Capstone Project

**ON**

**Customer Churn Prediction**

# 

**By,**

**Shobha Kumari Choudhary**

**Date: - 14th April ’24**



**Contents**

**1. Introduction ----------------------------------------------------🡪 1**

**2. EDA and Business Implication -----------------------------🡪 5**

**3. Data Cleaning and Pre-processing -----------------------🡪 18**

**4. Model building -------------------------------------------------🡪 19**

**5. Model validation ----------------------------------------------🡪 21**

**6. Final interpretation / recommendation ----------------🡪 25**

**List of Figures**

**Fig No - 1: Outlier Check -------------------------------------------------------------------------🡪 4**

**Fig no - 2: Count plot ------------------------------------------------------------------------------🡪 5**

**Fig no - 3: Count plot ------------------------------------------------------------------------------🡪 6**

**Fig no - 4: Pie Chart --------------------------------------------------------------------------------🡪 6**

**Fig no - 5: Continuous variable plot ----------------------------------------------------------🡪 7**

**Fig no - 6: Histogram of Tenure ----------------------------------------------------------------🡪 8**

**Fig no - 7: Boxplots ---------------------------------------------------------------------------------🡪 9**

**Fig no - 8: Count plot w.r.t Target ---------------------------------------------------------------🡪 10**

**Fig no - 9: Count plot w.r.t Target --------------------------------------------------------------🡪 10**

**Fig no-10: Histograms ------------------------------------------------------------------------------🡪 11**

**Fig no-11: Heatmap ----------------------------------------------------------------------------------🡪 12**

**Fig no-12: Pair plot ----------------------------------------------------------------------------------🡪 13**

**Fig no-13: Count plot of City Tier---------------------------------------------------------------🡪 14**

**Fig no-14: City Tier Vs Revenue ----------------------------------------------------------------🡪 14**

**Fig no-15: City Tier Vs Payment mode -------------------------------------------------------🡪 15**

**Fig no-16: Facet grid male customer ----------------------------------------------------------🡪 16**

**Fig no-17: Facet grid Female customer ------------------------------------------------------🡪 16**

**Fig no-18: Facet grid Account user count Male -------------------------------------------🡪 17**

**Fig no-19: Facet grid Account user count Female ---------------------------------------🡪 17**

**Fig no-20: Missing Value Visualization ------------------------------------------------------🡪 18**

**Fig no-21: Boxplot post Outlier Treatment -------------------------------------------------🡪 18**

**Fig No: 22 Model Comparison Table ----------------------------------------------------------🡪 20**

**Fig no - 23: CM LightGBM -------------------------------------------------------------------------🡪 22**

**Fig no - 24: ROC Curve LightGBM --------------------------------------------------------------🡪 22**

**Fig no 25: Feature Imp LightGBM ---------------------------------------------------------------🡪 23**

**Fig no 26: ROC Curve Comparison -------------------------------------------------------------🡪 24**

**Introduction of the business problem**

**1.a) Defining problem statement:**

**An E Commerce company or DTH (you can choose either of these two domains) 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.**

**You have been assigned to develop a churn prediction model for this company and provide business recommendations on the campaign.**

**Your campaign suggestion should be unique and be very clear on the campaign offer because your recommendation will go through the revenue assurance team. If they find that you are giving a lot of free (or subsidized) stuff thereby making a loss to the company; they are not going to approve your recommendation. Hence be very careful while providing campaign recommendation.**

* **Here the objective is to Build a Model to identify the Customers who will potentially Churn and provide useful business recommendations. Retaining a loyal customer is far more important than acquiring a new one. Hence predictive analysis of customer Retention is absolutely necessary in all business.**

**1.b) Need of the study/project:**

* **There are so many competitions available in the market for any product /services we use, so it is very necessary to study about customers likes & dislikes, their preferences about the product, their needs etc.**
* **There is a need to Study about this project in detail because Churning of customers impacts the Revenue of the company.**
* **Customer churn prediction using machine learning will help us to identify risky customers and understand why customers are willing to leave.**

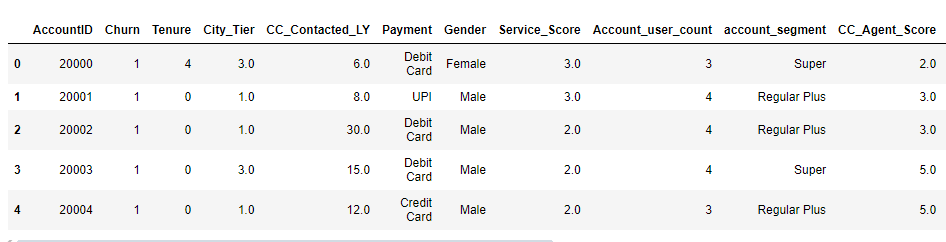
**1.c) Understanding business/social opportunity:**

* **The Goal of this project is to predict the type of customers that has the potential to churn by identifying various features available that can help in minimizing the customers churn rates and get the right business insights.**
* **We need to find a way to Retain the customers as long as possible.**
* **By knowing the customer needs, the kind of services they prefer, we need to keep the customers satisfied.**
* **By knowing the reason that are affecting customer churn will help us predict the churn and then avoid it.**
* **There can be many reasons for customers churn - like poor customer Services, another one could be prices.**
* **attracting new customers is much more expensive than retaining existing ones.**

