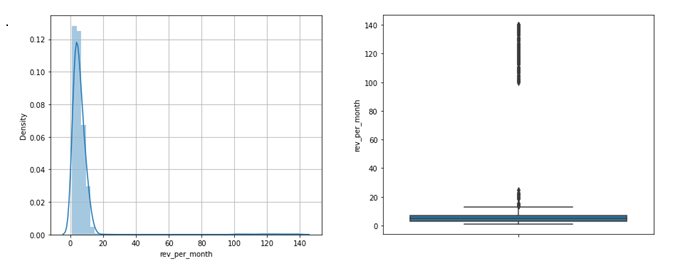


Figure XIII - Revenue Per Month

Average month revenue generated by account are between 0 and 20, but there are some extreme values present representing accounts generating >100.

### **Revenue Growth YoY**

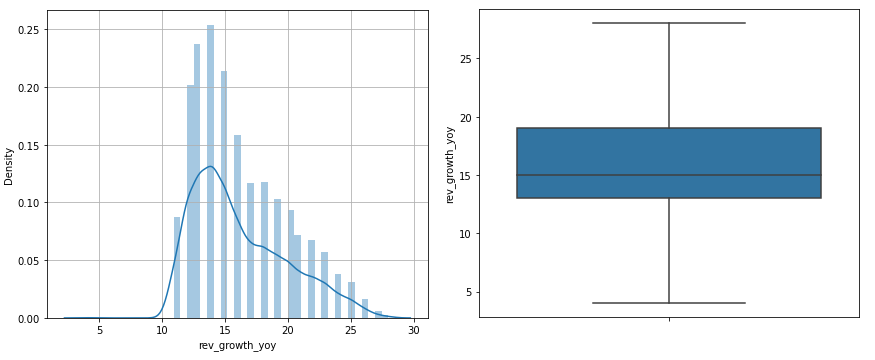


Figure XIV - Revenue Growth YoY

In collected data, we don’t have any outlier present where account is generating significantly higher or lower revenue compared to its’ last year revenue. All the data lies between 5 and 30.

### **Coupon used in last 12 month**

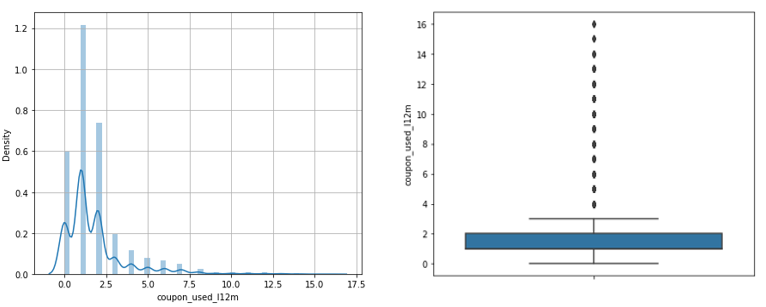


Figure XV - Coupon Used in Last 12 Month

Customers have used no coupon to 3 coupons. There are some outliers present in the data, but they are not truly extreme values, for an account customer is eligible to use more coupons than others.

### **Day Since CC Connect**

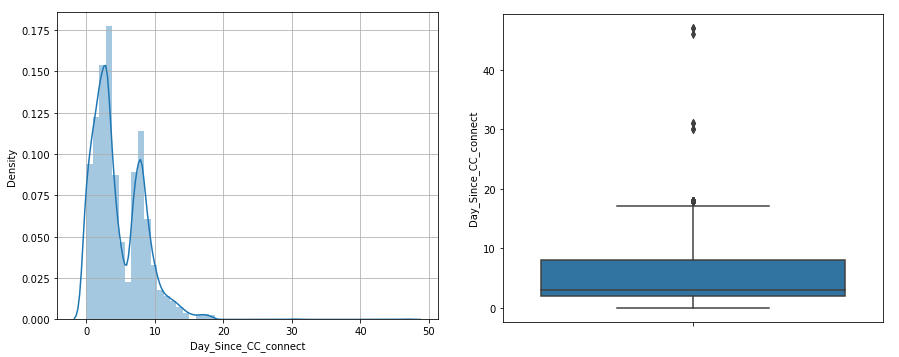


Figure XVI - Days Since CC Connect

Some of the customers haven’t contacted customer care since a month. And most of them contacted in 0 – 20 days.

### **Cashback in last 12 month**

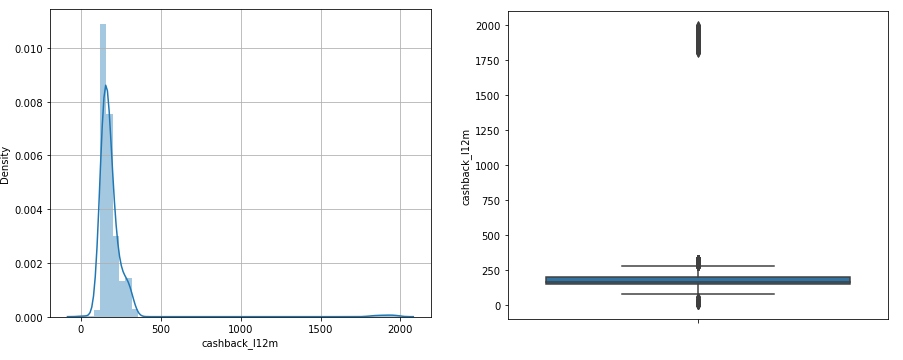


Figure XVII - Cashback in Last 12 Month

E commerce companies usually offers cashback to attract customers, and in our collected data, we can see some high cashback to customers that’s greater than 1750.

### **Login Device**

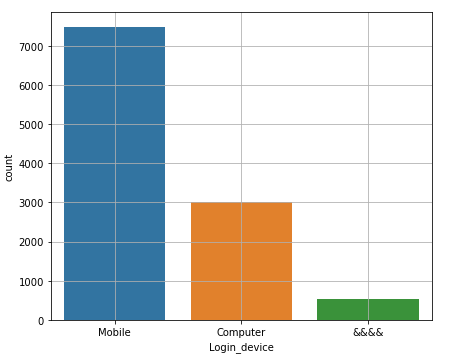


Figure XVIII - Login Device

We have some un-clear login type as &&&& but looking at the rest of the data we can infer that most of the customers are using mobile device.

