CAPSTONE PROJECT
Final Report

Report by
HARI HARAN

01st Oct, 2023

Great Learning

`

`

INTRODUCTION

a) Understanding the Problem Statement :

An E Commerce company is facing a lot of competition in the current market and it has become a challenge to retain the existing customers in the current situation. Hence, the company wants us 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 a single account can have multiple users. So if they loose one account then the company may be loose more than one customer. This can adversely affect the company's revenue stream.

We have been assigned to develop a churn prediction model for this company and provide business recommendations on the campaign. Our suggestions should be very crisp & clear on the campaign offer. We have to provide best possible solutions and recommendations to retain their customer base which can be verified by their revenue assurance team. Our report must have recommendations and insights must suggest sound improvements, if any, to retain their existing customer base.

b) Need of the study/project :

- Any business thrives on the revenue generation and predicting the future of the business and it's dimension and direction with maximum accuracy. But like any other business to achieve those, they need a plan to retain their customer base.
- In order to maintain their customer and in-turn the revenue flow, this study/project is important for the current client to plan their business's direction ahead future in terms of their services, products in different segments, customer care performance, customer-centric issues, revenue generation, various offers (if needed), etc .
- The outcome of this study/project will give us an understanding of company's current scenario and its future along with the risks it may face in upcoming years.
- This study/project will generate some insights and shade light on various aspects of the client's business which will help them to minimise any risk and increase accuracy to achieve maximum efficiency.

c) Understanding business :

In this case study/project ,

- Each customers has been assigned with an unique Account ID and any single ID may many customers (like family or friends) with different genders and marital-status.
- Customers accounts are segmented depending on the tenure of their account, city tier, based on the device they use, basis on their spending .
- Business runs on retaining their loyal customers with value-added services and prompt customer service
- Customers do have multiple different payment mode options, which is user friendly.
- The various promotional offers or cashback offers also helps to retain customer and may attract new ones.
- Customer feedback on their recent purchases or cashback or interaction with customer-care helps to understand their customer's perspective. Positive and negative feedback both gives clarity.

d) About the given data

Understandings from the given data,

1. The given data is collected for 11260 different unique account ID.
2. The given data has 19 variables(or features) with one variable ‘ Churn’ as the dependant variable and others independent variables.
3. The data has been collected for 12 months for each of the 11260 account IDs.
4. The customers data is based on their account tenure, account_segment, Login_device, gender and marital_status.
5. There are various payment modes.
6. The data also contain customer ratings on the basis of customer care services, ratings based on the services provided by the customer, customer contact and complains raised and over a period of 12 months.
7. The data on various payment modes is also available along with any coupons or cashback, which suggests the company provides offer to their customers.
8. rev_per_month shows the monthly revenue generation based on the spending of each account
9. rev_growth_yoy shows yearly growth in revenue based on these accounts.

Variable	Description
AccountID	account unique identifier
Churn	account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_LY	How many times all the customers of the account has contacted customer care in last 12months
Payment	Preferred Payment mode of the customers in the account
Gender	Gender of the primary customer of the account
Service_Score	Satisfaction score given by customers of the account on service provided by company
Account_user_count	Number of customers tagged with this account
account_segment	Account segmentation on the basis of spend
CC_Agent_Score	Satisfaction score given by customers of the account on customer care service provided by company
Marital_Status	Marital status of the primary customer of the account
rev_per_month	Monthly average revenue generated by account in last 12 months
Complain_ly	Any complaints has been raised by account in last 12 months
rev_growth_yoy	revenue growth percentage of the account (last 12 months vs last 24 to 13 month)
coupon_used_l12m	How many times customers have used coupons to do the payment in last 12 months
Day_Since_CC_connect	Number of days since no customers in the account has contacted the customer care
cashback	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