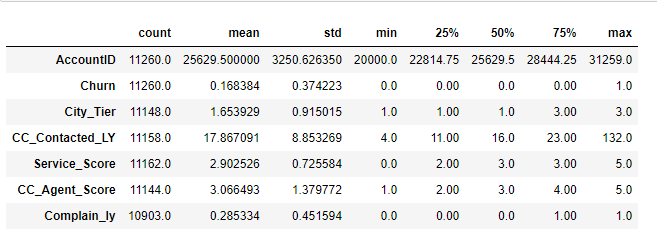
**Data Report**

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

* **Sample of the Dataset:**



* **Shape of the Dataset:**
* **The Number of Rows present in the Dataset are: 11260**
* **The Number of Columns present in the Dataset are: 19**
* **Statistical Summary:**

****

* **Observations:**
* **Columns like Tenure, Account\_user\_count, rev\_per\_month, rev\_growth\_yoy,**
* **coupon\_used\_for\_payment, Day\_Since\_CC\_connect and cashback are continuous**

**columns but missing from the Describe output being a numeric column. This**

**indicates the presence of some Anomalies or some Special Character in these**

**columns.**

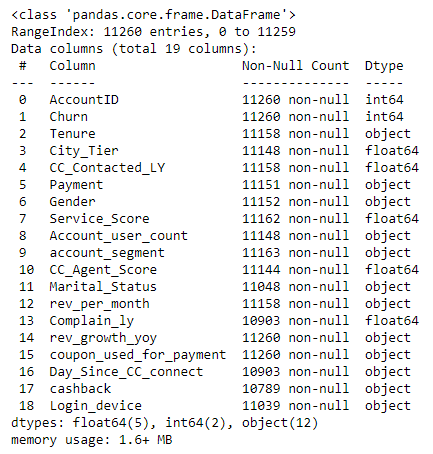
* **Columns like City\_Tier, Service\_Score, CC\_Agent\_Score and Complain\_ly seems**

**Contain Categorical data.**

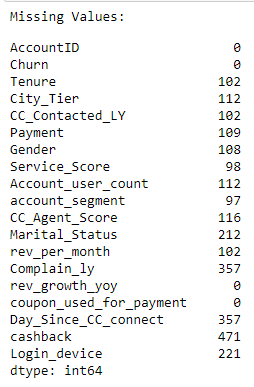
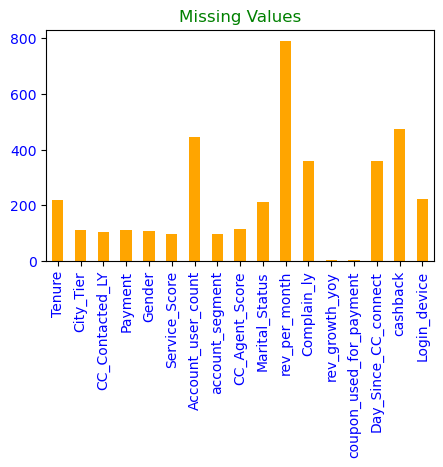
* **Average times Customers contacted CC is around 17 times and Max 132 times.**
* **Average complains raised by the customers is around 28%.**

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

* **Dataset Info:**

****

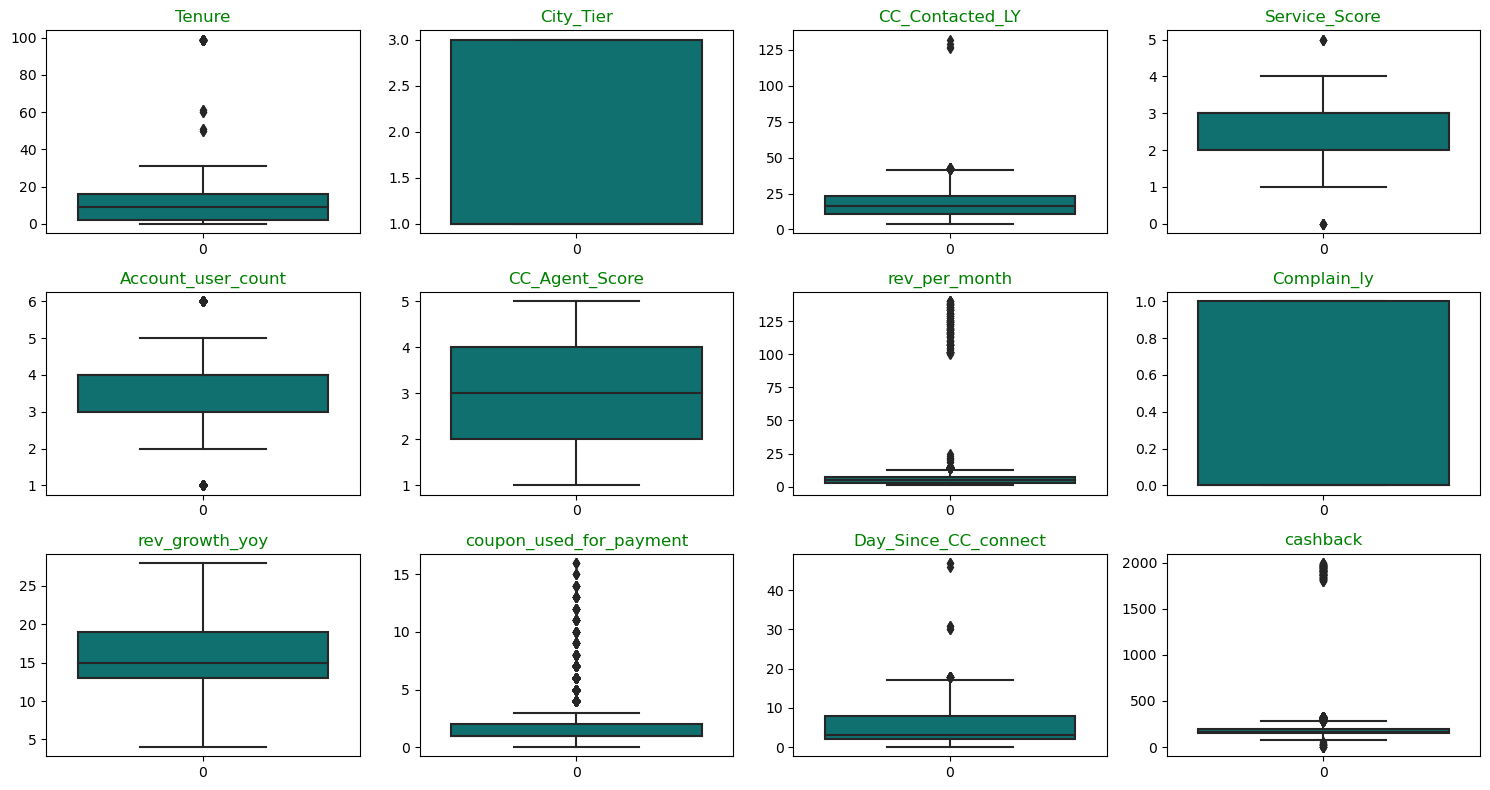
* **Observations:**
* **Null values are present in the dataset.**
* **There are 5 float64, 2 int64 and 12 object Datatypes.**
* **Columns like Tenure, Account\_user\_count, rev\_per\_month, rev\_growth\_yoy, coupon\_used\_for\_payment, Day\_Since\_CC\_connect and cashback contains integer values but their Datatype is object here.**
* **This may be due to some Anomalies or any special character present in the data.**
* **We need to inspect such columns for any Anomalies/Special character present in it.**
* **Duplicates:**
* **There are No Duplicate Rows present in the Dataset.**
* **Missing Values:**