## **Bi-Variate/Multi-Variate Analysis**

### **Correlation Plot**

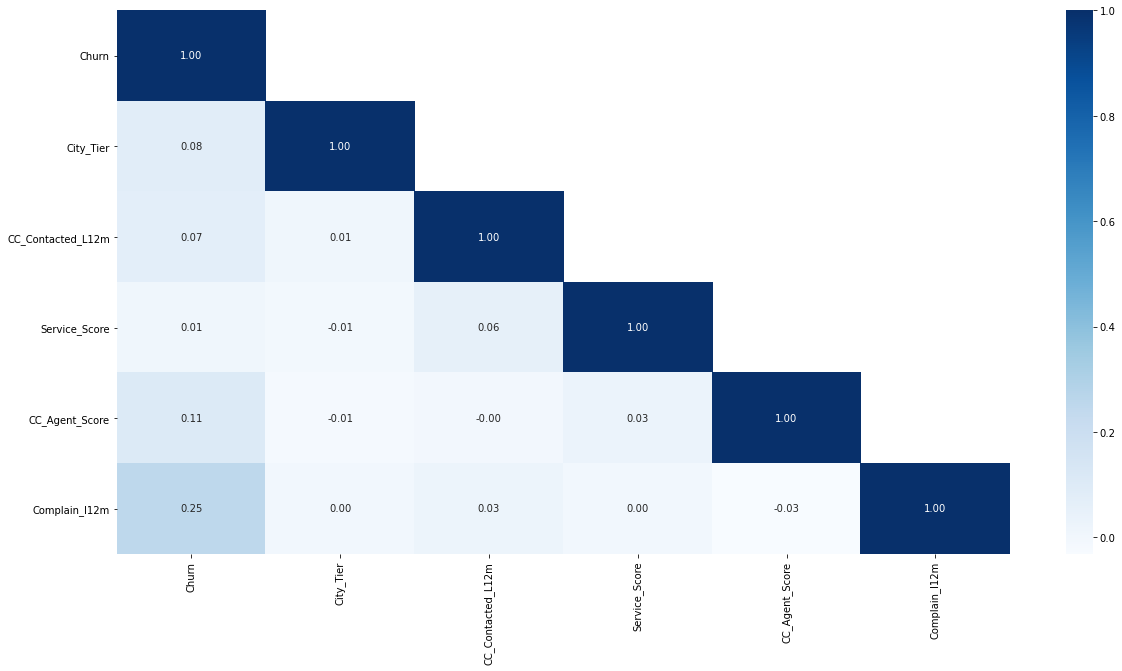
****

Figure XIX - Heat Map (Categorical)

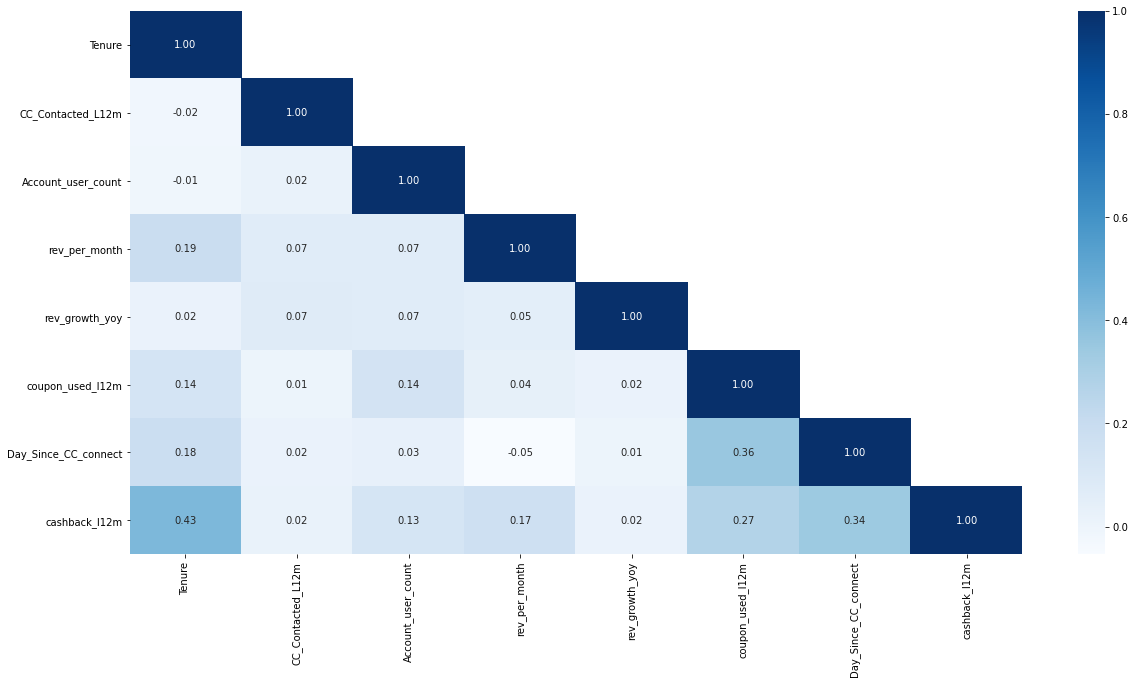


Figure XX - Heat Map (Numerical)

There isn’t any significant correlation between variables, but since we had only 6 numeric features, correlation was calculated only for those leaving some object features of collected data.