Table1 : Data Dictionary

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

Data info: Shape: 11260(rows) x 90 (columns)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11260 entries, 0 to 11259
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Churn                    11260 non-null  int64
 1   Tenure                   11158 non-null  object
 2   City_Tier                11148 non-null  float64
 3   CC_Contacted_LY          11158 non-null  float64
 4   Payment                  11151 non-null  object
 5   Gender                   11152 non-null  object
 6   Service_Score            11162 non-null  float64
 7   Account_user_count       11148 non-null  object
 8   account_segment          11163 non-null  object
 9   CC_Agent_Score           11144 non-null  float64
 10  Marital_Status           11048 non-null  object
 11  rev_per_month            11158 non-null  object
 12  Complain_ly              10903 non-null  float64
 13  rev_growth_yoy           11260 non-null  object
 14  coupon_used_for_payment  11260 non-null  object
 15  Day_Since_CC_connect     10903 non-null  object
 16  cashback                 10789 non-null  object
 17  Login_device             11039 non-null  object
dtypes: float64(5), int64(1), object(12)
memory usage: 1.5+ MB
```

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

Table 2 : Data Info and Null Values

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Churn	11260.0	NaN	NaN	NaN	0.168384	0.374223	0.0	0.0	0.0	0.0	1.0
Tenure	11158.0	38.0	1.0	1351.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
City_Tier	11148.0	NaN	NaN	NaN	1.653929	0.915015	1.0	1.0	1.0	3.0	3.0
CC_Contacted_LY	11158.0	NaN	NaN	NaN	17.867091	8.853269	4.0	11.0	16.0	23.0	132.0
Payment	11151	5	Debit Card	4587	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	11152	4	Male	6328	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Service_Score	11162.0	NaN	NaN	NaN	2.902526	0.725584	0.0	2.0	3.0	3.0	5.0
Account_user_count	11148.0	7.0	4.0	4569.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
account_segment	11163	7	Super	4062	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CC_Agent_Score	11144.0	NaN	NaN	NaN	3.066493	1.379772	1.0	2.0	3.0	4.0	5.0
Marital_Status	11048	3	Married	5860	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rev_per_month	11158.0	59.0	3.0	1746.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Complain_ly	10903.0	NaN	NaN	NaN	0.285334	0.451594	0.0	0.0	0.0	1.0	1.0
rev_growth_yoy	11260.0	20.0	14.0	1524.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
coupon_used_for_payment	11260.0	20.0	1.0	4373.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Day_Since_CC_connect	10903.0	24.0	3.0	1816.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cashback	10789.0	5693.0	155.62	10.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Login_device	11039	3	Mobile	7482	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 3 : Data Description

Note : We have dropped "AccountID"

a) Understanding of attributes (variable info, renaming if required):

From Table 1 and Table 2:

- Data types :

Float type	5 variables (Continuous independent variable)
Obejct type	12 variables (Categorical independent variable)
Integer type	1 variable (**Churn** Continuous dependent variable)

- **Table 2** also shows ***null values*** present in the data for each variable. These null variables needs to be treated for EDA.
- Renaming is not required at its pretty clear from the data dictionary about the description of each variable.