****

* **Checking for Anomalies/Discrepancies:**
* **Column Tenure contains a Special character '#' as an entry making it an object datatype.**
* **Column Gender contains four unique values as 'Male', 'Female',' M','F', which needs to be replaced with 'Male' and 'female'.**
* **Account\_user\_count also contains special character '@' as one of the entries.**
* **Column account\_segment contains some Anomalies like 'Regualr plus' and 'Super Plus' is written in two different ways which needs to be corrected.**
* **Column rev\_per\_month Contains '+' plus which needs to be replaced.**
* **Column rev\_growth\_yoy and Day\_Since\_CC\_connect contains '$' character.**
* **Column coupon\_used\_for\_payment contains #, $ and \* characters.**
* **Column Login\_device Contains &&&& as one of its categories.**
* **We need to inspect the Cashback column for any non-digit character present in it.**
* **We saw that Cashback column contains special character '$' due to which it is considered as 'object' datatype column.**

**We will Replace the special characters present in the dataset with np. Nan and later will impute the Null values.**

* **Checking for Outliers:**



**Fig No - 1: Outlier Check**

* **Observations:**

➢ **There are many features showing the presence of Outliers.**

**Exploratory data analysis**

## **Univariate Analysis**

* **Count plot of all Categorical Columns:**

### 

### ****Fig no - 2: Count plot****

**Insights:**

* **Most Customers Preferred Payment mode are Debit card and Credit Card.**
* **Very few customers make payment through UPI.**
* **Male customers are more as compared to female.**
* **There are more customers that belong to account\_segment Regular plus and Super.**
* **Very less customers belongs to Regular segment based on their spend.**
* **Married customers are more in number as compared to Singles and Divorced.**
* **Maximum of the customers prefers to Login Via Mobile as it is convenient to carry and they can Login from any Location.**
* **Very less customers Login via Computer.**

### 

### ****Fig no - 3: Count plot****

**Insights:**

* **Maximum customers belong to Tier-1 cities followed by Tier-3 cities and very few belongs to Tier-2 cities.**
* **Service score given by Maximum customers is 3, which states that customers are not fully satisfied by the service provided by the company.**
* **More numbers of Accounts are linked with 4 members, followed by 3 and 5 members per Account.**
* **Very few accounts are linked with 1,2 and 6 Members.**
* **Satisfaction score given by customers on an Average is 3 for Customer Care Services.**
* **Also, Excellent Score of 4 and 5 for Customer care services are almost equal in number. So, they are Satisfied customers too.**
* **It seems very less Complaint has been Raised by the customers in last 12 months.**
* **Distribution of Target Variable:**

### 

### ****Fig no - 4: Pie Chart****

**Insights:**

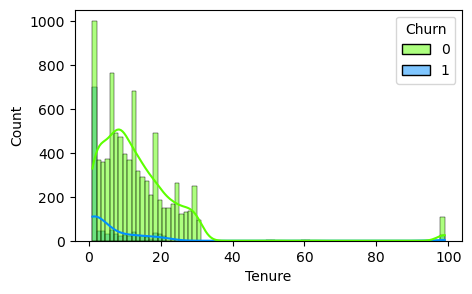
* **16.8% of the customers have churned whereas 83.2% of the customers have not churned.**
* **seems Imbalance in Target class.**
* **Histogram and Boxplot of all continuous variables**