### **Pair Plot**

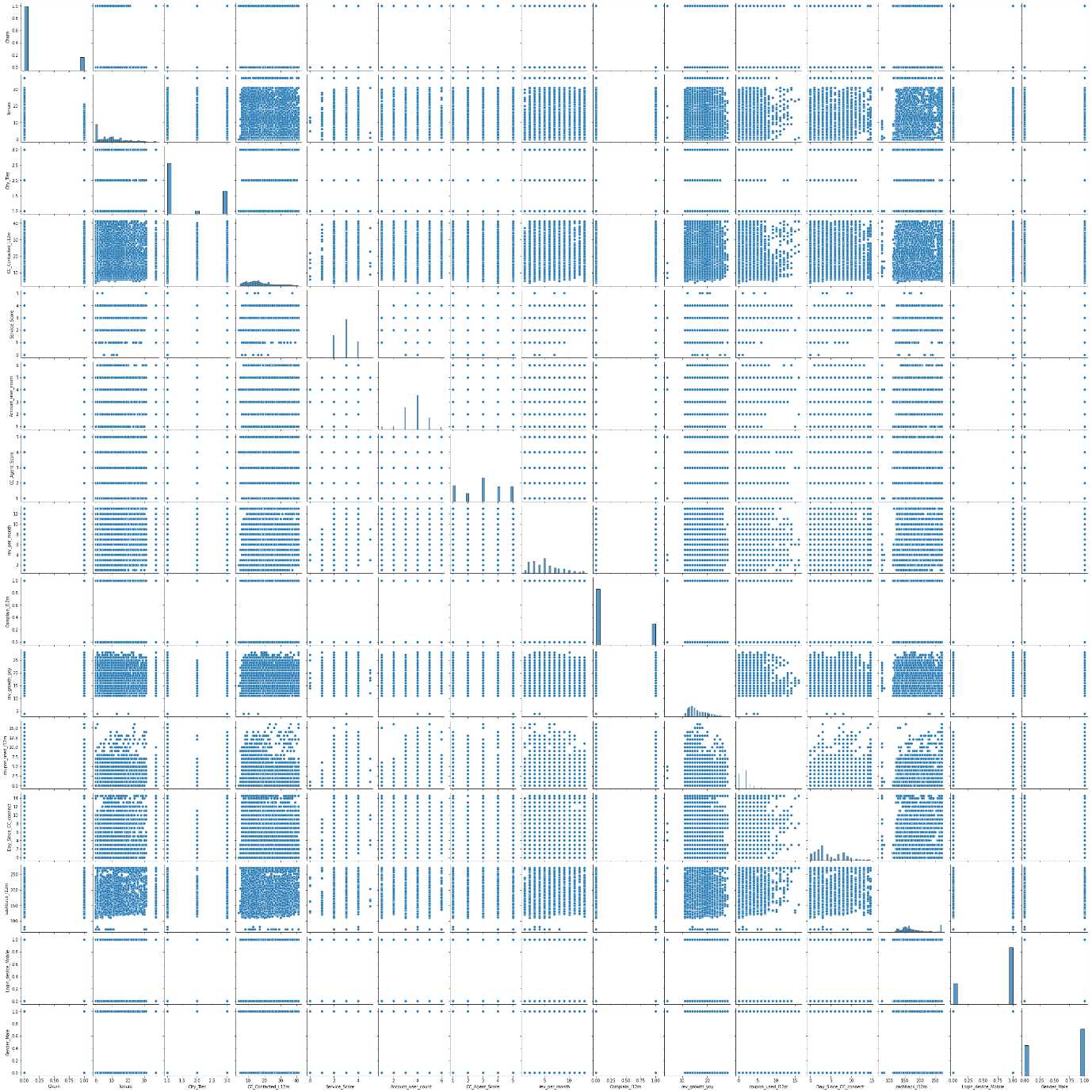


Figure XXI - Pair Plot

### **City Tier vs. Churn & Service Score vs. Churn**

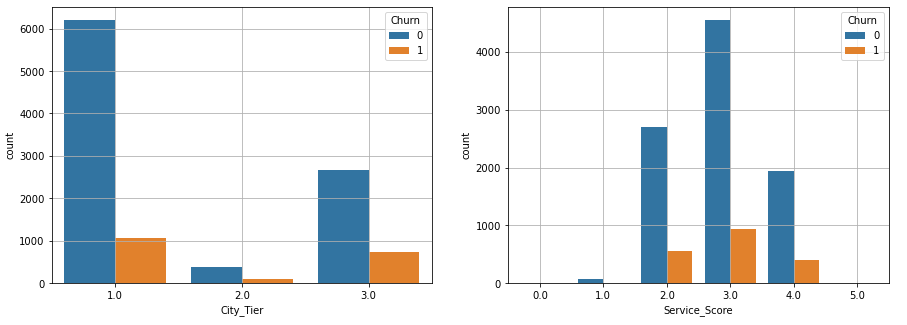


Figure XXII - City Tier vs. Churn & Service Score vs. Churn

We don’t have any significant evidence if people rating low to company service are leaving, but tier 3 cities seem to have higher churn rate compared to tier 1 or 2 cities.

### **CC Agent Score vs. Churn & Complain in Last 12 month vs. Churn**

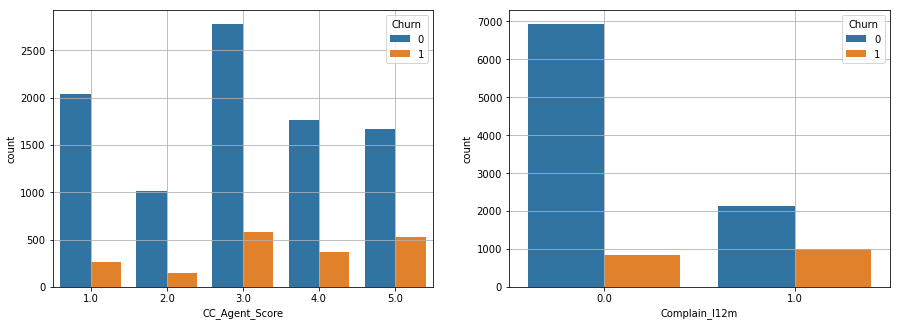


Figure XXIII - CC Agent Score vs. Churn & Complain in Last 12 month vs. Churn

Customers giving high satisfaction score to agent also seem to leave the company.

We have high churn rate on the accounts from which complains were raised in last 12 months.

### **Gender vs. Churn & Payment vs. Churn**

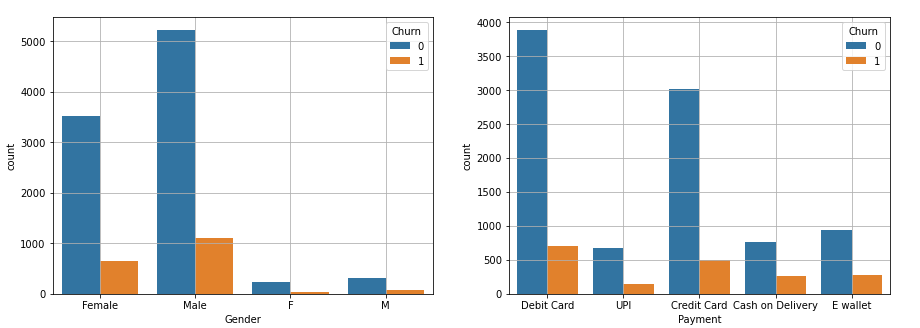


Figure XXIV - Gender vs. Churn & Payment vs. Churn

We cannot conclude much from above plots, as account churned are high for males and debit card.

### **Account per user vs. Churn & Account segment vs. Churn**

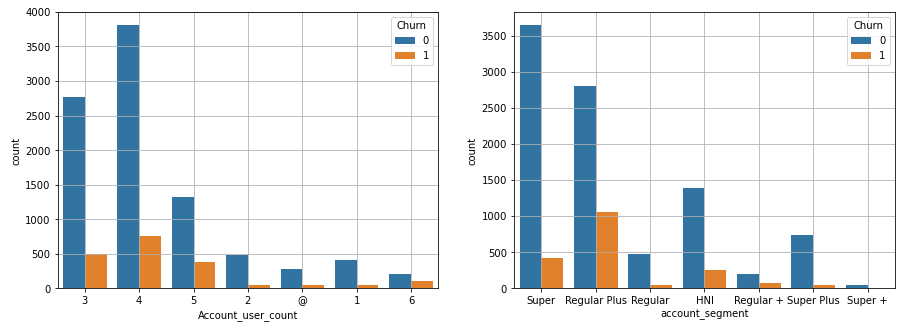


Figure XXV (3.b) - Account/user vs. Churn & Account segment vs. Churn

Customers with spending habits categorized as “Regular Plus” have high churn rate. They seem to be switching the E commerce companies.

**Marital Status vs. Churn & Login Device vs. Churn**

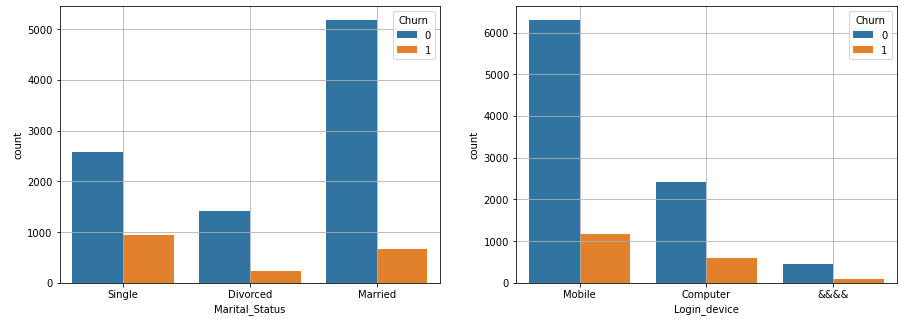


Figure XXVI - Marital Status vs. Churn & Login Device vs. Churn

No conclusion can be made by looking at both graphs. High number of churns are on high number of data category.

## **Insights Derived from EDA**

* **Tier-1** cities have more E-commerce usage compared to other cities.
* Customers have given **average rating (3/5)** to company and customer care agents.
* Most of the customers are **using debit or credit cards** for payment.
* On average an account being shared by **3 to 4 users** (customers).
* Most of the customers are using **mobile devices** for login/shopping.
* **Tier-3** cities have **high churn rate.**
* Customers who have **raised complains are likely to churn**.
* **Single** customers seem to **switch** E commerce companies, shows churn rate high.