- There are **259 duplicate values**. These can be ignored as there can be variables with similar values like same city_tier or tenure.
- Some of the categorical and continous variables have bad or missing values, which needs to be treated before balancing the model and model-building approach.

b) There are two types of variables in the given data :

1. Categorical Variables :

- 'Payment'
- 'Gender'
- 'account_segment'
- 'Marital_Status'
- 'Login_device'

2. Continous Variables :

- 'Churn'
- 'Tenure'
- 'City_Tier'
- 'CC_Contacted_LY'
- 'Service_Score'
- 'Account_user_count'
- 'CC_Agent_Score'
- 'rev_per_month'
- 'Complain_ly'
- 'rev_growth_yoy'
- 'coupon_used_for_payment'
- 'Day_Since_CC_connect'
- 'cashback'

EDA & Business Implications

Univariate Analysis ,

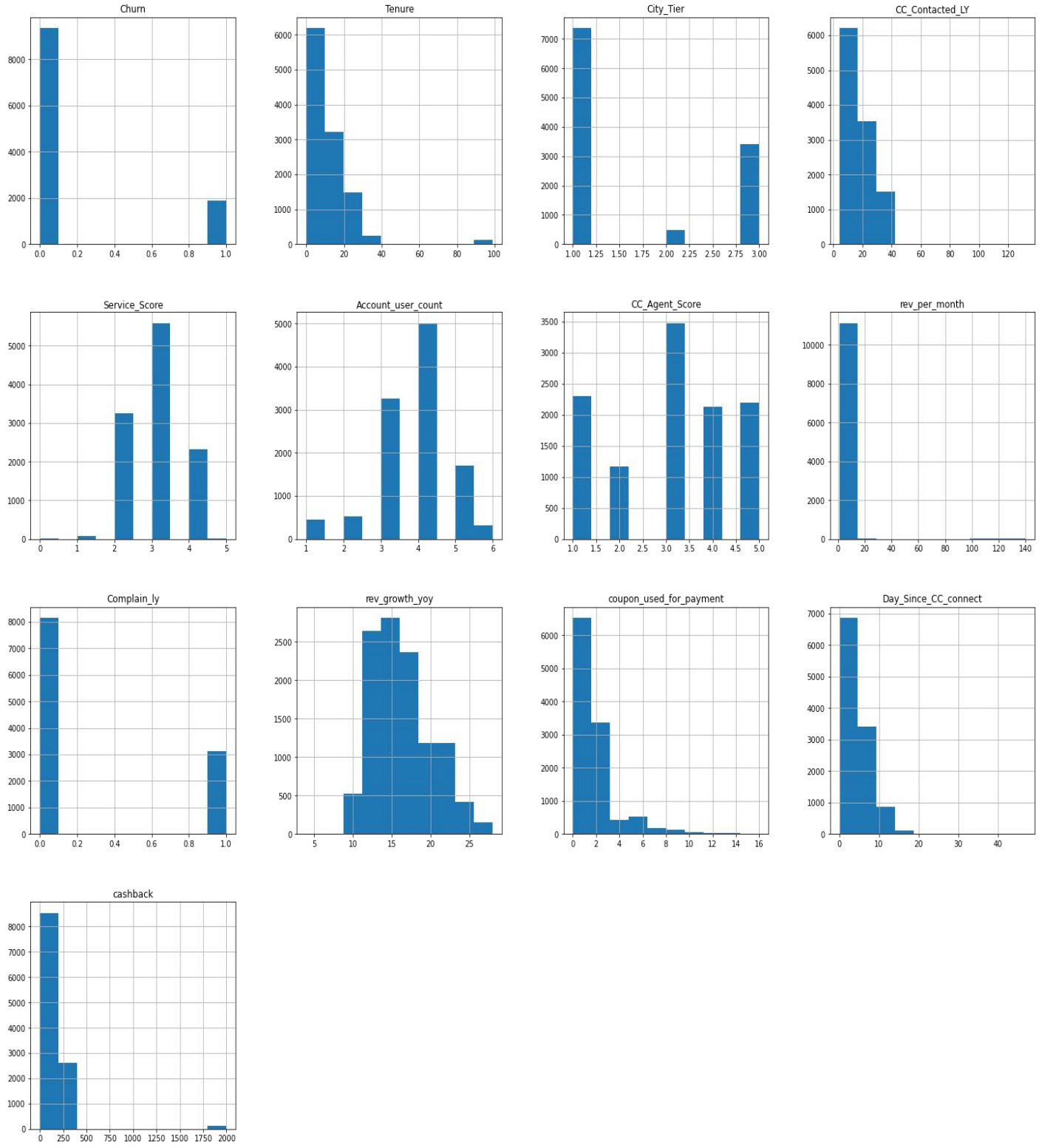

Fig.1 - Histogram for all Continuous Variables

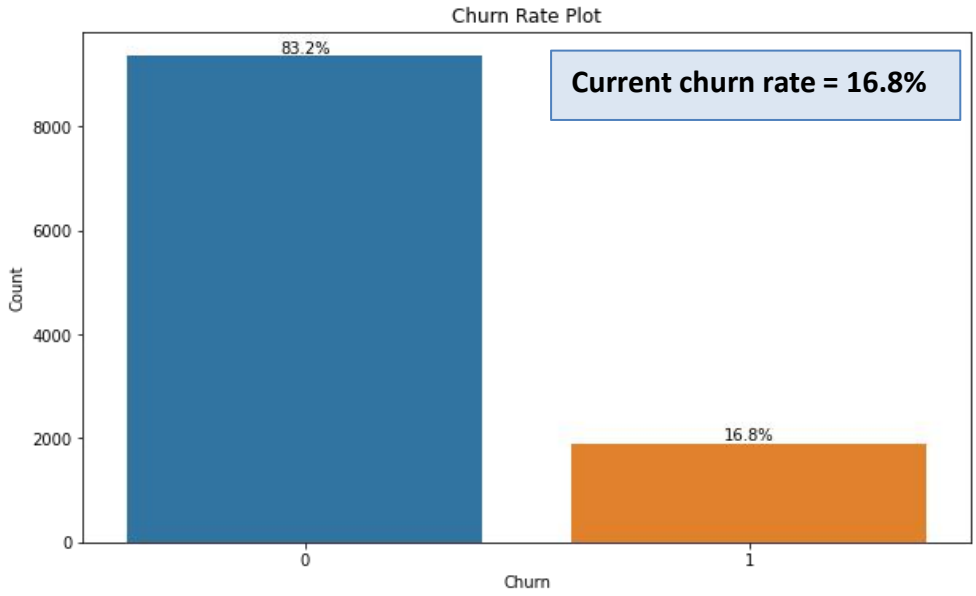

Fig.2 - Churn Rate

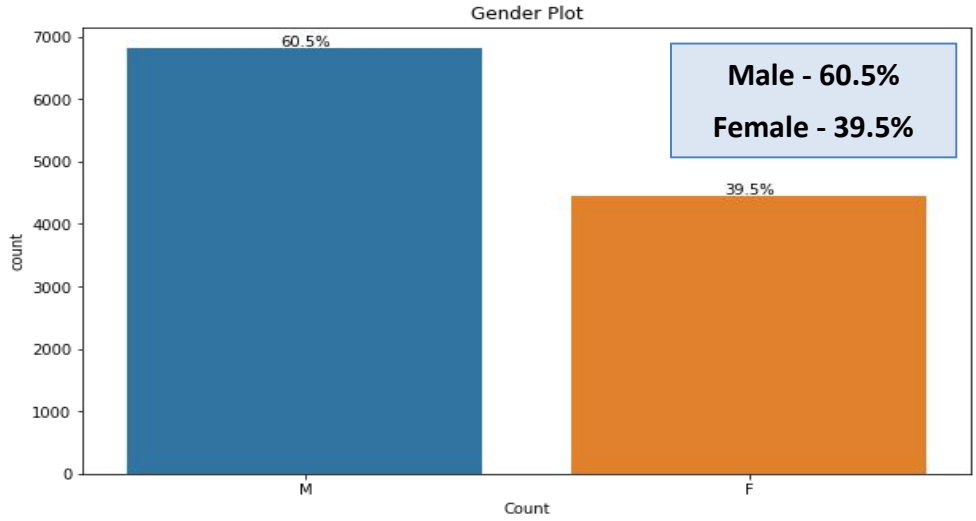

Fig.3 - Gender Plot

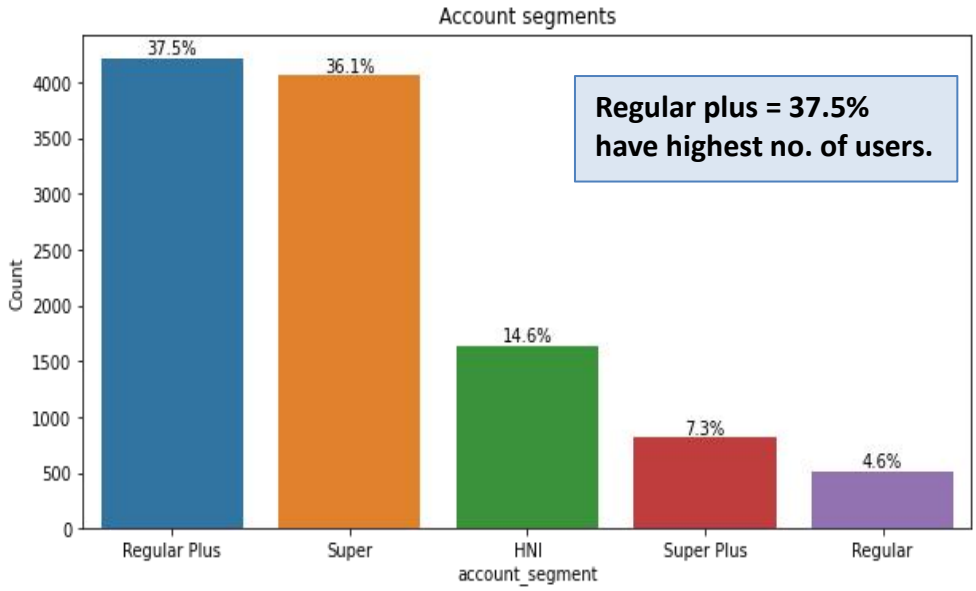

Fig.4 - Account Segments

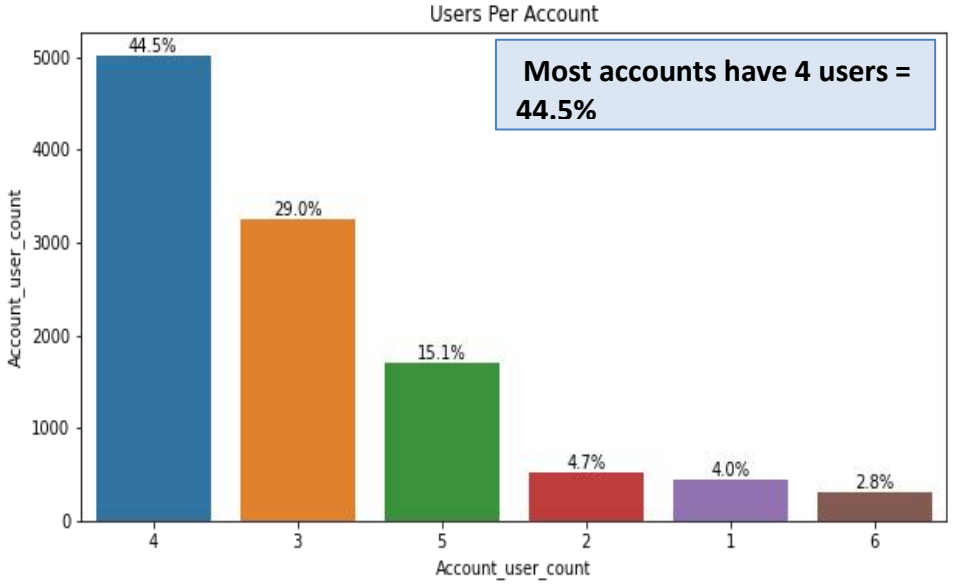

Fig.5 - Users per Account

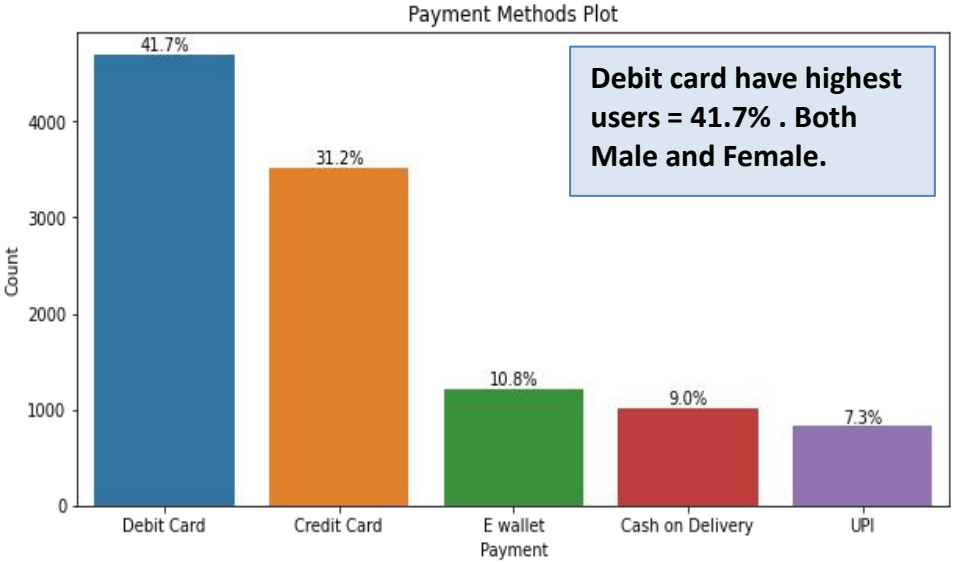

Fig.6 - Payment Methods

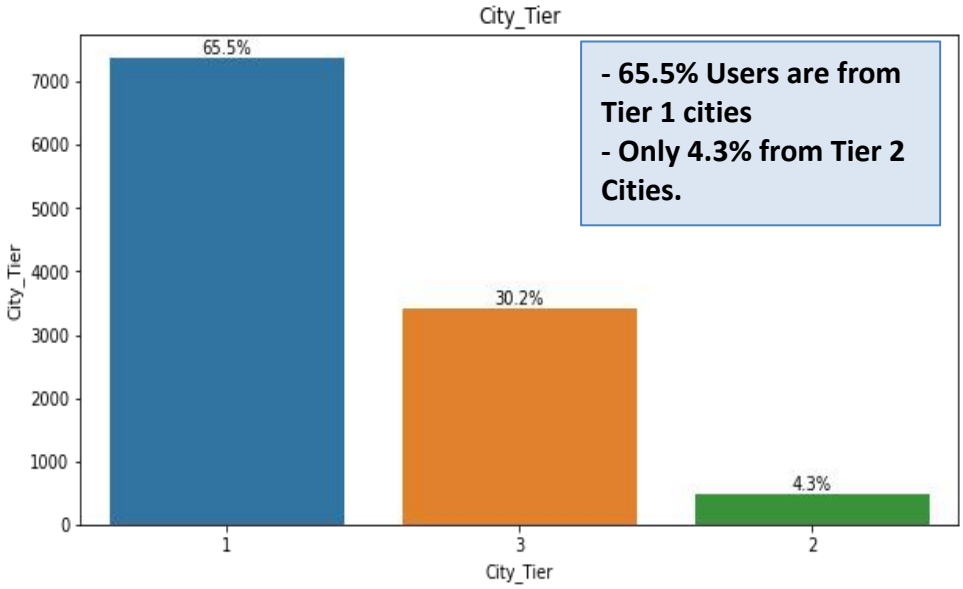

Fig.7 - City Tiers

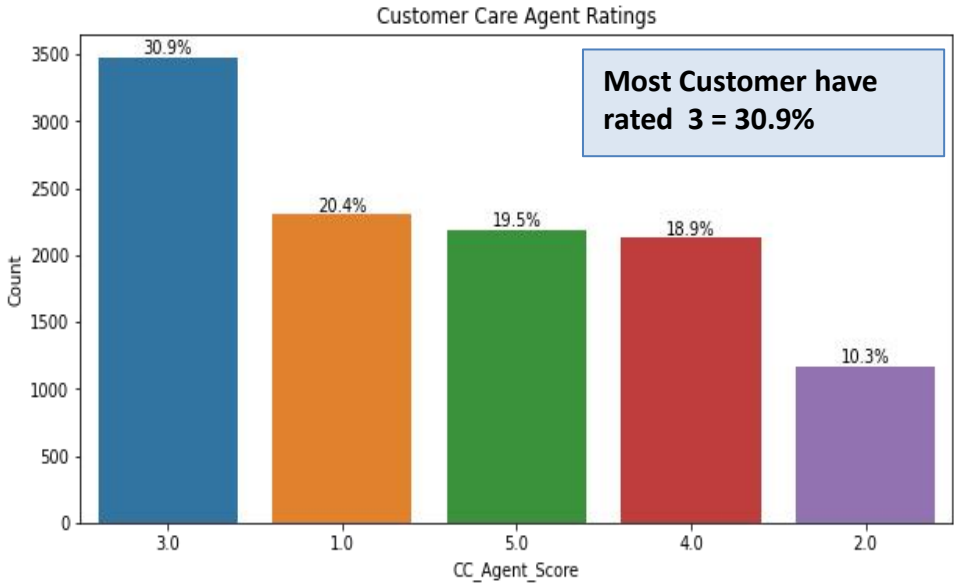

Fig.8 - Customer Care Agents Ratings

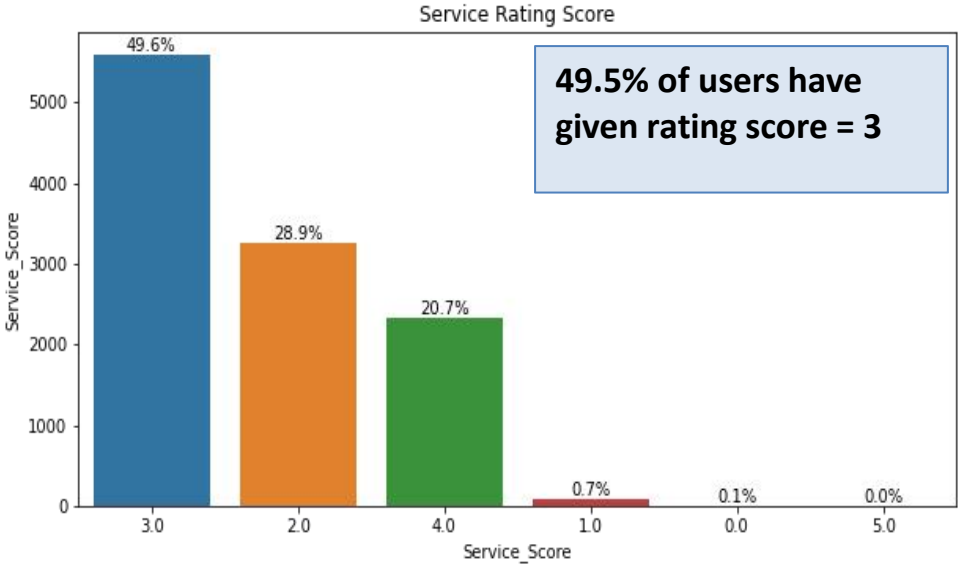

Fig.9 - Service Ratings Score

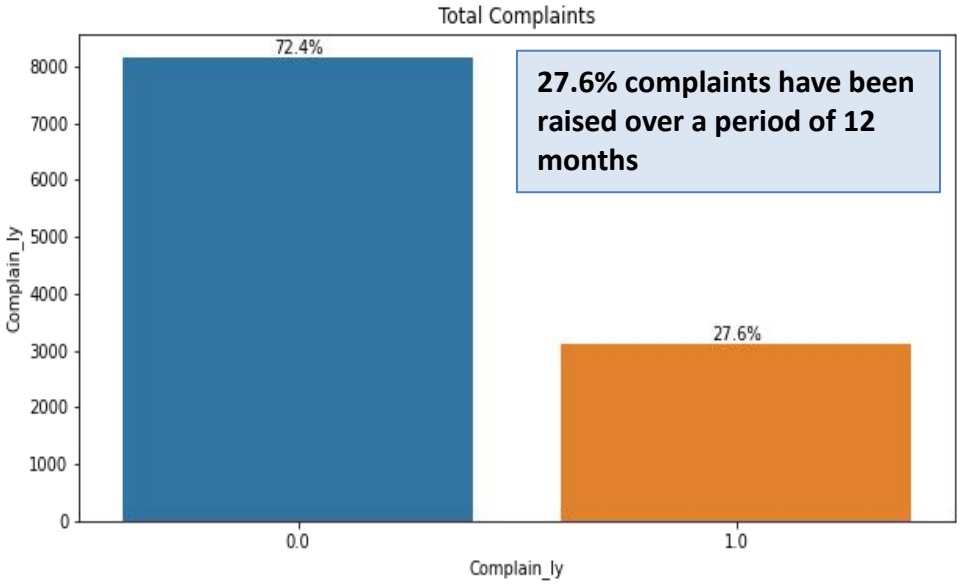

Fig.10 - Total Complaints registered

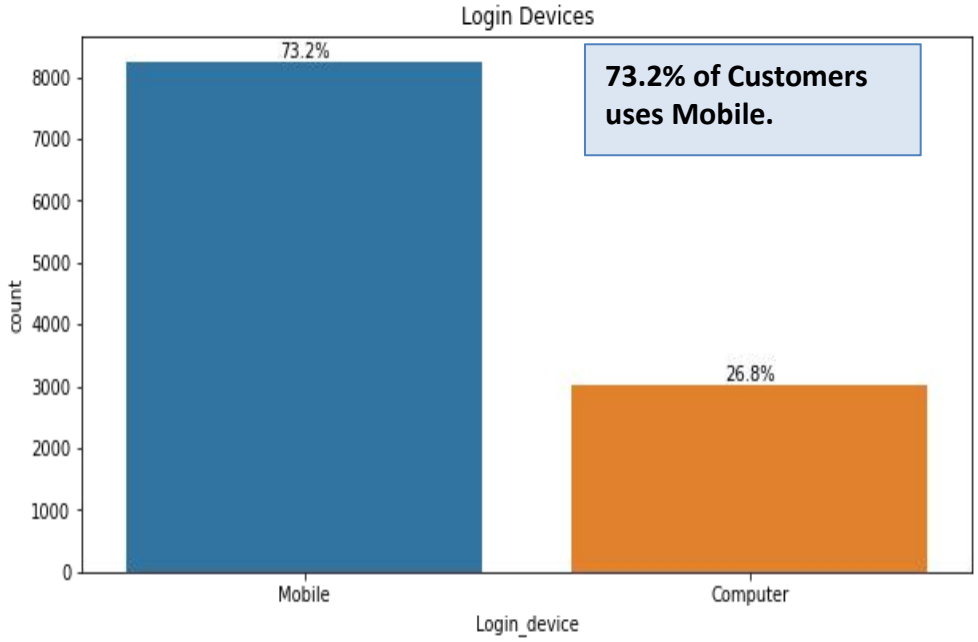

Fig.11 - Login Devices Used

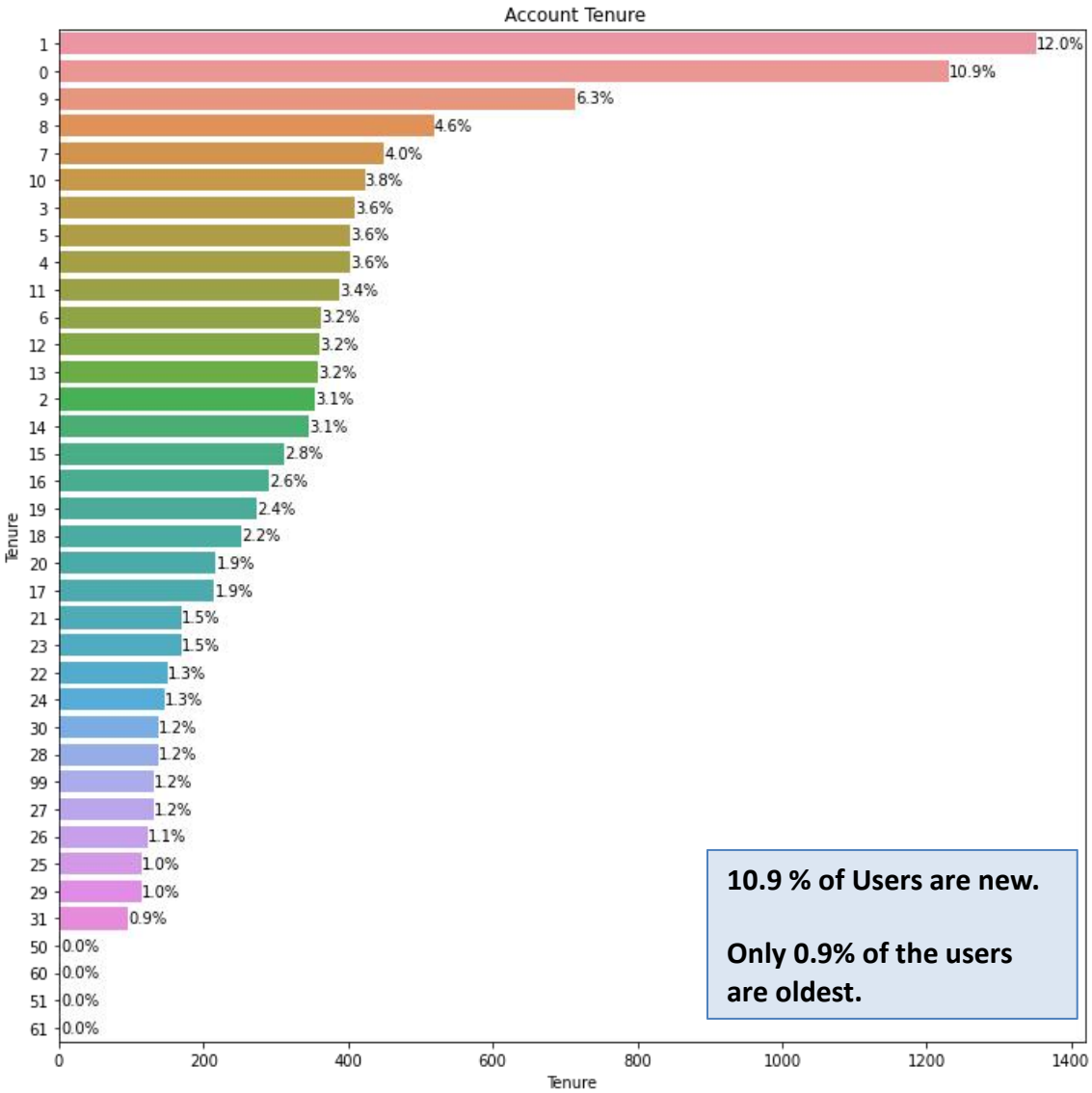

Fig.12 - Account tenures

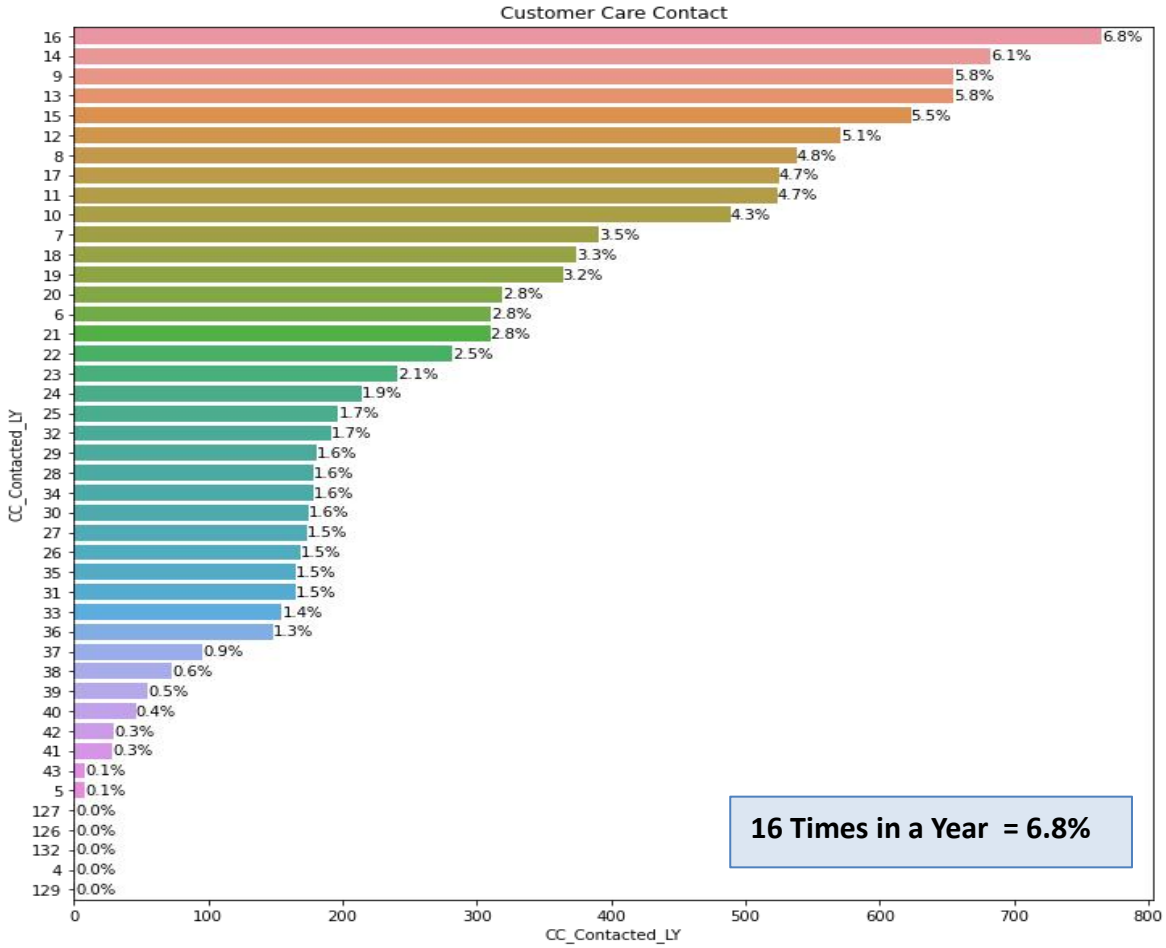

Fig.13 - Customer Care Connect

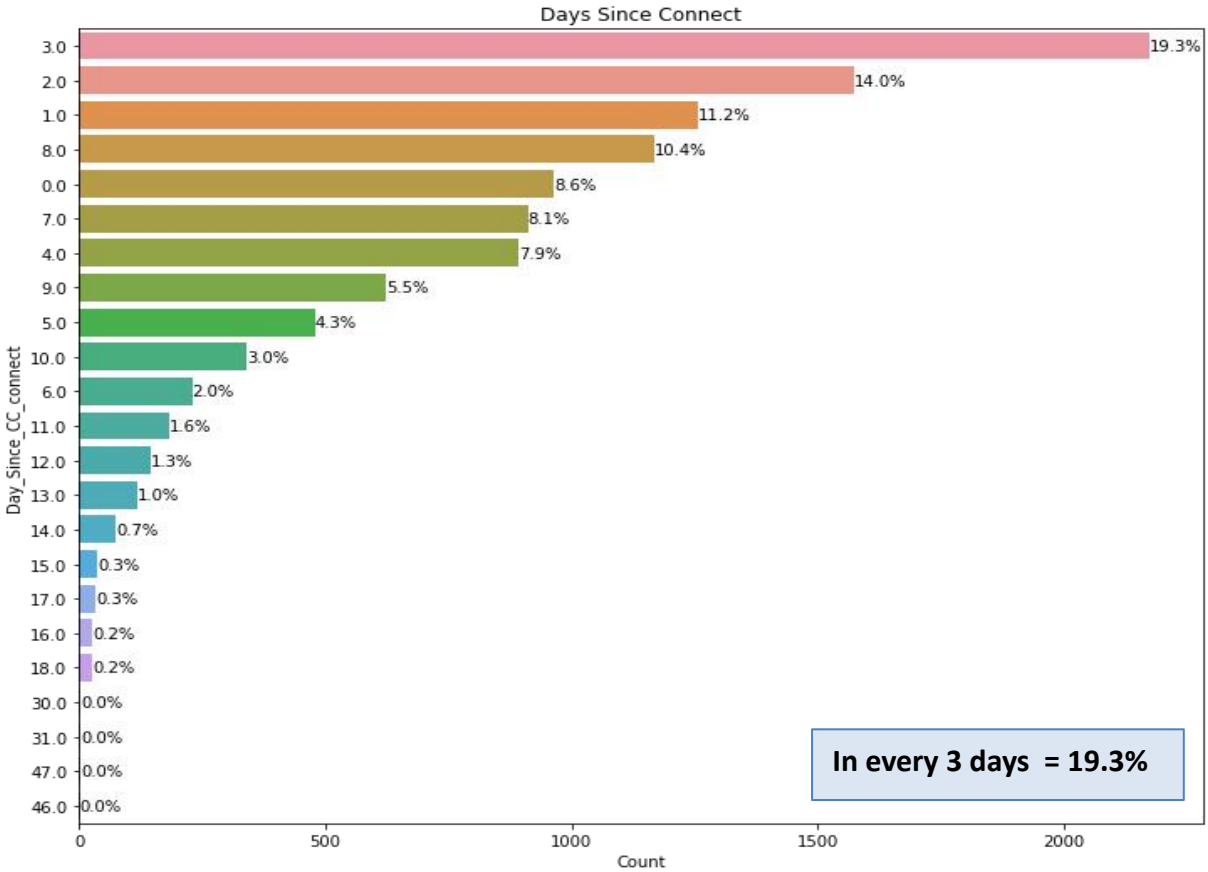

Fig.14 - Customer Care Connect Frequency

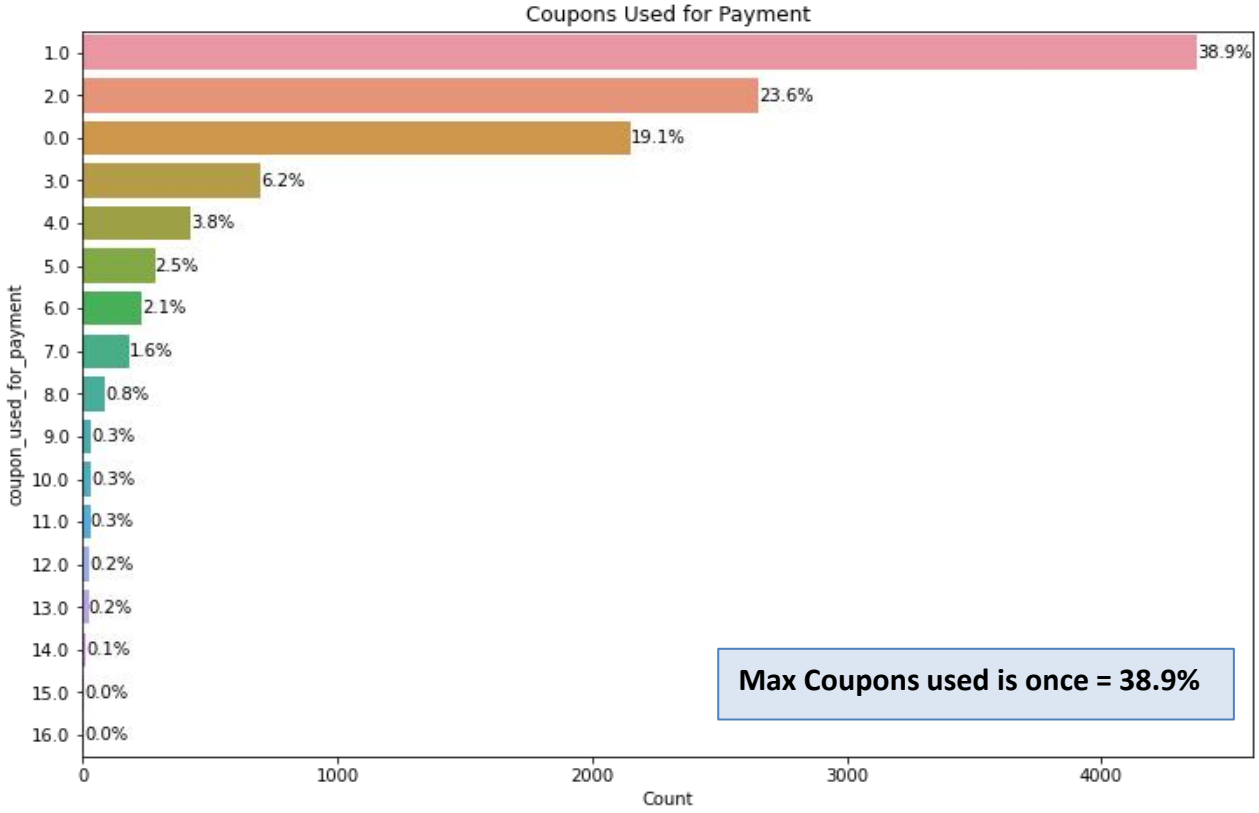

Fig.15 - Coupons Used for Payment

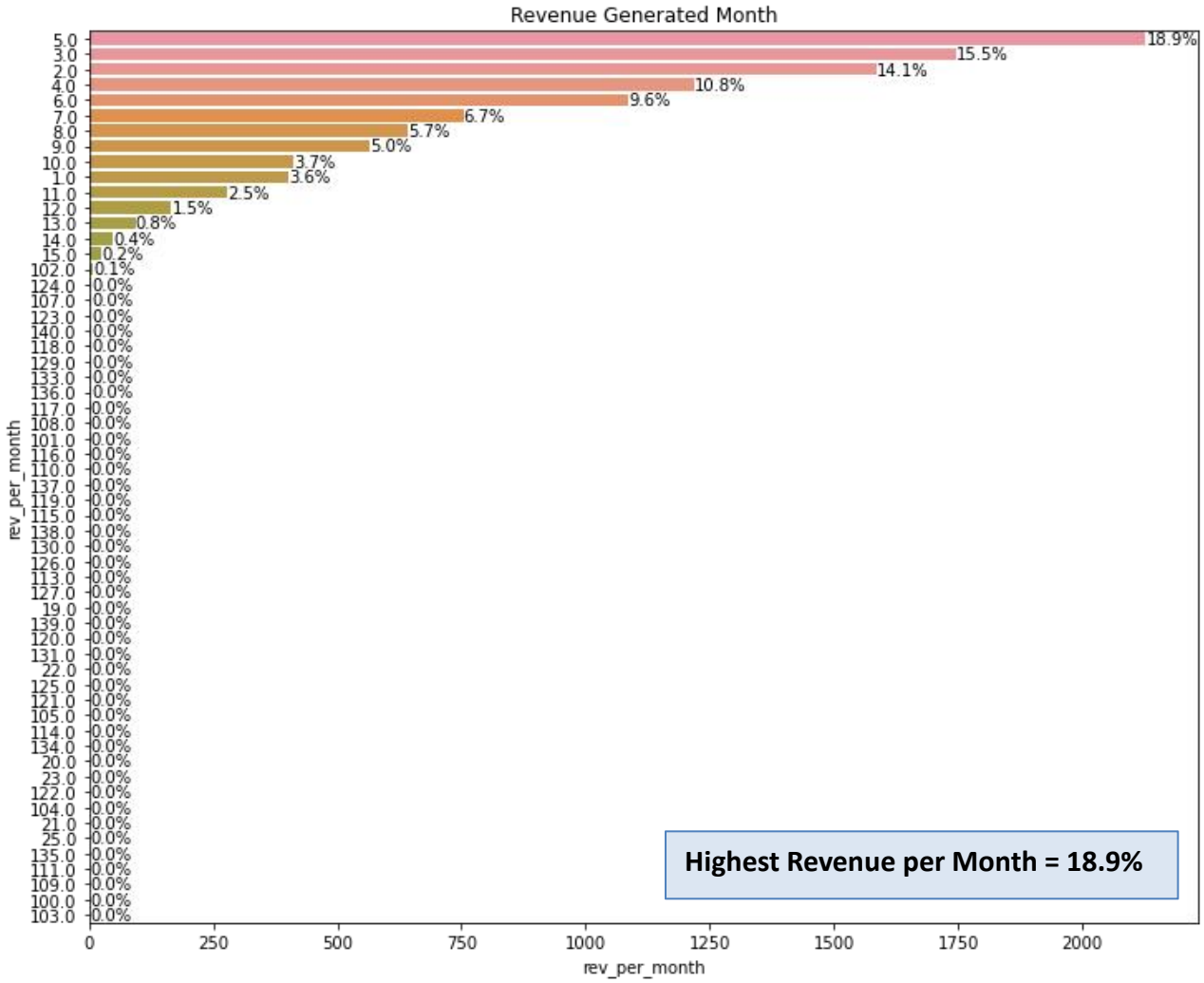

Fig.16 - Revenue Growth Per month

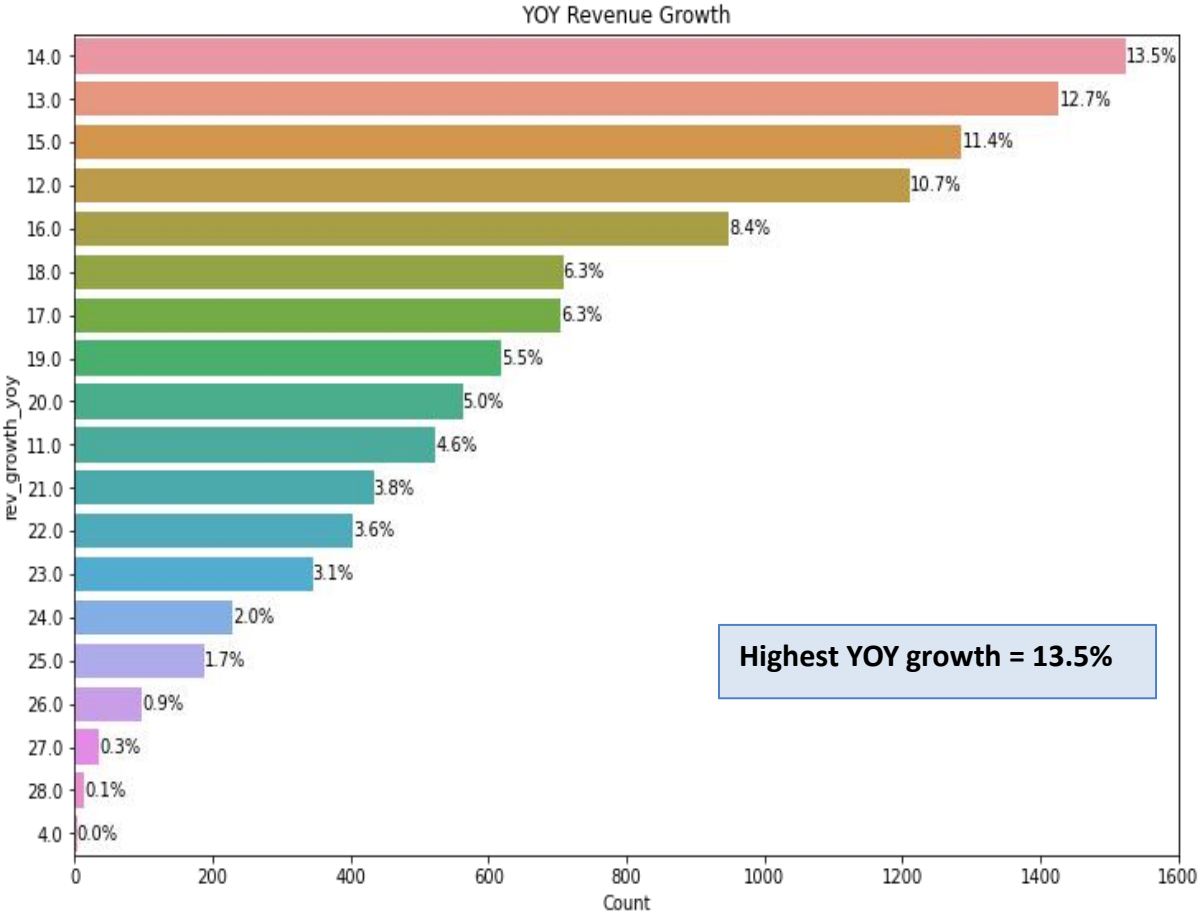

Fig.17 - Yearly Revenue Growth (YOY)

Bivariate Analysis,

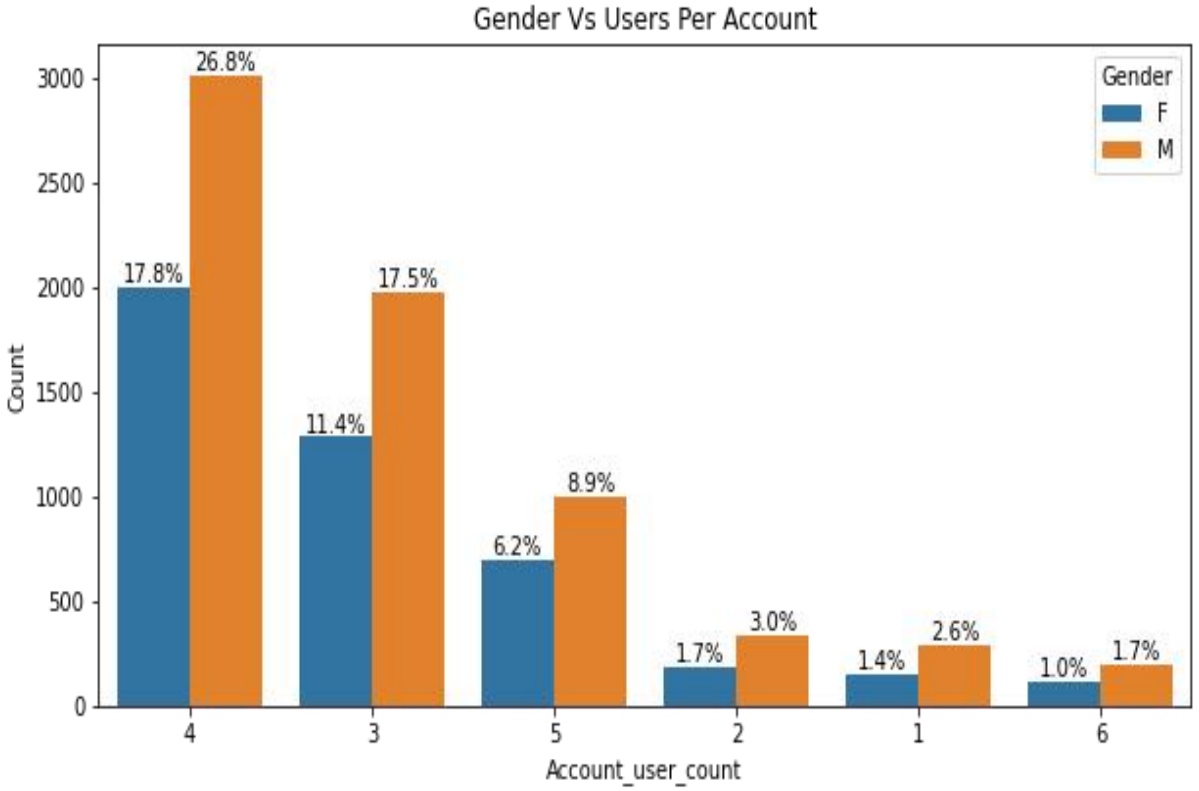

Fig.18 - Gender vs Account Users.

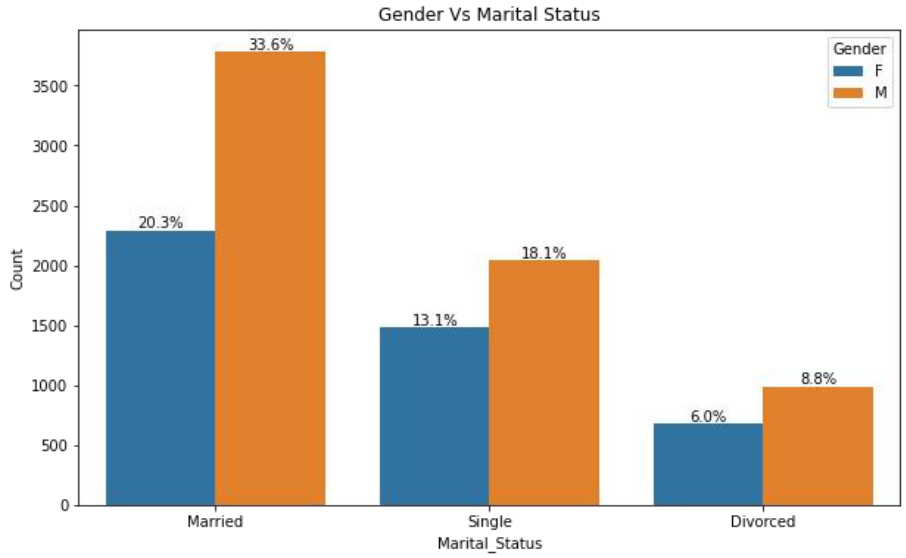

Fig.19 - Gender vs Marital Status

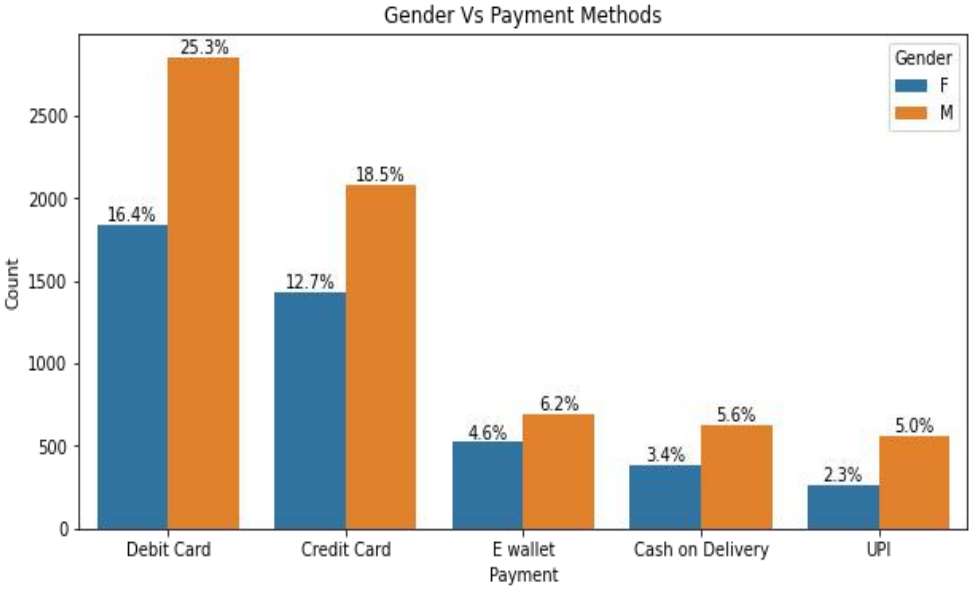

Fig.20 - Gender vs Payment Methods

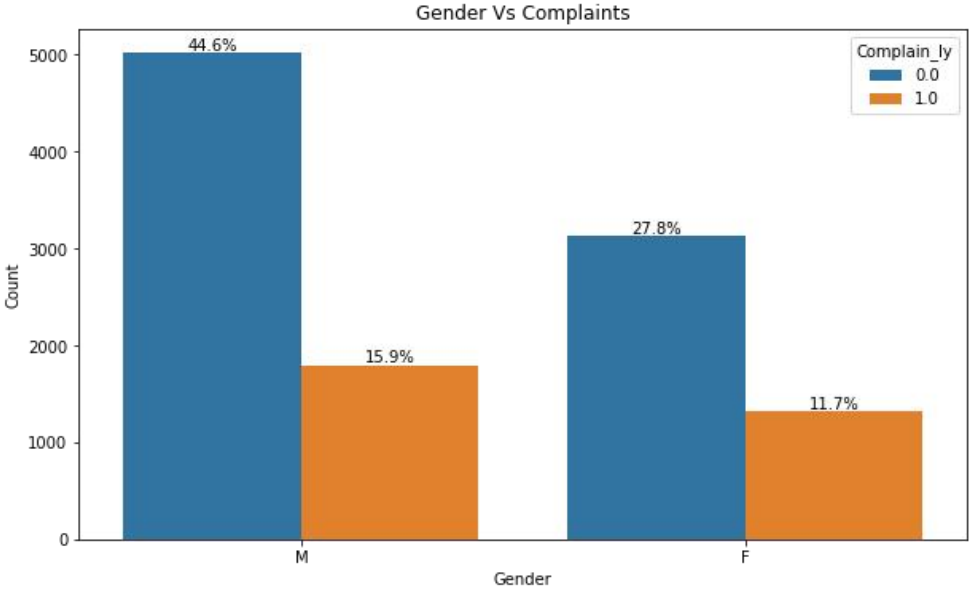

Fig.21 - Gender vs Complaints

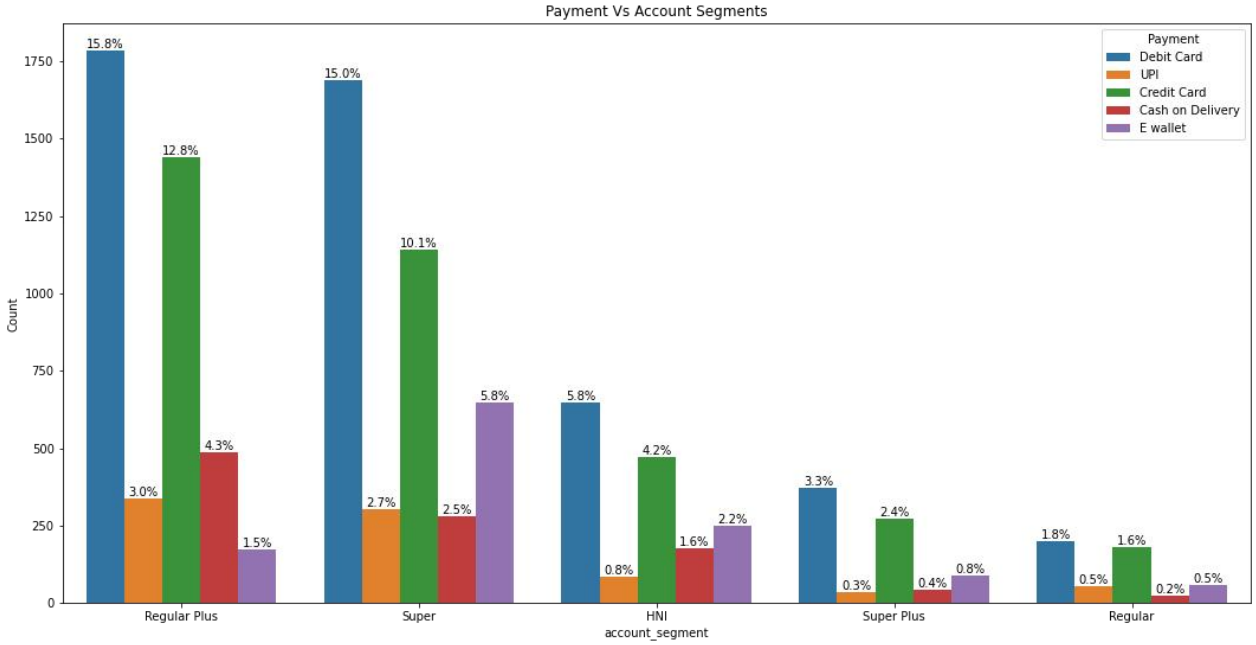

Fig.22 - Payment vs Account Segments

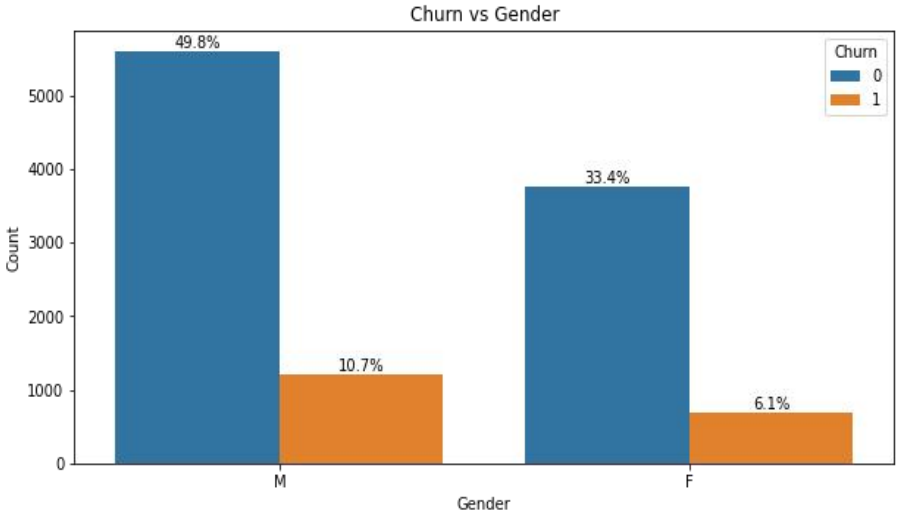

Fig.23 - Churn Vs Gender

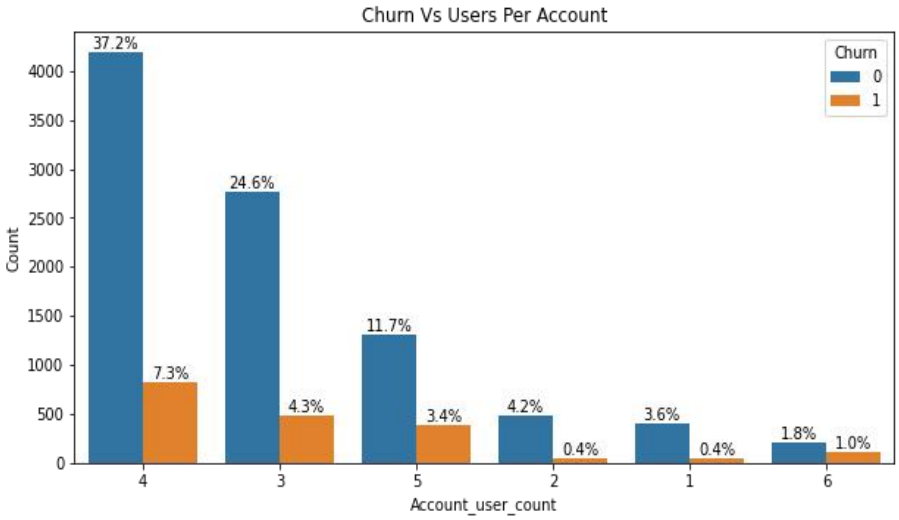

Fig.24 - Churn Vs Users Per Account

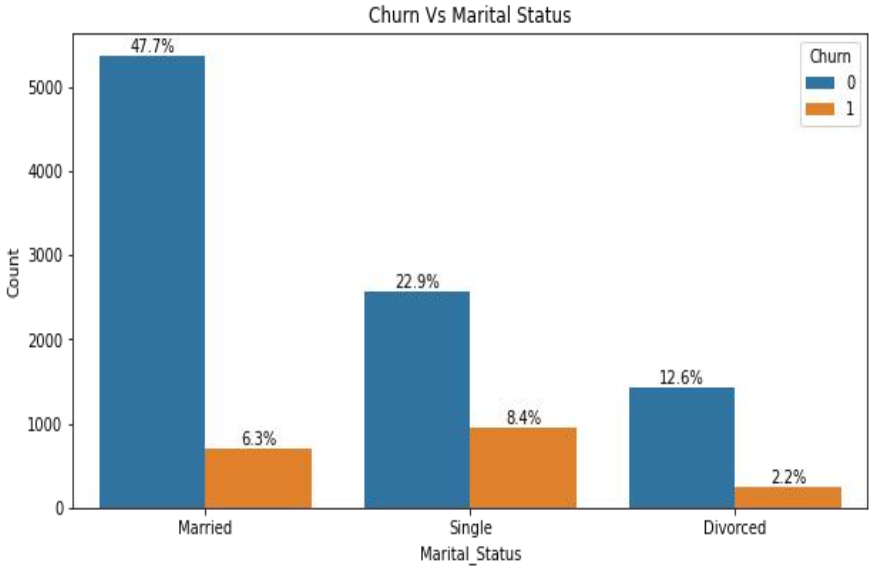

Fig.25 - Churn Vs Marital Status

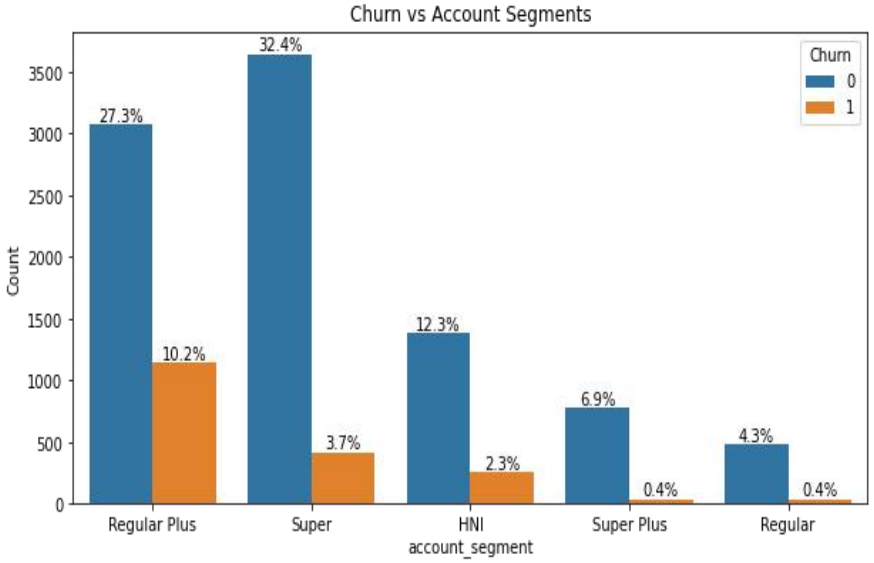

Fig.26 - Churn Vs Account Segments

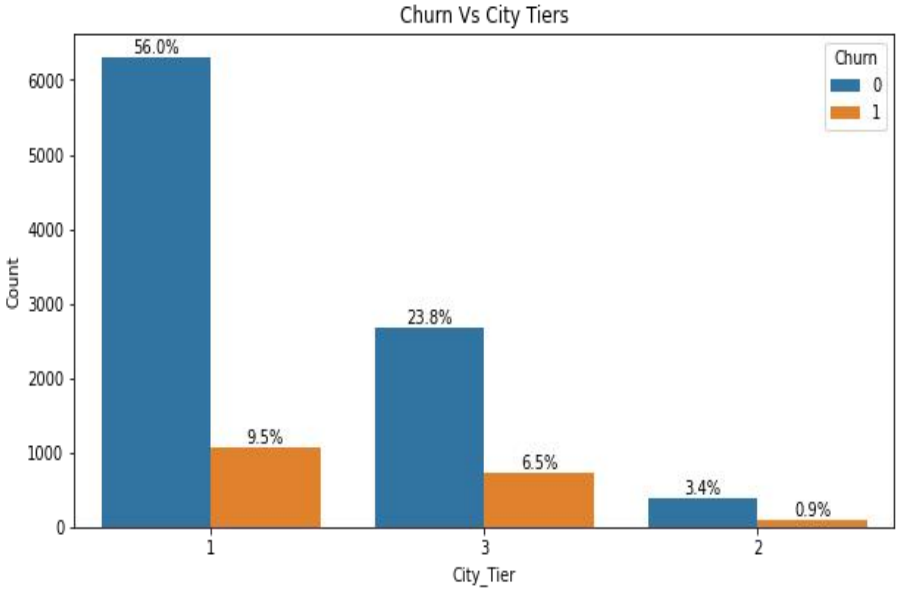

Fig.27 - Churn Vs City Tiers

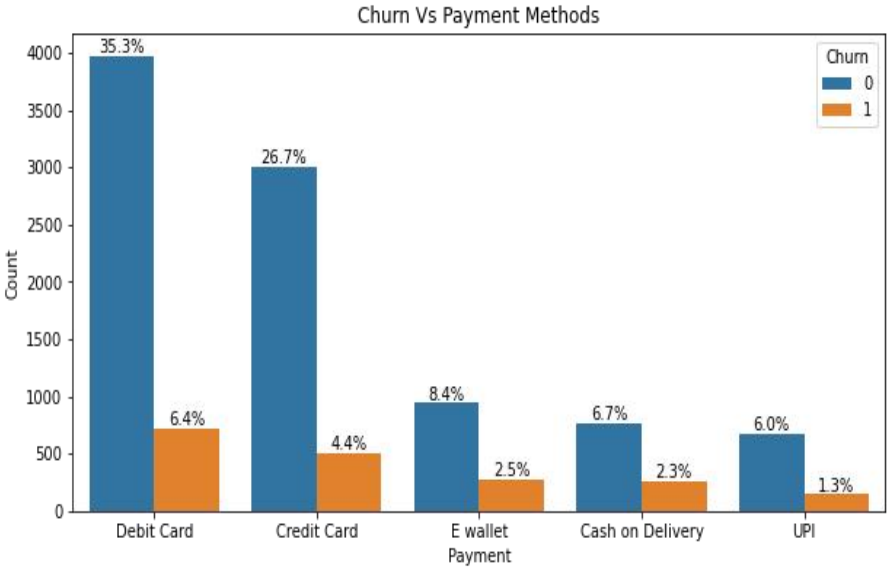

Fig.28 - Churn Vs Payment Methods

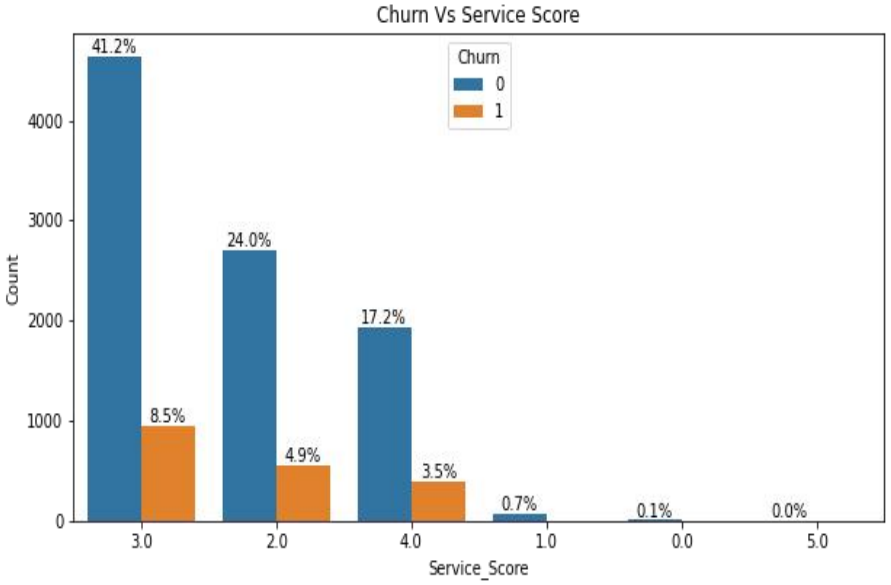

Fig.29 - Churn Vs Service Score

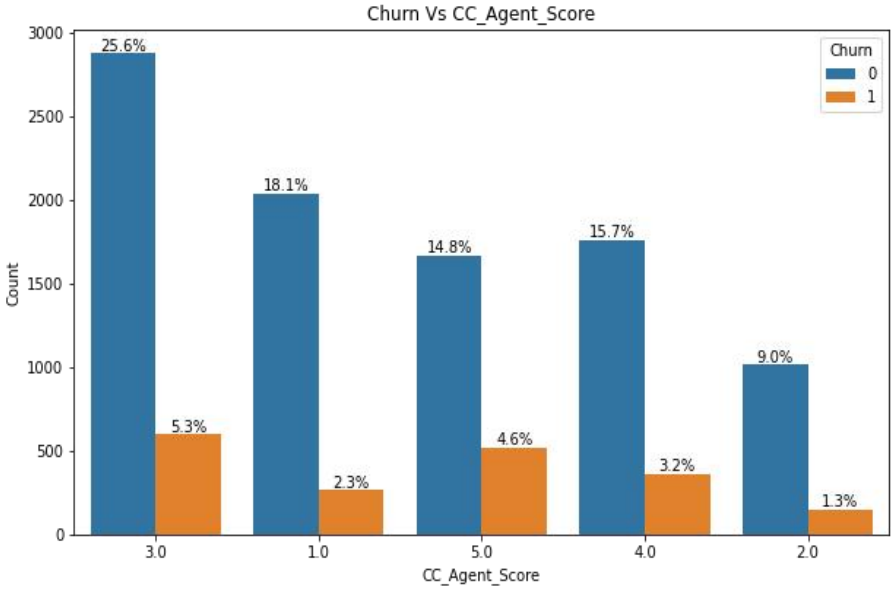

Fig.30 - Churn Vs CC Agent Score

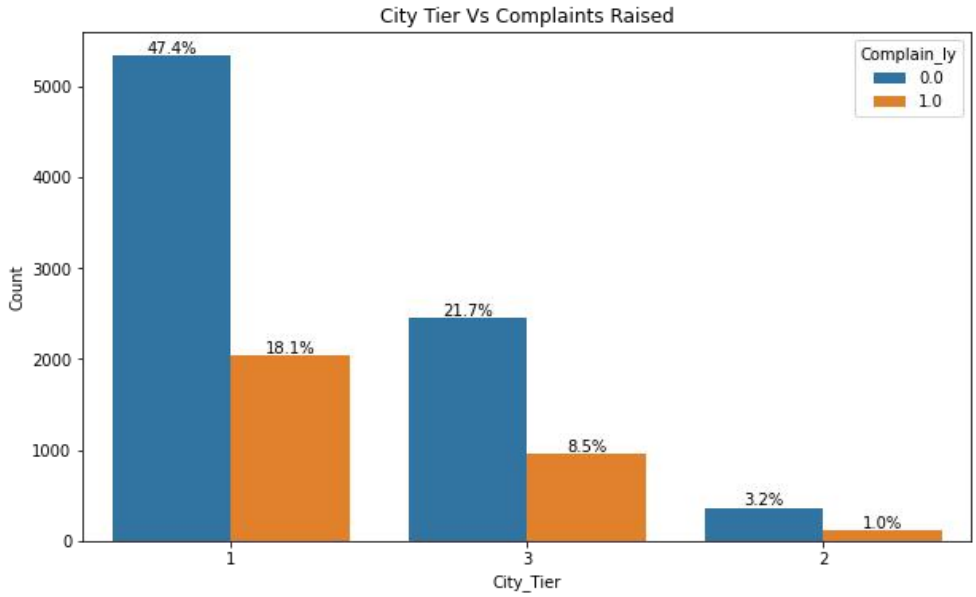

Fig.31 - City tier Vs Complaints Raised

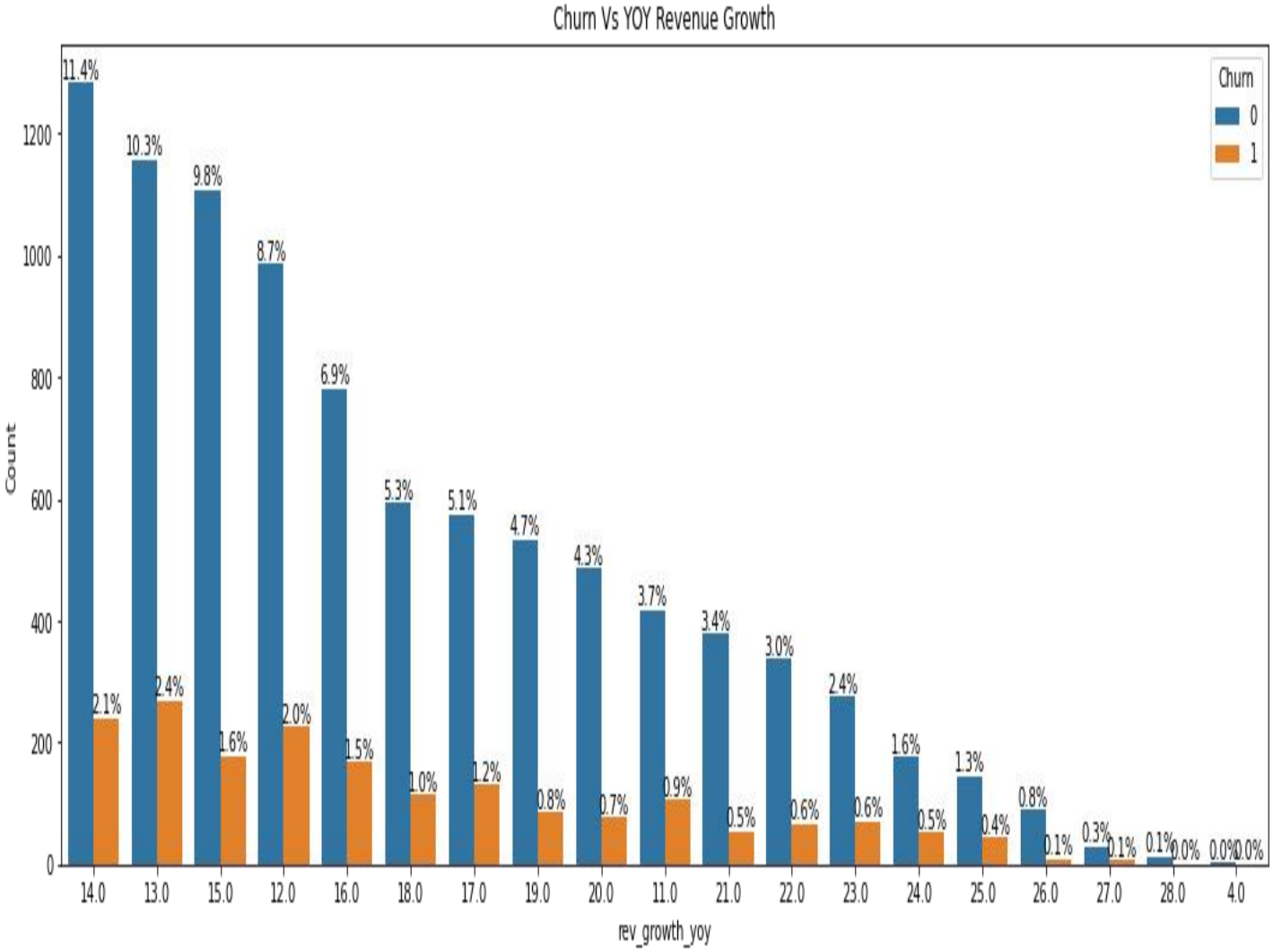

Fig.32 - Churn Vs Revenue Growth (YOY)

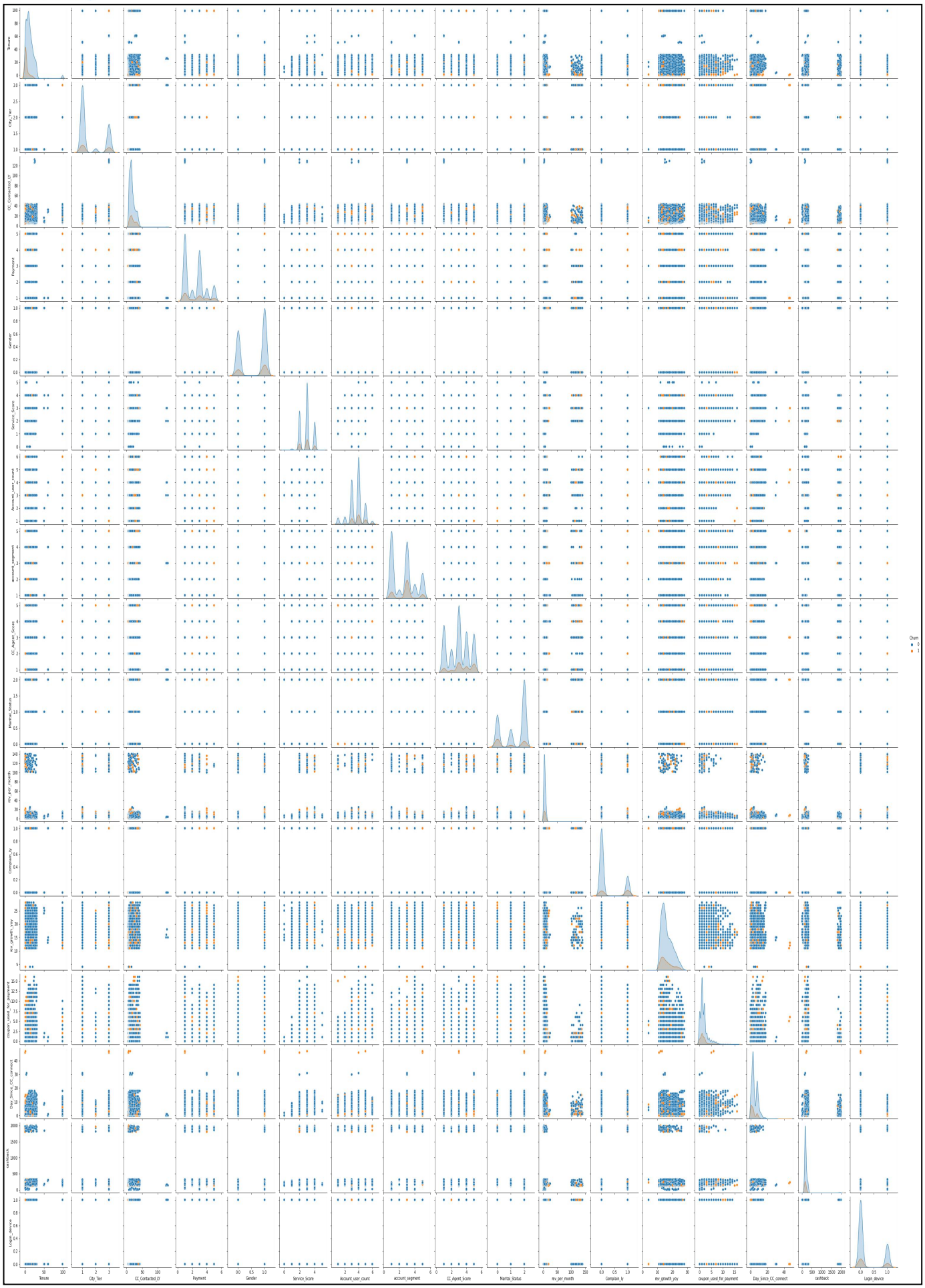

Fig.33 - Pair Plot - Churn (hue)

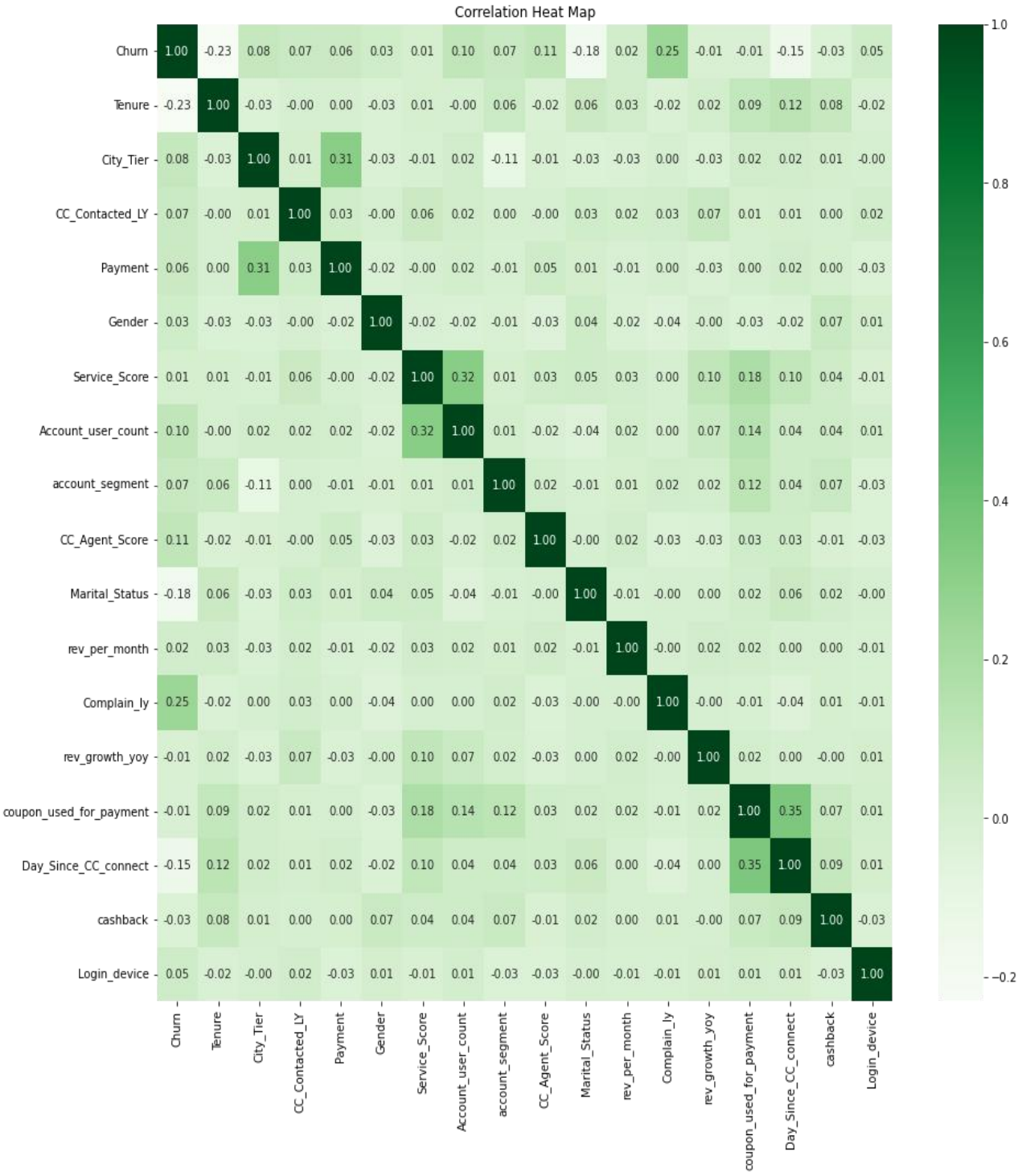

Fig.34 - Correlation Heat Map

From correlation heat map,

1. **Negative Correlation with Churn :**Tenure, Marital status, Revenue growth YOY , Coupons used for payment , Cash Back, Days_since_CC_connect