### 

### ****Fig no - 5: Continuous variable plot****

**Insights:**

* **All the Variables are Right Skewed showing the presence of Outliers.**
* **Maximum customers have a Tenure of less than a month.**
* **There are also some customers having a Tenure of more than 50 months, Max up to 100 months.**
* **Maximum number of customers have contacted Customer care 11 to 23 times in last 12 months.**
* **The Median of the Monthly average revenue generated by the company is around 10k (Assuming the Currency is in Thousands).**
* **Also, this feature is highly Right Skewed showing monthly avg revenue generated by the company more than 100k.**
* **There is an Approx 16% growth in revenue on an average, generated by the account in the last year compared to the previous years.**
* **On an Average, 2 times coupons were used to do the payment.**
* **Also, it seems some customers used the coupons more than 4 times to max 16 times.**
* **Avg no of days since customers have not contacted CC is around 5 days.**
* **On an Average, Cashback generated by the customers is around Rs 200/- in the last 12 months.**
* **Also, some customers generated cashback of more than Rs 1750/- monthly.**



**Fig no - 6: Histogram of Tenure**

**Insights:**

* **Customers having a Tenure of less than 15 months, Churns more.**
* **There are some Loyal Customers having Tenure of more than 80 months.**
* **Distribution seems Overlapping.**