# **Data Cleaning and Pre-processing**

## **Missing values % and Outlier Proportion**

|  |  |  |
| --- | --- | --- |
| **Features** | **Missing Values %** | **Outlier %** |
| Churn | 0.00% | 0.00% |
| Tenure | 0.91% | 1.23% |
| City\_Tier | 0.99% | 0.00% |
| CC\_Contacted\_L12m | 0.91% | 0.37% |
| Payment | 0.97% | 0.00% |
| Gender | 0.96% | 0.00% |
| Service\_Score | 0.87% | 0.00% |
| Account\_user\_count | 0.99% | 0.00% |
| account\_segment | 0.86% | 0.00% |
| CC\_Agent\_Score | 1.03% | 0.00% |
| Marital\_Status | 1.88% | 0.00% |
| rev\_per\_month | 0.91% | 1.64% |
| Complain\_l12m | 3.17% | 0.00% |
| rev\_growth\_yoy | 0.00% | 0.00% |
| coupon\_used\_l12m | 0.00% | 12.26% |
| Day\_Since\_CC\_connect | 3.17% | 1.15% |
| cashback\_l12m | 4.18% | 8.76% |
| Login\_device | 1.96% | 0.00% |

Table 3 - Missing Values and Outlier Percentage

* Cashback data has 4.18% missing values followed by *Days since CC Connect* and *Complain raised in last 12 months* with 3.17% of missing values
* Feature *Coupon used in last 12 months* having 12.26% of outliers and Cashback data has 8.76%

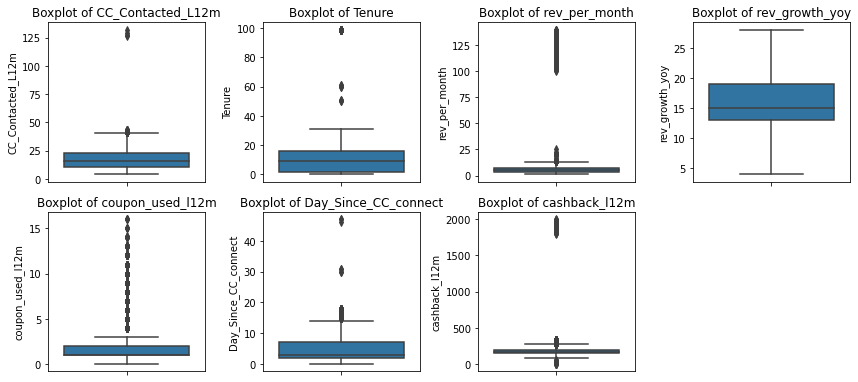
****

Figure XXVII - Outlier Analysis

### **Imputation**

* We have imputed missing values in categorical features with **Mode** and for continuous features, if have no outlier than **Mean** else **Median**
* For outliers, *Coupon used* and *Cashback* can be high/low for any customer and should not be treated consider of authenticity of data. Rest of the features having outliers were imputed as
  + Values lower than Q1 – 1.5\*IQR as (Q1 - 1.5\*IQR)
  + Values higher than Q3 + 1.5\*IQR as (Q3 + 1.5\*IQR)

## **Variables Transformation**

Features identified as numerical but should be categorical –

* + Churn – Have values either 0 or 1
  + City Tier – Have values either 1, 2 or 3, for tier – 1, tier – 2, or tier – 3 cities.
  + Service Score – Have values from 0 to 5
  + CC Agent Score – Have values from 1 to 5
  + Complain\_l12m – Have values either 0 or 1

Invalid value fix

* + **Gender –** Updated “M” to “Male” and “F” to “Female”
  + **Account User Count –** Fixed “@” with median of account\_user\_count field’s values
  + **Account Segment –** Updated “Regular +” to “Regular Plus” and “Super +” to “Super Plus”
  + **Login Device–** Fixed “&&&&” with mode of Login\_device field’s values

## **Variables Removal or Addition**

* AccountID has been removed as that’s only an identification for borrower and should not be used for analysis
* Categorical variables (which are non-numeric) were label encoded to feed values in models.
  + Payment
  + Account Segment
  + Marital Status

# **Model building**

We want to predict customers who may churn in future, hence for assisting business we will be creating predictive models.

Since our target variable is binary, we have created below classifier models for churn prediction –

* Logistic Regression
* Linear Discriminant Analysis
* Naïve Bayes
* K Nearest Neighbours
* Random Forest
* Artificial Neural Network

## **Churn proportion in data**

We have **16.83%** account churned and rest **83.16%** are non-churned.

For model building, we are using 70% data as train data and 30% data will be used for validating those models.

## **Classifier Models’ Building**

### **Logistic Regression**

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

We have built logistic regression model with default parameters.

* + Penalty – l2
  + Tolerance – 0.0001
  + Solver – lbfgs
  + Max Iteration – 100

**[LR Model] Feature Importance –**

Top 5

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| Complain\_l12m | 1.476819 |
| Marital\_Status\_Single | 0.542461 |
| Account\_user\_count | 0.290301 |
| City\_Tier | 0.279265 |
| Payment\_E wallet | 0.259718 |

Table 4 - Feature Importance of Logistic Regression

Logistic Regression gives **Complain in last 12 months** most importance in decisions/churn prediction.

**[LR Model] Confusion Matrix –**

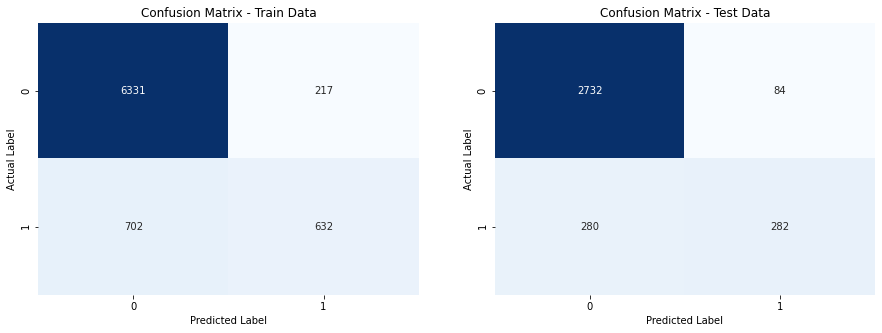


Figure XXVIII - Confusion Matrix of Logistic Regression

**[LR Model] Classification Reports –**

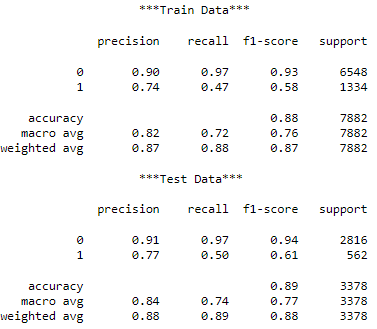


Figure XXIX - Classification Report of Logistic Regression

Logistic Regression model predicts churn with 88% on test data and 89% on train data…

### **Linear Discriminant Analysis**

LDA is mainly used in classification problems where you have a categorical output variable. It allows both binary classification and multi-class classification.

We have built Linear Discriminant Analysis model with default parameters.

* + Solver – svd
  + Tolerance – 0.0001

**[LDA Model] Feature Importance –**

Top 5

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| Complain\_l12m | 1.909802 |
| account\_segment\_Regular | 0.972582 |
| Marital\_Status\_Single | 0.909184 |
| account\_segment\_Super Plus | 0.37308 |
| Account\_user\_count | 0.37069 |

Table 5 - Features Importance of Linear Discreminent Analysis

Linear Discriminant Analysis gives **Complain in last 12 months** most importance in decisions/churn prediction.