2. **Positive Correlations with Churn :** City Tier, CC_Contact - Customer Care contact, Payment, Gender, Service Score, Account_user_count,Account Segment,CC_agent_score, rev_per_month,Complain_ly, Login_device

The **negative correlation** indicates it will effect the churn rate positively. And **positive correlation** indicates that it will decrease the churn rate.

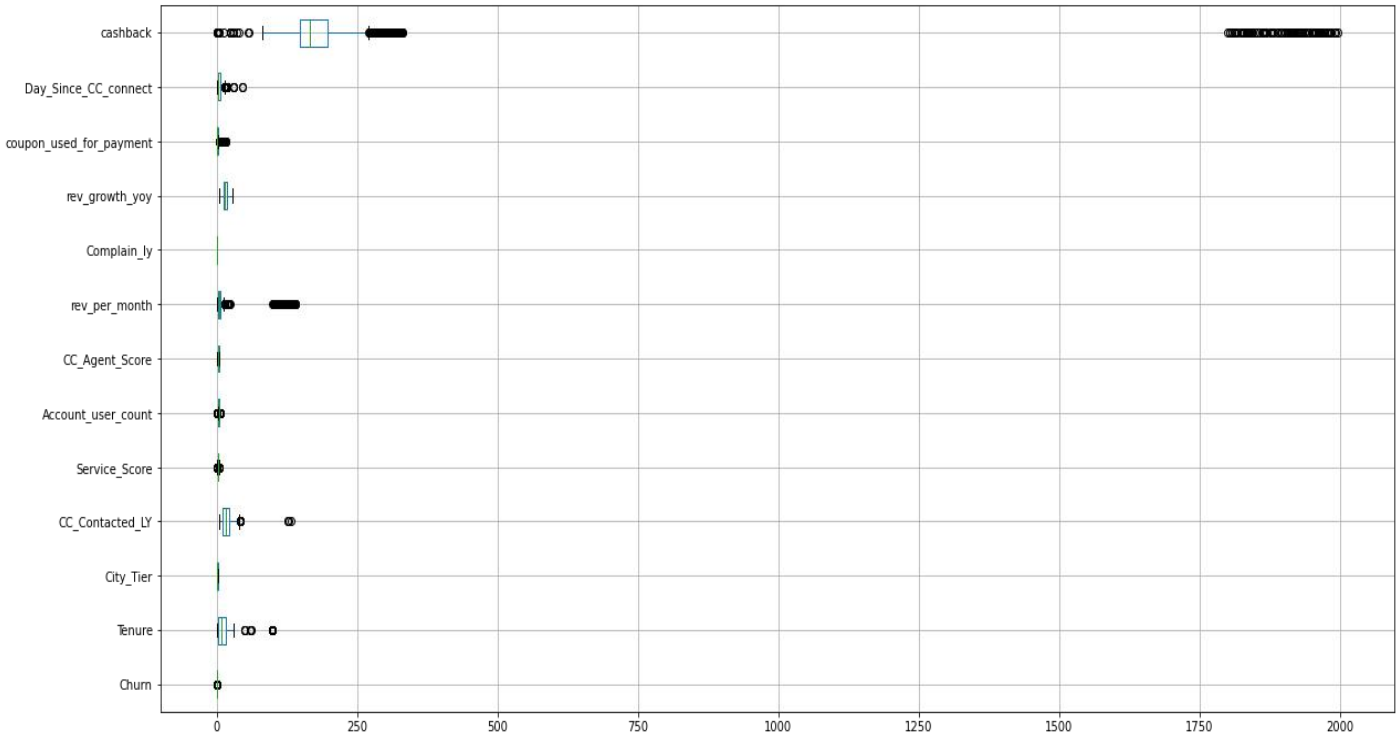

Fig.35 - Box Plot For Continuous Variables

From the above EDA:

- Data is imbalanced since the **churn rate** is only **16.8%** . The data needs to be balanced for modelling approach. **Ref - Fig.2**
- Most users are with **60.5%** of male and only **39.5%** of females users. Business is attracting less females users. Business needs to work on attracting female customers too. **Ref - Fig.3**
- The data has highest Regular Plus subscription of **37.5%** . Customers or users are not attracted to other subscription plans. Regular is the lowest only **4.6%**. **Ref - Fig.4**
- Super plus seems to be higher price range, may also generate good revenue. But it seems it's not better than regular plus or super plans. Business needs to work on this.
- Most males are married and that means they have users in their family both male and female. **Ref - Fig.19.**
- Business can also include or collect data of the **age-range of their customers**. This would help them curate certain offers or products for depending on their gender and age.
- 4 users per account = **44.5%** and lowest is 6 users per account = **2.6%**. **Ref - Fig.5**
- Debit card have highest users = 41.7% . Both Male and Female. Business can offer cashback on other digital platform payment methods. **Ref - Fig.6**
- Tier 2 cities have lowest customer base only **4.3%** compared to Tier 3 and Tier 1 cities. Management of the Business needs to address this issue. As, it can tap into huge potential market. **Fig.7**
- **Customer care Agent score** and **service score** both have obtained highest 3 ratings, which implicates that the services are averages. It may also be due to quality of products. **Ref. - Fig. 8 and Fig 9**
- **Coupons** used are also very low .The cashback and coupons offered by the business doesn't look appealing to customers. 1 = 38.9% and 0 = 19%. **Ref. - Fig 15**
- YOY growth and Revenue per month can be observed in **Fig. 16** and **Fig. 17.** It has decreased over longer period of time. Churn is effective the YOY growth positively with each year **2-1% decrease** in revenue growth. **Ref. - Fig. 32.**
- **Ref. - Fig. 23 to Fig 32** shows how **churn** is affected by other variables in the data.
- **Fig.33 & Fig.34 - Pair Plot & Correlation heat map** - shows how all variables are affecting and related to churn.
- **Fig.35 shows presence of outliers in the data** which needs to be treated before model building approach.

DATA CLEANING & PRE-PROCESSING

Outlier treatment:

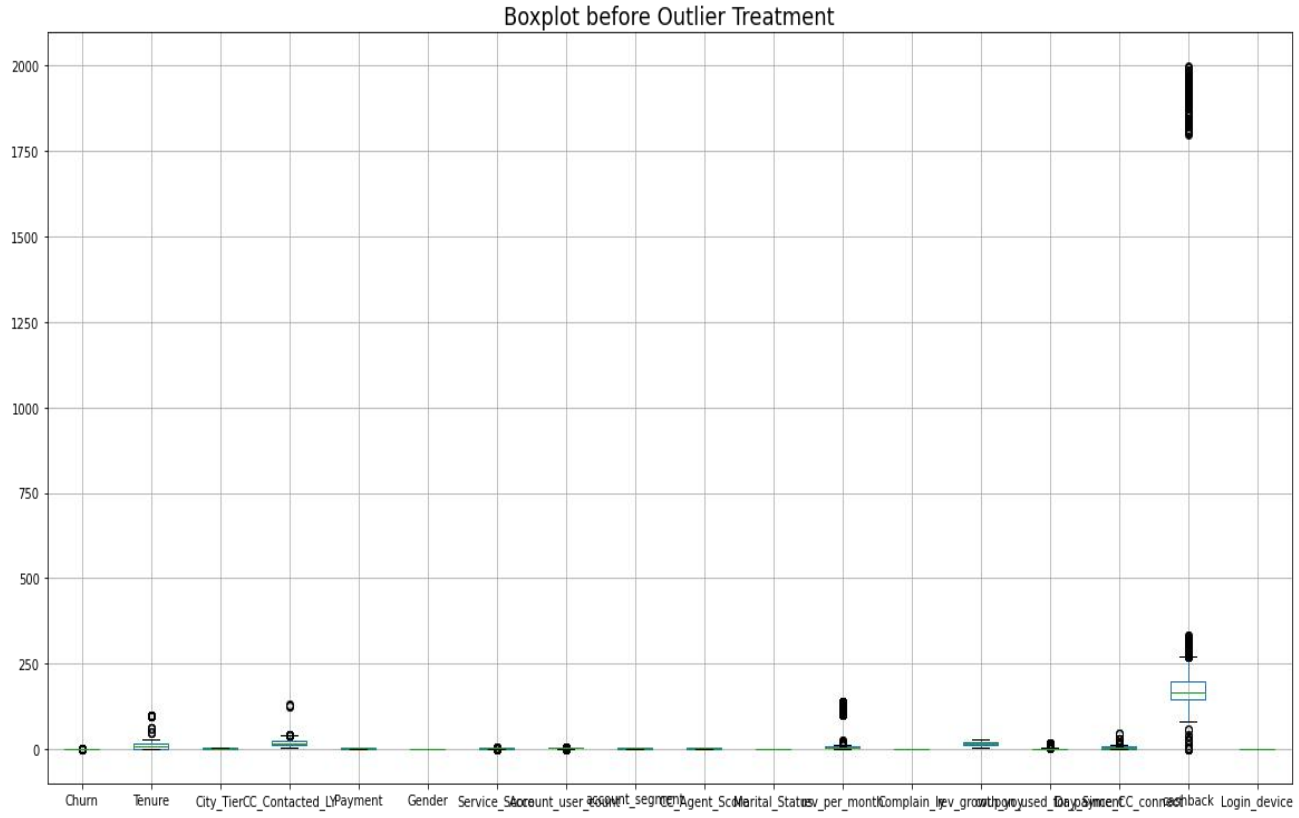

Fig.36 - Boxplot Before Outlier treatment

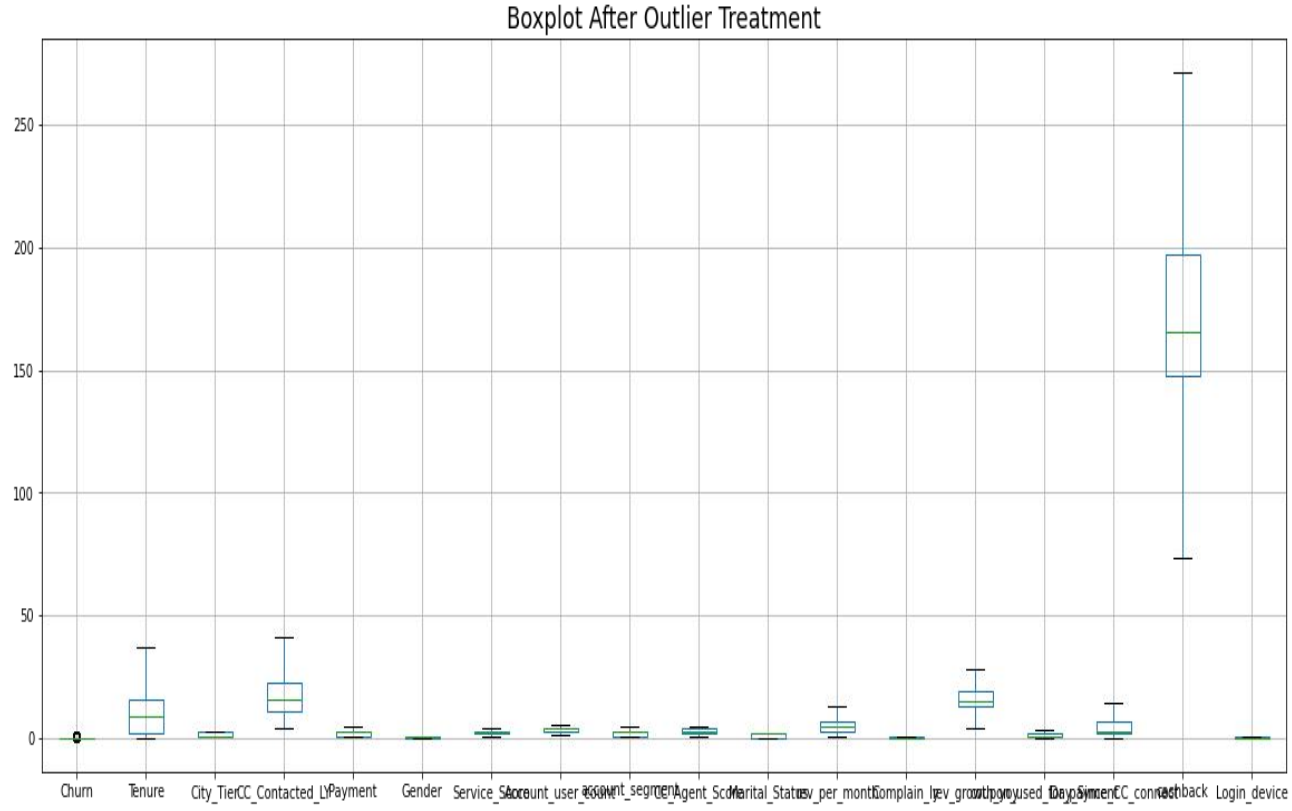

Fig.37 - Boxplot with After Outlier treatment

Variables	Before Outlier Treatment		After Outlier Treatment	
	Kurtosis	Skewness	Kurtosis	Skewness
Churn	1.142	1.773	1.142	1.773
Tenure	23.916	3.94	-0.034	0.817
City_Tier	-1.375	0.753	-1.375	0.753
CC_Contacted_LY	8.331	1.434	-0.223	0.805
Payment	-1.017	0.454	-1.017	0.454
Gender	-1.816	-0.43	-1.816	-0.43
Service_Score	-0.648	0	-0.744	0.014
Account_user_count	0.726	-0.434	-0.085	-0.344
account_segment	-1.089	0.277	-1.089	0.277
CC_Agent_Score	-1.106	-0.141	-1.106	-0.141
Marital_Status	-1.594	-0.46	-1.594	-0.46