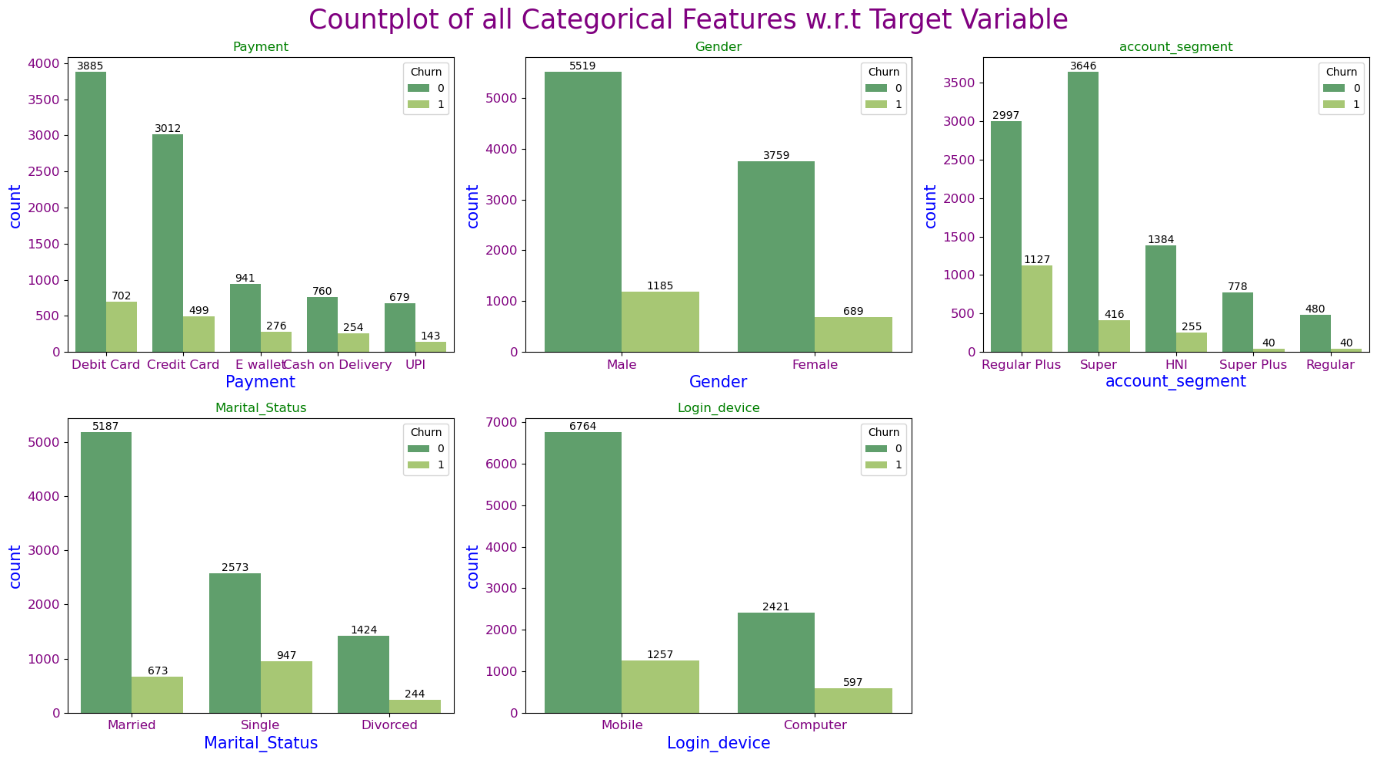
**Bivariate Analysis**

### 

**Fig no - 7: Boxplots**

**Insights:**

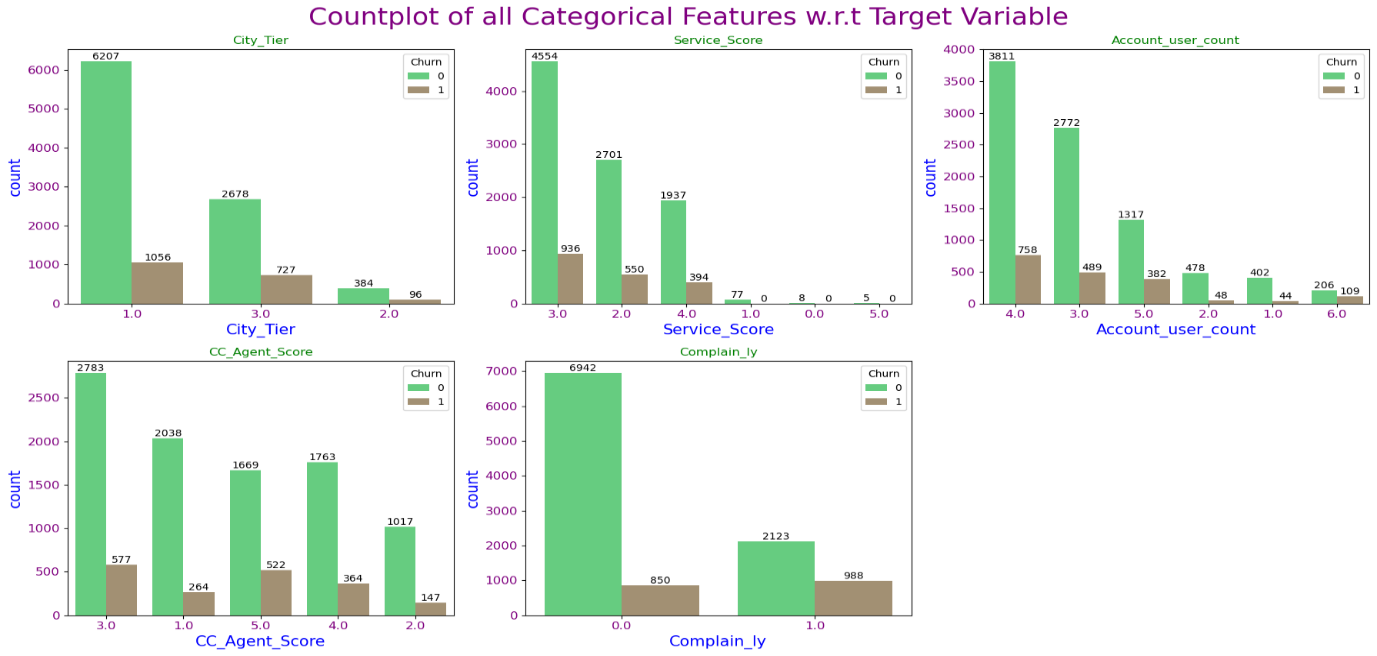
* **It seems that when the customers contact the Customers care more their queries gets resolved due to which they don't churn and keep using the services more hence reducing the Churn rate.**
* **So, we can say Customer Care is also playing an Important role in Retaining Customers.**
* **We see that median value of the Tenure and Days\_since\_CC\_connect for Churners is less compared to that of non-Churners.**
* **The Distribution of Coupon\_used\_for\_payment, Cashback and rev\_per\_month is same for both Churners and Non-churners.**
* **There is no difference in median rev\_growth\_yoy between churners and non-churners.**
* **Count plot of all Categorical Variables Vs Target Variables**