**[LDA Model] Confusion Matrix –**

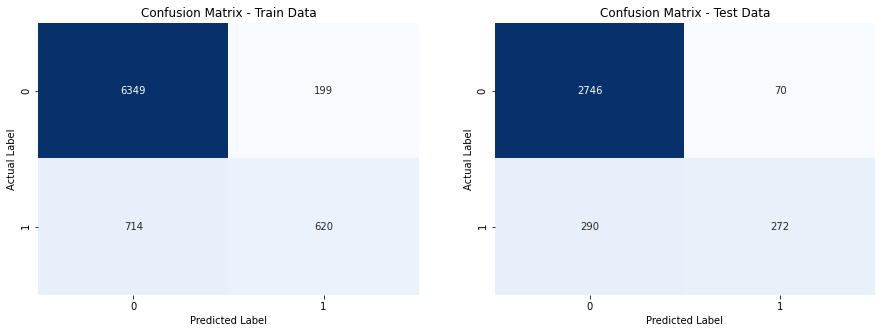


Figure XXX - Confusion Matrix of LDA Model

**[LDA Model] Classification Reports –**

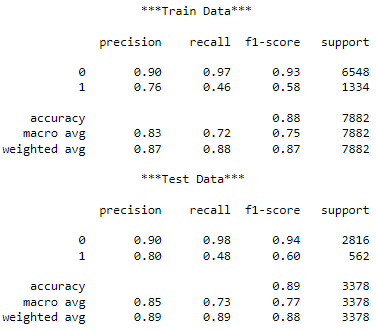


Figure XXXI - Classification Report of LDA Model

LDA Model predict churn with 88% on train data and 89% on test data, but precision and recall has not been improved compared to logistic regression model.

### **Naïve Bayes**

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

We have built Naïve Bayes model with default parameters.

**[Naïve Bayes Model] Confusion Matrix –**

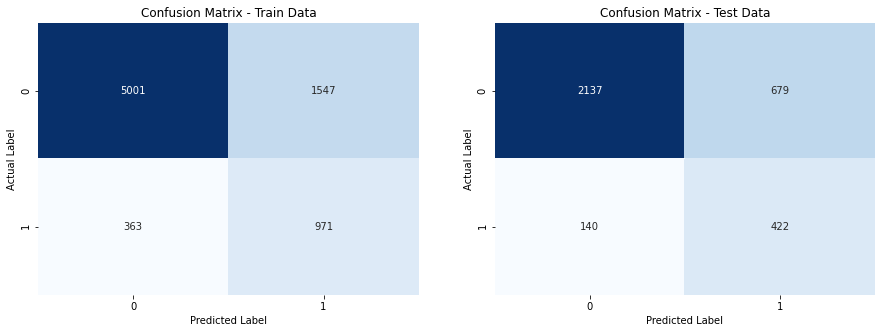


Figure XXXII - Confusion Matrix of Naive Bayes

**[Naïve Bayes] Classification Reports –**

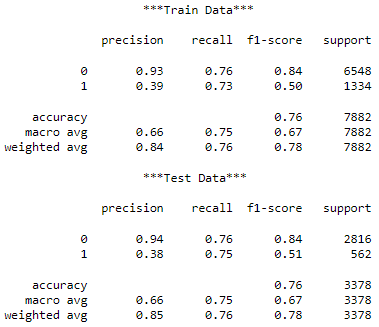
****

Figure XXXIII - Classification Report of Naive Bayes Model

Naïve Bayes predicts churn with 76% on train and test data but performs poorly compared to Logistic Regression and Linear Discriminant Analysis.

### **K Nearest Neighbours (kNN)**

KNN aims for pattern recognition tasks. K-Nearest Neighbour also known as KNN is a supervised learning algorithm that can be used for regression as well as classification problems.

For K-Nearest Neighbours (kNN) model, we were needed to bring data in uniform scale…

Post that we have created model with default parameters.

* + Weights – Uniform
  + Metric – Minkowski
  + Leaf Size – 30

**[kNN Model] Confusion Matrix –**

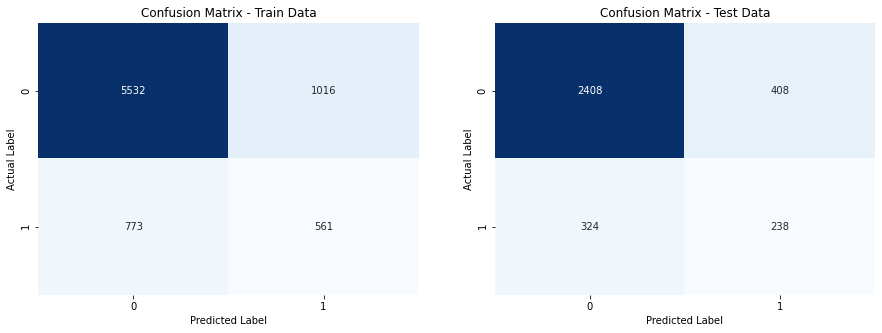


Figure XXXIV - Confusion Matrix of kNN Model

**[kNN Model] Classification Reports –**

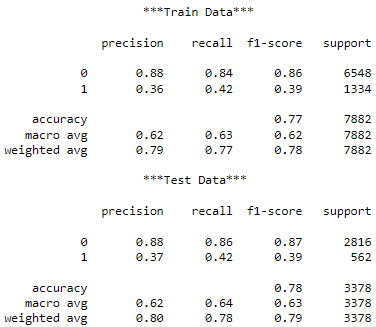


Figure XXXV - Classification Reports of kNN Model

kNN Model predicts churn with 77% on train data and 78% on test data. But logistic regression and linear discriminant model are still performing better than kNN.

### **Random Forest**

Random forests or random decision forests is an ensemble learning method for classification, it uses multiple decision trees to predict target variable.

We have built Random Forest model with default parameters –

* + N Estimators – 100
  + Criterion – Gini
  + Minimum Sample Split – 2
  + Min Sample Leaf – 1

**[RF Model] Feature Importance –**

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| Tenure | 23.788704 |
| cashback\_l12m | 8.138678 |
| CC\_Contacted\_L12m | 7.246582 |
| Day\_Since\_CC\_connect | 6.868372 |
| rev\_growth\_yoy | 6.347791 |
| rev\_per\_month | 5.865139 |
| Complain\_l12m | 5.853284 |
| CC\_Agent\_Score | 5.402495 |

Table 6 - Features Importance of Random Forest

Random forest model gives **Tenure** highest importance followed by cashback received by customers in last 12 months.

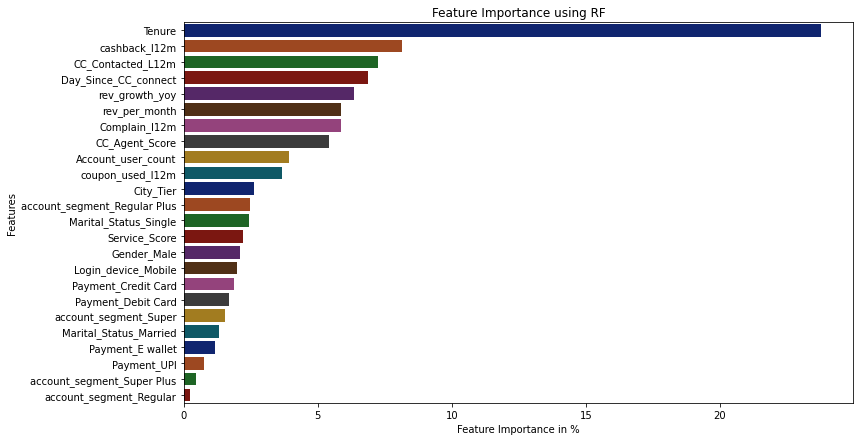


Figure XXXVI - Features Importance of Random Forest (Plot)