rev_per_month	93.894	9.443	0.065	0.809
Complain_ly	-0.999	1.001	-0.999	1.001
rev_growth_yoy	-0.219	0.753	-0.219	0.753
coupon_used_for_payment	9.105	2.576	-0.739	0.46
Day_Since_CC_connect	5.572	1.322	-0.216	0.755
cashback	84.846	8.966	-0.082	0.903
Login_device	-0.903	1.048	-0.903	1.048

Table 4 : Kurtosis and Skewness

Steps and Methods involved :

- **Median and Mode method** is applied to treat the missing values.
- **Box plot** is a good method to identify outliers.
- **Interquartile Range (IQR) method** is applied to detect and subsequently, treat the outliers in the given data. The outlier treatment is necessary to cap the value range of variables with outliers.
- The duplicate values are not dropper as some rows can have similar values.
- Account ID is dropper as it's just represents each cutomer. It won't affect the modelling approach.
- Next step is to apply **SMOTE balancing** method to balance the data.

APPLYING SMOTE METHOD :

Here, balancing is necessary to get an unbiased result and insights from clustering or modeling. In any scenario, whenever data is unbalanced the results obtained are mostly biased towards certain features only which may result in inaccurate predictions and insights. So its is imperative to balance the data before any further analysis.

The Data seems to be unbalanced in case of our target variable - ***" Churn"***

Value counts of Churn :

0	9364
1	1896

To apply **SMOTE technique** we will split the data into 70:30 - Train (70)and Test (30)

Shape : x_train (7882, 17); x_test (3378, 17); y_train (7882,); y_test (3378,)

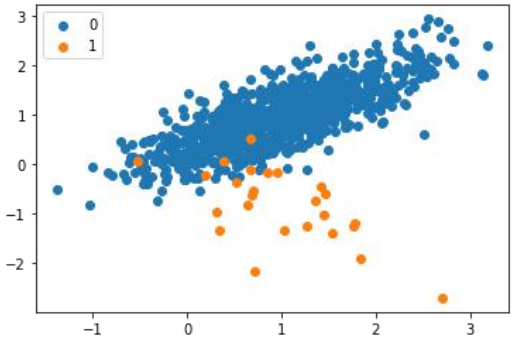

Fig.38 - Unbalanced - Scatter plot - SMOTE

Applying SMOTE gives us,

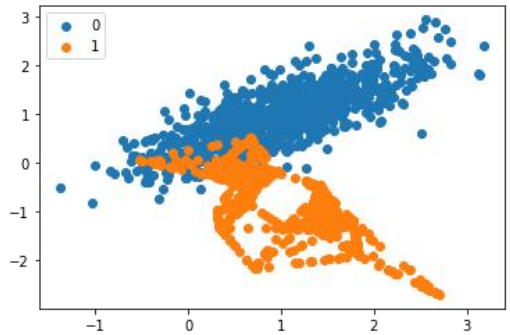

Fig.39 - Balanced - Scatter plot - SMOTE

CLUSTERING :

Applying Hierarchical Clustering,

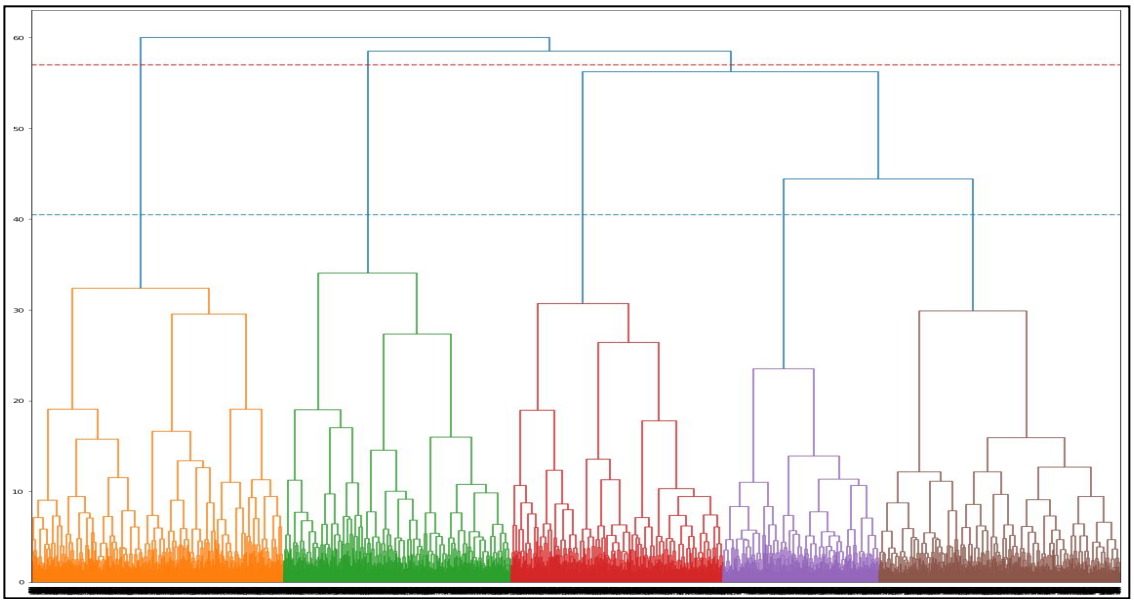

Fig.40 - Dendogram

Applying K-Means Clustering,

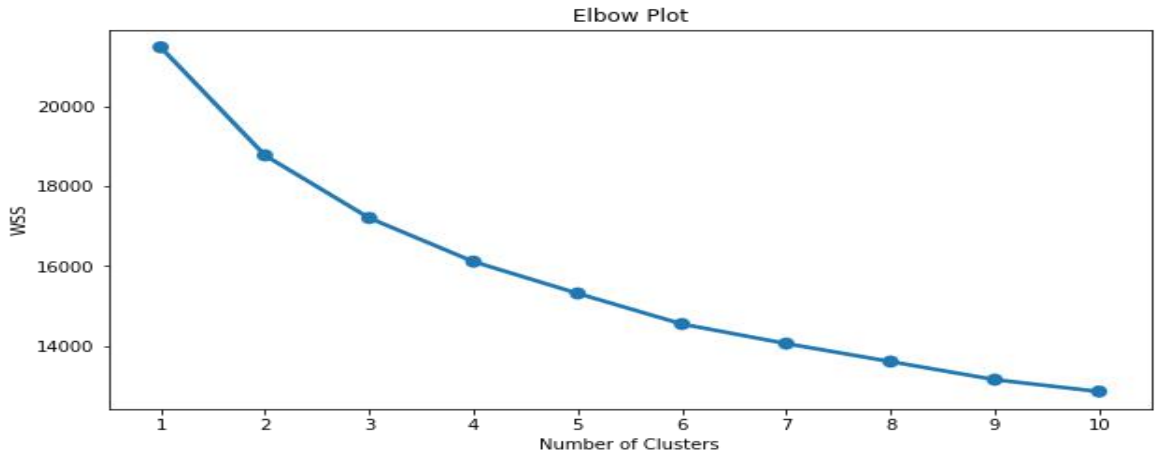

Fig.41- Elbow Plot

Silhouette Score :

- The Average Silhouette Score for 2 clusters is 0.13298
- The Average Silhouette Score for 3 clusters is 0.12104
- The Average Silhouette Score for 4 clusters is 0.11139
- The Average Silhouette Score for 5 clusters is 0.129
- The Average Silhouette Score for 6 clusters is 0.11596
- The Average Silhouette Score for 7 clusters is 0.13784
- The Average Silhouette Score for 8 clusters is 0.11529
- The Average Silhouette Score for 9 clusters is 0.11276
- The Average Silhouette Score for 10 clusters is 0.11665

The best practise in choosing number of clusters is minimum of 5-6 but in the above case we will 3.
Number of clusters = 3

Data Head with KMeans_Labels,

Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status	rev_per_month	Complain_ly	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback	Login_device	KMEANS_LABELS
1	4	3	6	1	0	3	3	1	2	0	9	1	11	1	5	159	0	2
1	0	1	8	2	1	3	4	3	3	0	7	1	15	0	0	120	0	1

Data Tail with KMeans_Labels,

Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status	rev_per_month	Complain_ly	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback	Login_device	KMEANS_LABELS
0	23	3	11	3	1	4	5	1	4	2	7	0	16	2	9	179	1	2
0	8	1	22	3	1	3	2	1	3	2	5	0	13	2	3	175	0	1

Cluster Profiling :

KMeans_Labels value Counts :

1	4832
2	3411
0	3017

Variables	coun	mean	std	min	25%	50%	75	max
Churn	11260.0	0.168384	0.374223	0.0	0.0	0.0	0.0	1.0
Tenure	11260.0	10.251421	8.888905	0.0	2.0	9.0	16.0	37.0
City_Tier	11260.0	1.647425	0.912763	1.0	1.0	1.0	3.0	3.0
CC_Contacted_LY	11260.0	17.815009	8.564140	4.0	11.0	16.0	23.0	41.0