**Fig no - 8: Count plot w.r.t Target**

**Insights:**

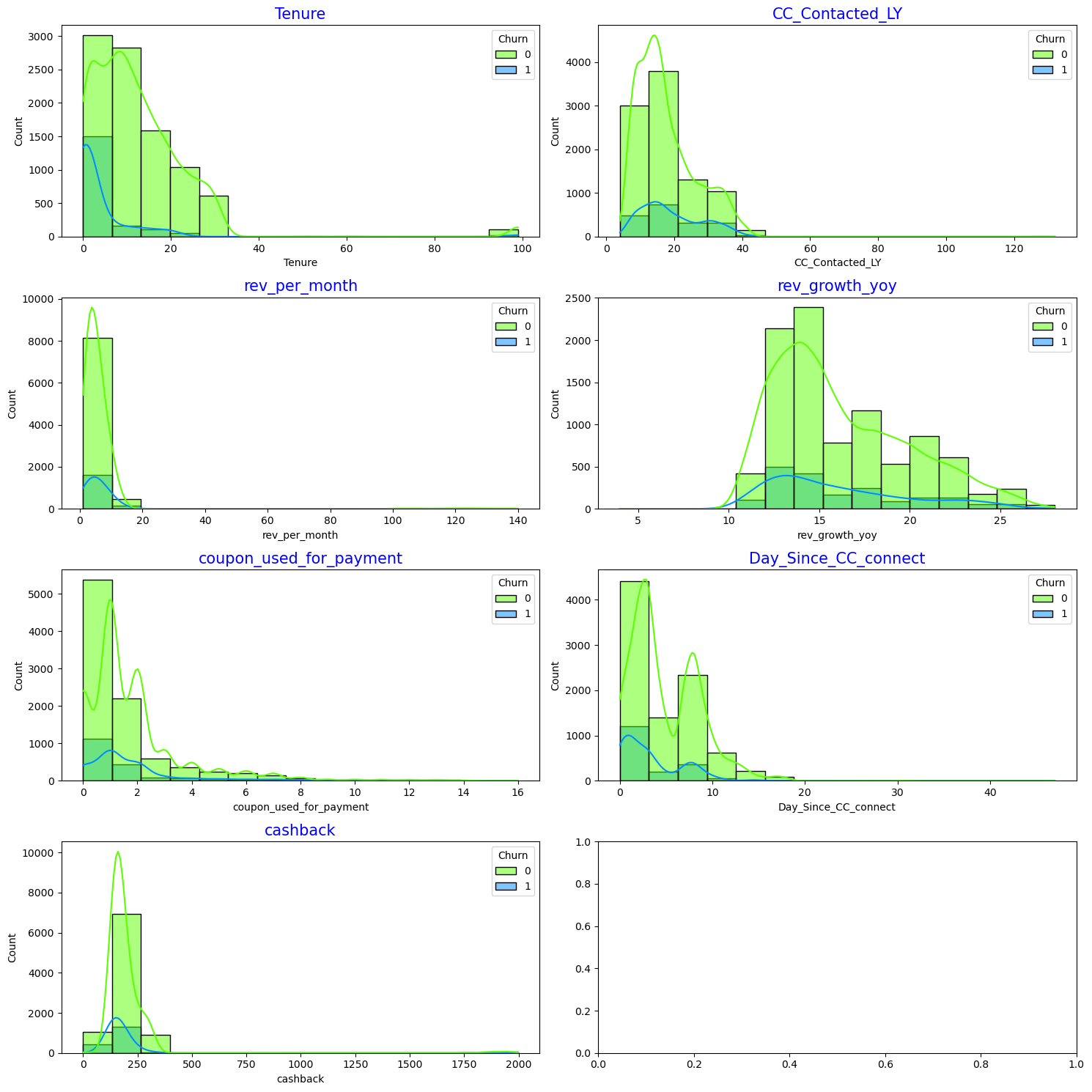
* **The proportions of Churners are more of Male customers as compared to Female.**
* **Most of the churners make payment via Debit Card.**
* **Non-churners preferrable mode of payment are Debit cards and Credit Cards.**
* **Most of the churners belongs to Regular plus and Super account segment.**
* **Maximum non-Churners belongs to Super Segment followed by Regular plus.**
* **Most of the churners are single.**
* **Non-Churners are Maximum Married Couples.**
* **Most Preferred Login Device for both Churners and Non-Churners is Mobile since it's handy.**



**Fig no - 9: Count plot w.r.t Target**

**Insights:**

* **Maximum non-churners belong to City Tier-1 followed by Tier-3.**
* **Churning rate of Customers from Tier-2 is very less, means customers tends continue the services seems they are more satisfied.**
* **Most of the Churners are from Tier-1.**
* **Maximum Service score given by customers is 3 by both churners and non-churners.**
* **Account tagged with 3,4 and 5 customers have more churning rate.**
* **Customers tagged with 2,1 and 6 accounts are mostly non-churners.**
* **Customers who have given an agent score of 3 and above show the Maximum churn rate.**
* **Maximum non-churners have given a Agent score of 3 and 1.**
* **Very few Churners have raised the complaints in the last 12 months. If they would have raised the complaint, then may be their queries would have been resolved and they would not churn.**
* **It is evident from the fact that Customers who have raised the complaint maximum no of times are mostly non-churners. This shows that raising complaint have solved their issue and hence made them retain the use of service and decreasing churn Rate.**

**Checking separation across the numerical variables:** 

**Fig no-10: Histograms**

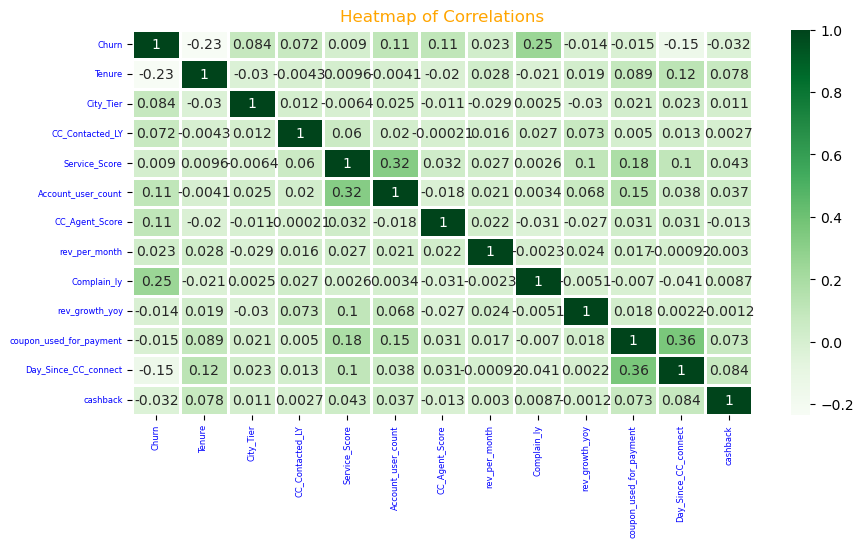
**Insights:**

* **Most of the Distributions are Right Skewed.**
* **Customers having Short-term Tenure churns more.**
* **There are some Loyal customers also having Tenure more than 80 months.**
* **Maximum number of times Customer care was contacted is between 4 to 23 times.**
* **maximum customers use 1-2 coupons for payment.**
* **Most of the cashback are generated is around 0 to 300.**
* **Maximum cashbacks are generated by non-churners.**