**[RF Model] Confusion Matrix –**

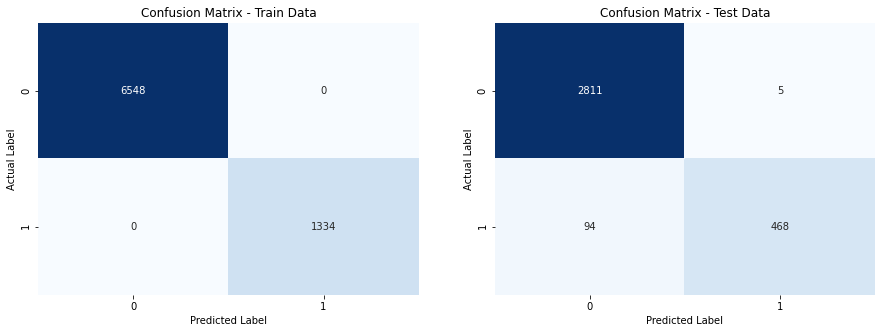


Figure XXXVII - Confusion Matrix of Random Forest

**[RF Model] Classification Reports –**

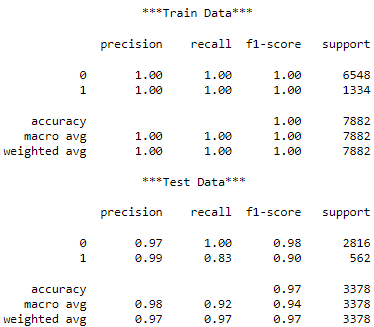


Figure XXXVIII - Classification Report of Random Forest

Random Forest model seems to be over-fitted (works best on train data but bad on test data)

The model predicts churn with 97% on test data which is best among created models.

### **Artificial Neural Network**

An artificial neural network is an interconnected group of nodes, inspired by a simplification of neurons in a brain.

We have built Neural Network Model with default parameters –

* + Activation – Relu
  + Solver – Adam
  + Hidden Layer Sizes – (100, --, --)

**[ANN Model] Confusion Matrix –**

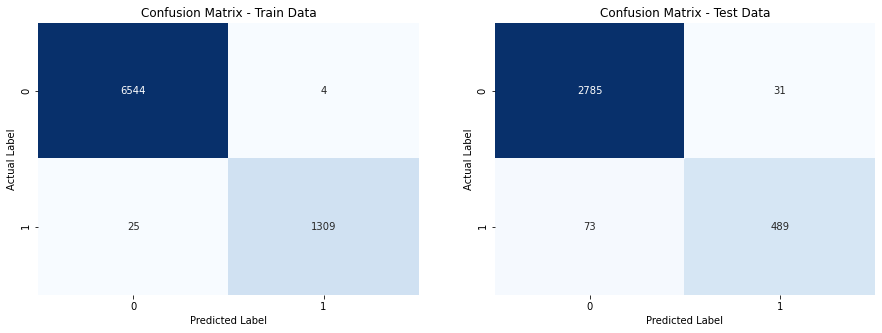


Figure XXXIX - Confusion Matrix of ANN Model

**[ANN Model] Classification Reports –**

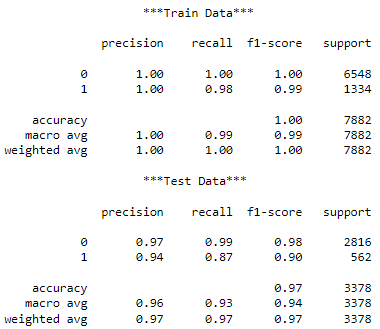


Figure XL - Classification Report of ANN Model

ANN model again over-fits the results, it gives 100% accuracy on train data and 97% accuracy on test data.

Random forest and Artificial neural network both are working better compared to other models, but Random Forest is form of ensemble technique hence gives us better prediction confidence.

## **Positive Class prediction, ROC – AUC Score & Interpretation**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Train Data** | | | **Test Data** | | |
| **Precision** | **Recall** | **F1-Score** | **Precision** | **Recall** | **F1-Score** |
| **Logistic Regression** | 74% | 47% | 58% | 77% | 50% | 61% |
| **Linear Discriminant Analysis** | 76% | 46% | 58% | 80% | 48% | 60% |
| **Naïve Bayes** | 39% | 73% | 50% | 38% | 75% | 51% |
| **KNN Model** | 36% | 42% | 39% | 37% | 42% | 39% |
| **Random Forest** | **100%** | **100%** | **100%** | **99%** | **83%** | **90%** |
| **Artificial Neural Network** | 100% | 98% | 99% | **94%** | 87% | 90% |

Table 7 - Positive Class Prediction of All Models

* Logistic regression and Linear discriminant analysis model gives us 89% accuracy and has AUC score around 87%.
* Native Bayes predicts with 76% accuracy and has perform poorly on our data.
* kNN model has high AUC score but on default probability threshold its accuracy drops to 78% (which is not good compared to LR and LDA model)
* Logistic Regression and LDA model give “Complain raised in last 12 month” feature most importance, whereas Random Forest gives most importance to “Tenure”.
* Random Forest and Neural Network overfits the result, perform good on training data and their accuracy gets dropped on test data.
* Since Random Forest has good AUC score compared to ANN model, and has better precision, so Random Forest works best in Churn prediction.

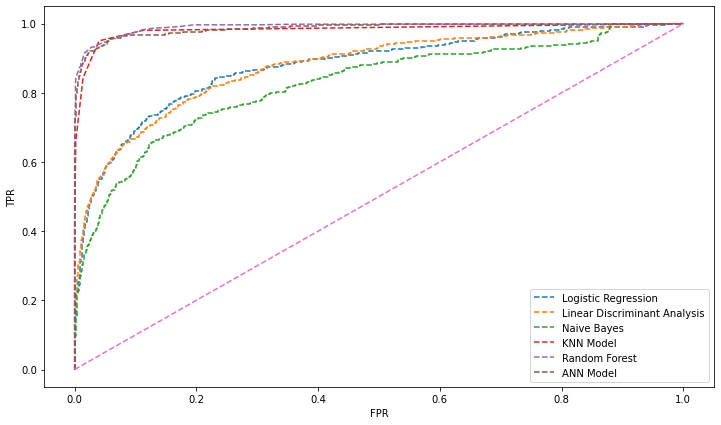


Figure XLI - ROC-AUC Plot All Models

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | |  |
| **Train Data** | **Test Data** | **ROC AUC Score** |
| **Logistic Regression** | 88% | 89% | 87.64% |
| **Linear Discriminant Analysis** | 88% | 89% | 87.58% |
| **Naïve Bayes** | 76% | 76% | 82.50% |
| **KNN Model** | 77% | 78% | 98.22% |
| **Random Forest** | 100% | 97% | **99.09%** |
| **Artificial Neural Network** | 100% | 97% | 98.63% |

Figure XLII - Accuracy & ROC-AUC Score All Models

## **Model Tuning**

### **Hyper Parameter Tuning**

Hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process.