Payment	11260.0	2.399112	1.379380	1.0	1.0	3.0	3.0	5.0
Gender	11260.0	0.604973	0.488878	0.0	0.0	1.0	1.0	1.0
Service_Score	11260.0	2.902931	0.721493	0.0	2.0	3.0	3.0	4.0
Account_user_count	11260.0	3.676998	0.952661	1.0	3.0	4.0	4.0	5.0
account_segment	11260.0	2.596092	1.407510	1.0	1.0	3.0	3.0	5.0
CC_Agent_Score	11260.0	3.065808	1.372663	1.0	2.0	3.0	4.0	5.0
Marital_Status	11260.0	1.226643	0.894745	0.0	0.0	2.0	2.0	2.0
rev_per_month	11260.0	5.250799	2.879616	1.0	3.0	5.0	7.0	13.0
Complain_ly	11260.0	0.276288	0.447181	0.0	0.0	0.0	1.0	1.0
rev_growth_yoy	11260.0	16.193073	3.757271	4.0	13.0	15.0	19.0	28.0
coupon_used_for_payment	11260.0	1.414032	0.996832	0.0	1.0	1.0	2.0	3.0
Day_Since_CC_connect	11260.0	4.540497	3.477415	0.0	2.0	3.0	7.0	14.0
cashback	11260.0	176.805417	43.581623	73.0	147.0	165.0	197.0	271.0
Login_device	11260.0	0.268028	0.442952	0.0	0.0	0.0	1.0	1.0
KMEANS_LABELS	11260.0	1.034991	0.754782	0.0	0.0	1.0	2.0	2.0

Table 5 : Data Description with K-Means

The Silhouette Coefficient for each sample is **-0.06985407665453326.**

KMEANS_LABELS	0	1	2
Churn	0.13	0.16	0.21
Tenure	10.89	10.35	9.55
City_Tier	1.05	1.07	3
CC_Contacted_LY	17.9	17.66	17.95
Payment	2.13	2.1	3.07
Gender	0	1	0.58
Service_Score	2.91	2.9	2.9
Account_user_count	3.69	3.64	3.71
account_segment	2.76	2.67	2.34
CC_Agent_Score	3.18	3.02	3.03
Marital_Status	1.2	1.28	1.17
rev_per_month	5.38	5.24	5.14
Complain_ly	0.28	0.27	0.28
rev_growth_yoy	16.17	16.29	16.07
coupon_used_for_payment	1.44	1.38	1.45
Day_Since_CC_connect	4.64	4.38	4.68
cashback	175.96	173.95	181.6
Login_device	0.25	0.28	0.27
freq	3017	4832	3411

Table 6 : Cluster Profile

Business Insights From clustering :

- From the above clustering method we found 3 important clusters.
- Max freq is of **cluster 1 (2nd cluster).**
- The most influential clusters are **cluster 1 and cluster 2.**

MODEL BUILDING & MODEL VALIDATION

In this part, we are building models, tuning and validation of the model performance based on metrics ,i.e, Accuracy, F1 Score, Recall, Precession, AUC score, confusion matrix, classification report and ROC-AUC curve plot. Based on this we will choose the best model which does not underfit or overfit along with **best accuracy.** Also, we will calculate **mean-square error** and **cross-validation method** to **validate the model**.

Splitting Data into Train and Test set :

The data is split into Train and Test dataset with 70:30 ratio and building models based on train-set and test-set. We will check accuracy between train and test dataset after the split.

Shape of training and test dataset :

Shape : x_train (7882, 17); x_test (3378, 17); y_train (7882,); y_test (3378,)

We will also use ***SMOTE balanced*** dataset on train and test dataset :

x_train for balanced set (13112, 17)
y_train for balanced set (13112, ***)***

Steps involved in model building :

- We have used **default data** set for
- Applied **GridSearch CV method** - hyper parameters to tune the model.
- Lastly with **Balanced Data** set Obtained from SMOTE technique
- Each model will be based on these three criteria.
- Comparing the results will give us the best model based on accuracy.

Objectives of the Model :

- Whether the customer will churn or not?
- Understanding factors affecting churn rate whether its payment method, city tier, gender, customer service, etc.
- Controlling the churn rate by controlling the factors affecting it positively.
- Based on this we will choose the best model which does not underfit or overfit along with **best accuracy**.