**Correlation Heatmap:**

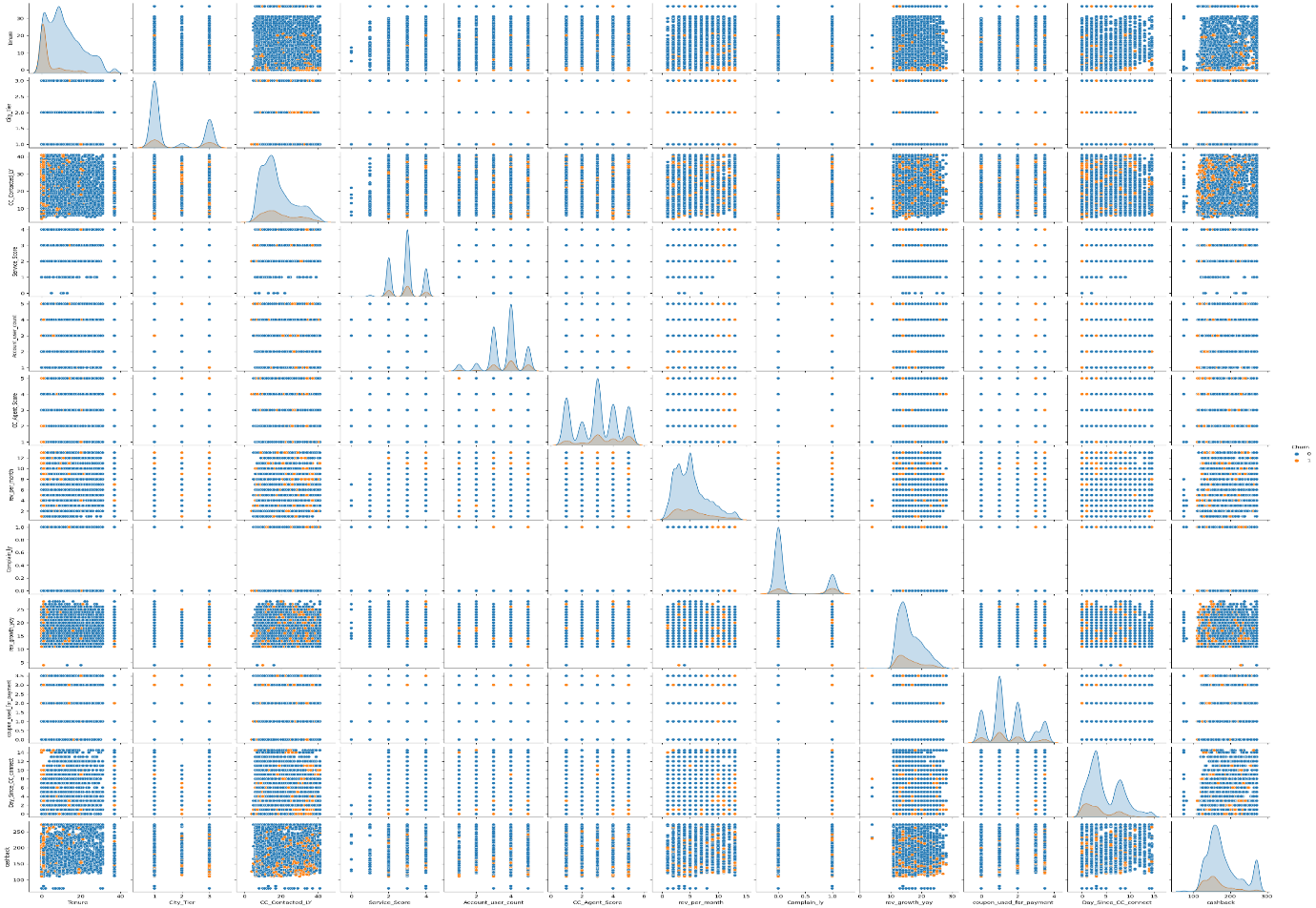
**Insights:**

* **Churn is negatively correlated with Tenure, having a correlation of -23%.**
* **The smaller the tenure value, the greater the churn rate.**
* **Correlation between Churn and Complain\_ly is around 25%. The greater the complaint rate, the greater the churn rate.**
* **Service score has 32% correlation with Account user count and 18% correlation with Coupon used for payment.**
* **Coupon used for payment has 36% correlation with Days since CC connect, 18% with Service score and 15% with Account user count.**



**Fig no-11: Heatmap**

* **Pair plot:**



**Fig no-12: Pair plot**

**Insights:**

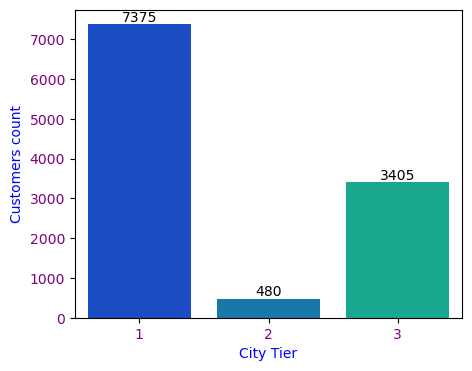
* **Churners and non-churners seem overlapping each other in almost all the features.**
* **There is no Linear Pattern observed.**
* **Customers Churn more with Lowest Tenure.**
* **Tier-1 and Tier-3 Customers Churn Rate is more compared to Tier-2 cities customers.**

## **Data Visualization using Segmentation**

**Let's Segment the data based on City Tiers, Payment Mode and Genders Respectively and will Draw useful insights if any.**