Parameters accepted by RF classifier Model and different values from which we are trying to identify optimal ones –

* + Minimum Samples Split (in each decision tree) – 10, 20 or 30
  + Maximum Depth (of each decision tree) – 15 or 20
  + Maximum Features (in each decision tree) – 7, 8 or 9
  + N estimators (Number of trees) – 100, 200 or 300

Optimal parameter values came out as –

* + Minimum Samples Split (in each decision tree) – 10
  + Maximum Depth (of each decision tree) – 15
  + Maximum Features (in each decision tree) – 8
  + N estimators (Number of trees) – 200

**[RF Tuned Model] Confusion Matrix –**

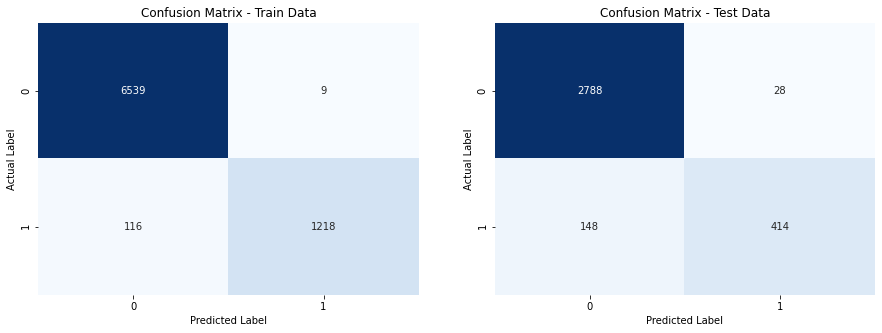


Figure XLIII - Confusion Matrix of RF Tuned Model

**[RF Tuned Model] Classification Reports –**

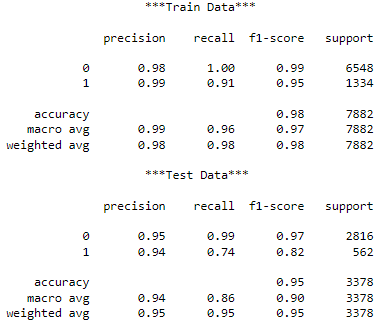


Figure XLIV - Classification Report of RF Tuned Model

With tuned parameter, RF model came down to 98% accuracy on training data and 95% on validation.

The model can be considered for prediction as we have reduced number of features being used and other parameters which were forming decision tree at maximum depth possible.

### **RF After SMOTE**

We had only 16.83% churned data and 83.16% non-churned. We have added synthetic data points on training data using SMOTE to see if it improves our model’s accuracy –

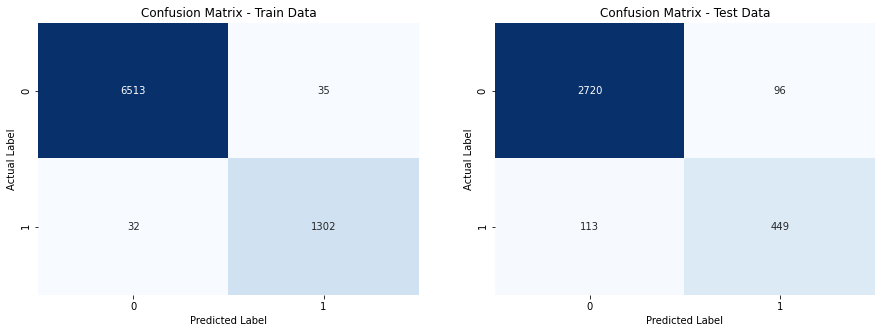


Figure XLV - Confusion Matrix after SMOTE

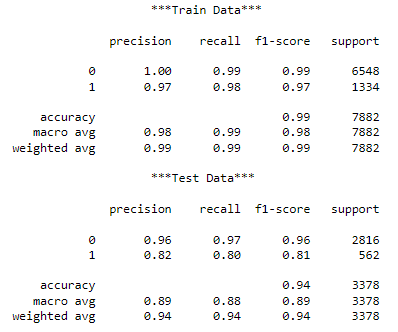


Figure XLVI - Classification Report after SMOTE

### **AdaBoosting**

It fits a sequence of weak learners on different weighted training data. It starts by predicting original data set and gives equal weight to each observation.

We have created Ada Boost model with base estimator of Tuned Random Forest model and train weak learners by 0.1 learning rate.

**[AdaBoost] Features Importance –**

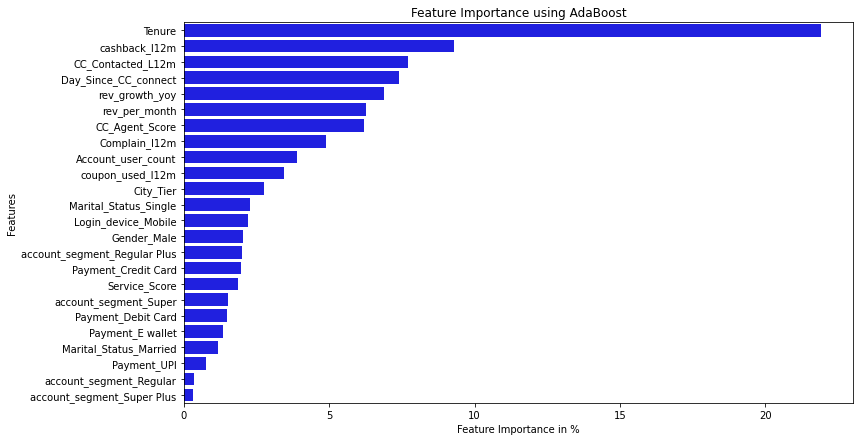


Figure XLVII - Feature Importance after AdaBoost

* Ada Boost has given highest priority to tenure followed by cashback received in last 12 months.

**[AdaBoost] Confusion Matrix –**

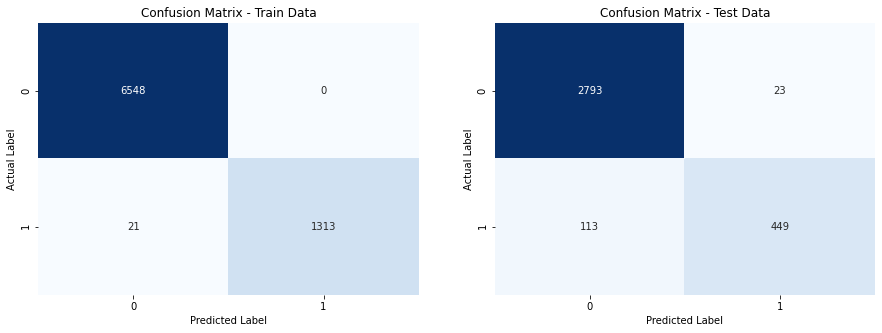


Figure XLVIII - Confusion Matrix of AdaBoost

**[AdaBoost] Classification Reports –**

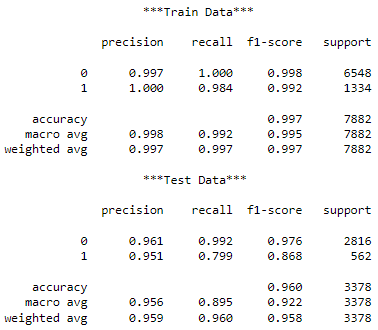


Figure XLIX - Classification Report of AdaBoost

AdaBoost has given us 99.7% on training data and 96% on test data. It has increases tuned model accuracy by 1%, and best model so far.

### **Gradient Boost**

Gradient boosting Regression calculates the difference between the current prediction and the known correct target value. This difference is called residual. After that Gradient boosting Regression trains a weak model that maps features to that residual.

We have created Gradient Boost model with base estimator of Tuned Random Forest model and train weak learners by 0.1 learning rate.