Why Best Accuracy Scores ?

- Our current business firm needs to predict both, whether the customer will churn or not.
- Churn = 1 and Not churn = 0
- Higher the accuracy score better the model

From the above steps and method, we have obtained **KNN (balanced data set)** as best possible model for the given data with **best accuracy**.

KNN Model :

KNN Model for SMOTE Balanced set :

Scores	Train	Test
Accuracy	97.87%	93.25%
Precision	96.11%	73.49%
Recall	99.79%	93.86%
F1 Score	97.91%	82.43%
AUC	99.97%	97.35%

TABLE 11 : KNN Scores - Balanced Set

	precision	recall	f1-score	support
0.0	1.00	0.96	0.98	6556
1.0	0.96	1.00	0.98	6556
accuracy			0.98	13112
macro avg	0.98	0.98	0.98	13112
weighted avg	0.98	0.98	0.98	13112

	precision	recall	f1-score	support
0.0	0.99	0.93	0.96	2808
1.0	0.73	0.94	0.82	570
accuracy			0.93	3378
macro avg	0.86	0.93	0.89	3378
weighted avg	0.94	0.93	0.94	3378

TABLE 12 : KNN Classification report (Balanced Set) -Train(Left) & Test (Right)

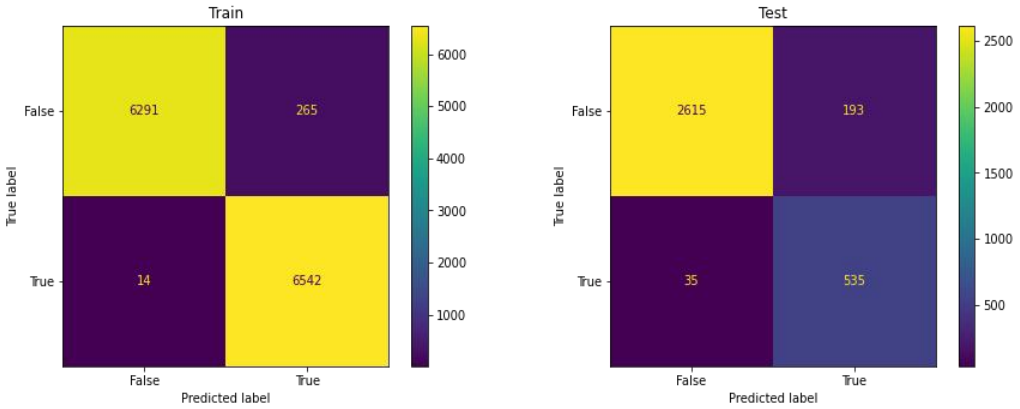

Fig.46 - KNN Confusion matrix (Balanced Set) -Train(Left) & Test (Right)

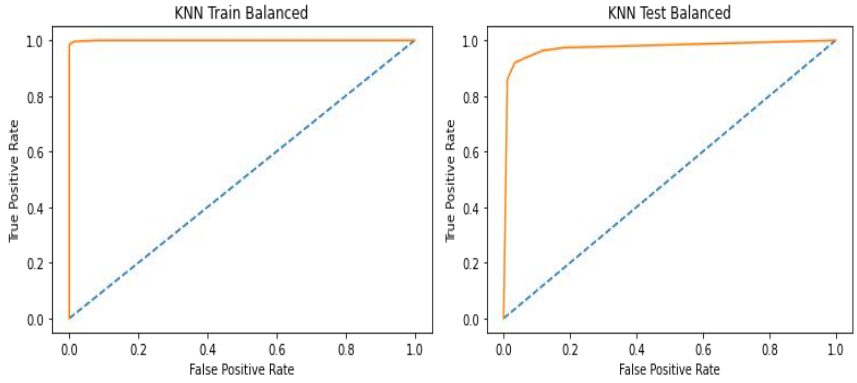

Fig.47 - KNN ROC curve (Balanced Set) -Train(Left) & Test (Right)

Model Validation - Mean Square Error and Cross-Validation Scores :

Mean-Square error % for train-set :: 2.13
Mean-Square error % for test-set :: 6.75

Cross-Validation score % for train-set :: [94.89 96.34 96.19 96.49 95.42 96.11 96.03 94.58 95.58 95.58]
Cross-Validation score % for test-set :: [89.05 89.35 86.98 90.83 91.12 89.05 89.94 88.17 89.61 86.35]

The **RMSE score** for both train and test are low, but train and test have visible gap.
From above results, we observe that the **cross validations scores** are almost similar and indicates that the model built is correctly.

Observations :
Based on the accuracy score between train and test set, this model is performing well. Since the model is based on **balanced data set**, It's a very good model when compared to other better performing models.

Model Comparison :

MODELS	Accuracy (in %)		Precision (in %)		Recall (in %)		F1-Score (in %)		AUC Score (in %)	
Data Set	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
LOGISTIC REGRESSION MODELS (LOG-REG)										
LogReg	88.7	88.34	77.99	75.58	45.7	45.61	57.63	56.89	87	87
LogReg_Tuned	88.77	88.63	76.92	75.98	47.51	47.72	58.74	58.62	87.27	87.18
LogReg_Balanced	80.99	78.3	79.45	42.42	83.6	80	81.47	55.44	88.07	87.18
LINEAR DISCRIMINANT ANALYSIS MODEL (LDA)										
LDA_Model	87.82	87.89	75.49	74.92	40.87	42.46	53.03	54.2	86.49	86.4
LDA_Tuned	87.88	87.8	75.94	74.53	40.95	42.11	53.21	53.81	86.49	86.38
LDA_balanced	80.43	76.97	78.19	40.83	84.4	81.23	81.18	54.34	87.18	87.07
KNN MODEL										
KNN Model	**97.72**	**95.5**	**95.33**	**90.04**	**90.87**	**82.46**	**93.05**	**86.08**	**99.53**	**96.93**
KNN Tuned	100	97.63	100	94.87	100	90.88	100	92.83	100	99.14
KNN Balanced	**97.87**	**93.25**	**96.11**	**73.49**	**99.79**	**93.86**	**97.91**	**82.43**	**99.97**	**97.35**
GAUSSIAN NAIVE BAYES MODEL (NB)										
NB Model	87.5	87.12	65.4	64.27	54.6	53.33	59.52	58.29	84.01	83.62
NB_Balanced	**75.64**	**70.78**	**72.98**	**34.12**	**81.44**	**78.6**	**76.98**	**47.58**	**84.65**	**82.83**
RANDOM FOREST MODEL (RF)										
RF_Model	100	97.25	100	98.57	100	84.91	100	91.23	100	99.27
RF_Balanced	100	96.95	100	93.64	100	87.89	100	90.68	100	98.99
ENSEMBLE MODELS										
Bagging	**99.35**	**95.62**	**99.77**	**96.27**	**96.38**	**77.02**	**98.04**	**85.58**	**99.99**	**98.72**