**[Gradient Boost] Confusion Matrix –**

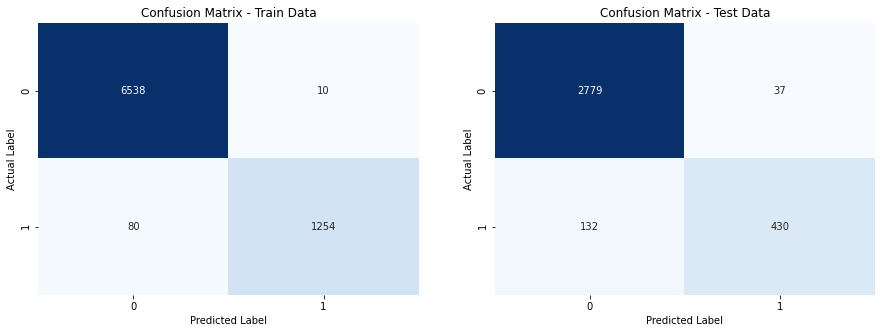


Figure L - Confusion Matrix of Gradient Boost

**[Gradient Boost] Classification Reports –**

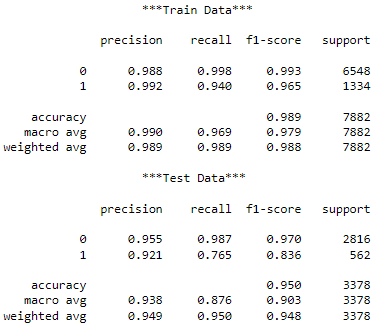


Figure LI - Classification Report after Gradient Boost

Gradient Boost has given us 98.9% accuracy on training data and 95% on testing data. It has slightly improved accuracy on training data but Ada Boost still our best model so far.

### **XG Boost**

XGBoost is a scalable and highly accurate implementation of gradient boosting that pushes the limits of computing power for boosted tree algorithms, being built largely for energizing machine learning model performance and computational speed.

We have created XGBoost model with default parameters.

**[XG Boost] Features Importance –**

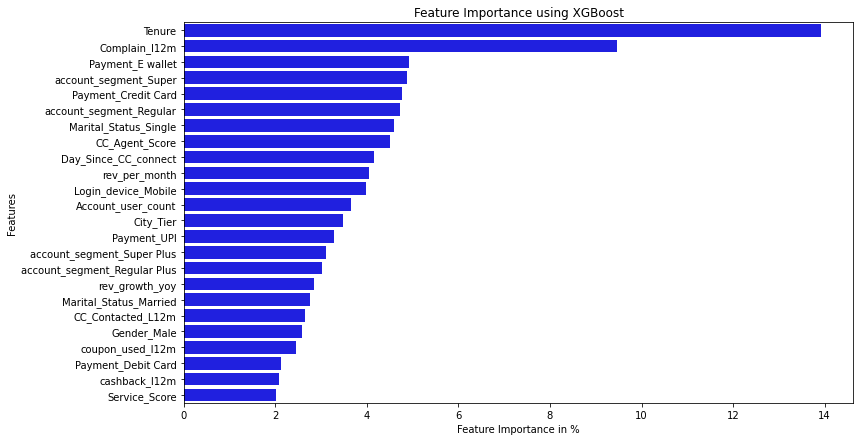


Figure LII - Feature Importance after XGBoost

XGBoost classifier gives Tenure and complain raised in last 12 months features high importance while predicting.

**[XG Boost] Confusion Matrix –**

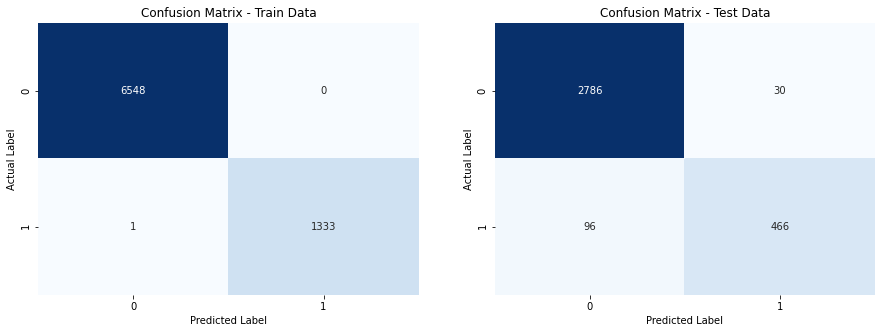


Figure LIII - Confusion Matrix of XGBoost

**[XG Boost] Classification Reports –**

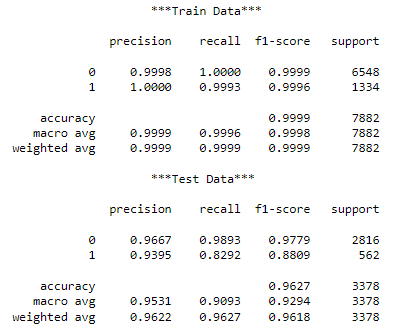


Figure LIV - Classification Report of XGBoost

XG Boost classifier predicts with 96.27% accuracy on test data, which is best among boosting models.

# **Model validation**

Primary metric for evaluation would be **Recall** as our model should have better positive prediction, and secondary metric would be **Precision** for positive class.

Since our data is biased towards non-churn accounts, accuracy would always be on higher side.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | | **Recall** | | **Precision** | |
| **Train Data** | **Test Data** | **Train Data** | **Test Data** | **Train Data** | **Test Data** |
| **Random Forest (Overfitted)** | 100.00% | 97.00% | 100.00% | 100.00% | 100.00% | 97.00% |
| **Random Forest (Tuned)** | 98.00% | 95.00% | 91.00% | 74.00% | 99.00% | 94.00% |
| **Random Forest (SMOTE)** | 99.00% | 94.00% | 98.00% | 80.00% | 97.00% | 82.00% |
| **AdaBoost** | 99.70% | 96.00% | 98.40% | 79.90% | 100.00% | 95.10% |
| **Gradient Boost** | 98.90% | 95.00% | 94.00% | 76.50% | 99.20% | 92.10% |
| **XG Boost** | **99.99%** | **96.27%** | **99.93%** | **82.92%** | **99.90%** | **93.95%** |

Table 8 - Models Comparison

## **Optimum Model**

* XGBoost model would be our best model for prediction that predicts with 96.27% accuracy.
* It has decent recall of **82.92%** and precision **93.95%** on positive class.

# **Final interpretation / recommendation**

## **Insights**

* **Tier-1 Cities** have more E-commerce usage compared to other cities.
* Customers have given **average ratings (3/5)** to company’s services and customer care agents’ resolution.
* Most of the customers are using **Debit or Credit cards** for Payment/Purchase.
* One account is shared among **3 to 5 customers**.
* Most of the customers are using **Mobile Devices** for Shopping.
* **Tier-3** City Customers have **high Churn rate**.
* Customers who have raised **complaints in last 12 months** are likely to Churn.
* **Average Tenure** of non-Churned account is **high**.
* **Single customers** seem to switch E-Commerce companies, shows high churn rate.
* Non-Churned customers seem to receive **high cashback.**

## **Recommendations**

* Company should focus more on **increasing tenure of customers** by giving them offers or better service.
* Company should escalate complains being raised by customers and **expedite assistance**.
* Company should **do surveys more often** to those who raised any complain to see whether they were satisfied with company’s services or not
* Try providing quick support and better service to **Single** customers, by routing calls to agent giving satisfactory assistance to them.
* Offer better services in **Tier-3** Cities.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* END \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*