ADB_Tuned	90.03	90.14	77.22	78.42	57.77	57.37	66.09	66.26	91.85	90.98
GB_Model	91.64	91.21	82.86	82.58	63.42	60.7	71.85	69.97	94.82	92.71

Table 13 : Model Comparison - All Model Scores

FINAL INTERPRETATION & RECOMMENDATION

Observations Based on Model Comparison (Ref. Table-33) :

- Based on above models **KNN_Model (Default set)** is the best optimum model. Although, Bagging model is performing well to but when compared to the recall of both **KNN and Bagging Model with default** set the **recall value** is better in KNN model. **(Highlighted with light green**).
- **KNN Model with balanced set** is performing well **(Highlighted with bright green)**. Since the data is unbalanced we should prefer an unbiased model with balanced data set.
- Although **KNN Tuned**, **RF_Model** and **RF_Balanced** have good scores but they seems to be **over-fitted** models hence cannot be recommended. (Scores between 70 ~ 100 is recommended thumb rule. Below 70% should not be considered as good model.) The **worst** performing model is **Gaussian Naive Bayes model (NB)** with balanced set.
- Model Building is an iterative method, more models can be built and tweaked or tuned to get optimized results. So more models can be built and improved further for best outcome.

Implication of Models on The Business :

- The above recommended models, if implemented can help the business improve their strategies.
- Business can monitor customer activities and predict the outcomes.
- The business can improve offers for customers and other services.
- Business will have clear idea of the buying behavior of their customers.
- Business can predict customer churn for each user account.
- Business monitor customer care and services and their long term implication on the business.
- Revenue growth predictions can be drawn out based on the churn rate.

Insights & Recommendations :

- Business can provide more offers to customers who are more than 5 years and spends a lot
- 3 - 5 years of loyalty programmes can be introduced to attract new customers.
- The cashback and coupon offers attracts good revenue but company needs to gather information if there are customers who are not aware of these benefits.
- Tier 2 cities have lowest customer base and hence company can focus in this region.
- Customer care experience is moderate and hence business provide training to their customer care representative.
- Customer care should act as customer-centric provided it doesn't adversely affect the business under any circumstances.
- Gift vouchers and gift coupons can be introduced for accounts with more than 1 user and additional data can be collected if they are family or friends.
- A curated offers can be given to for the single users to add their family and friends with cash back or discounts on every first purchase made by them.
- Referral programme can be introduced to attract more customers through existing customer with value offers or loyalty points
- More products can be introduced as there more Male customers than Female, keeping in view to attract more female customers.
- Most customers use debit card and credit card. Company emphasize more digital mode of transaction UPI and E wallet as they are a safe-way of transaction with proof of billing.
- 1 and 3 account_segments are the highest spenders. Business needs to focus on other segments as well.
- More value-added services can be implemented in terms customer care and services.
- Business can focus high spending loyal customers with good customer care and service ratings of 4 or 5.