**Segment on the basis of City Tiers**

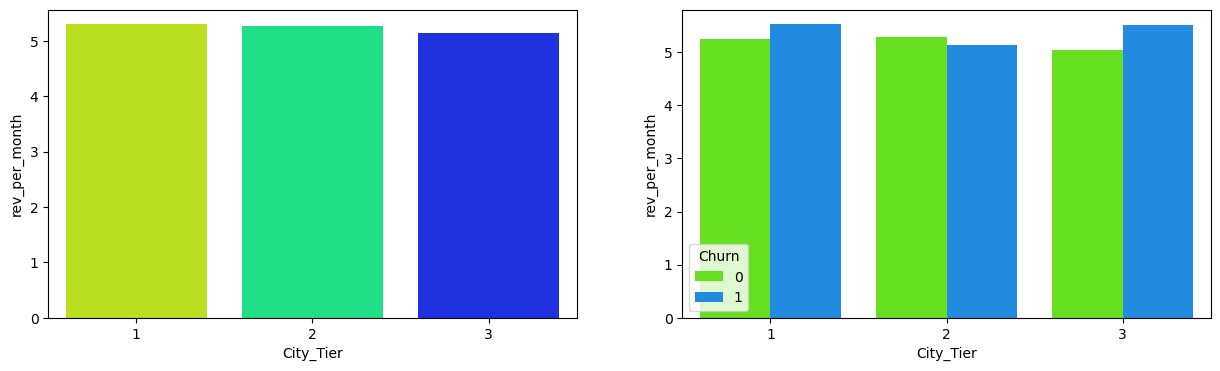
**There are 3 city tiers mentioned in the dataset. Generally, tier 1 cities are considered as the major metro cities where the people tend to use more DTH Services. So, accordingly, can we say that tier 1 city customers tend to generate more avg revenue as compared to tier 2 and tier 3 city customers. Let's visualize this and find it it's True or not?**



**Fig no-13: Count plot of City Tier**

**Insights:**

* **Count of customers are more in Tier-1 followed by Tier-3.**
* **Ver less customers belong to Tier-2.**
* **City Tier Vs Monthly Average Revenue:**



**Fig no-14: City Tier Vs Revenue**

**Insights:**

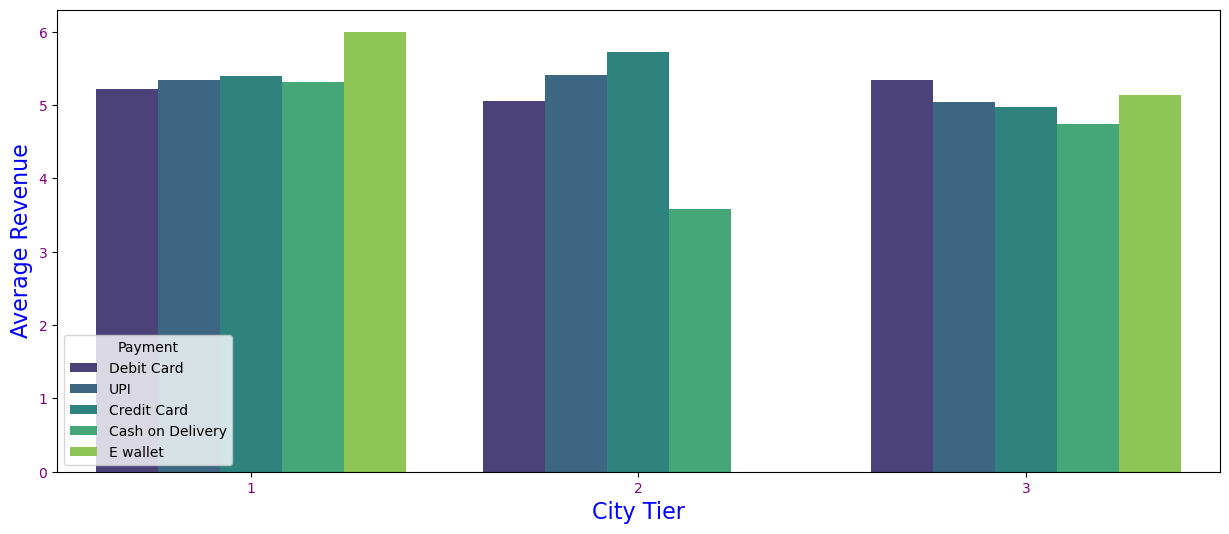
* **We can see that Avg Revenue Generation is same across all the City Tiers.**
* **Though the Customers using the Services in Tier-2 is very less as compared to Tier-1 & 3, But the Avg Revenue generated per month by the account is same across all the Tiers.**
* **It seems Customers are more satisfied in Tier-2 hence they Retain to use the Services, generating greater revenue.**
* **As we can see from the plot, City\_Tier 1 & 2 has a slightly higher mean of Avg Revenue per month as compared to City\_Tier 3 which are more or less the same. So, our assumption here that tier 1 city customers tend to generate more avg revenue cannot be validated looking at the plot.**

### Payment Mode

#### **There are different payment modes (CC, DC, COD, E-wallet & UPI). Depending on the city tiers, Let's Visualize and find the preferred payment modes used by the customers.**

### 

* **46.2 % of customers prefer Debit card as preferred payment mode in Tier 1 cities.**
* **47.08 % of customers prefer UPI as preferred payment mode in Tier 2 cities.**
* **35.39 % of customers prefer E wallet as preferred payment mode in Tier 3 cities.**



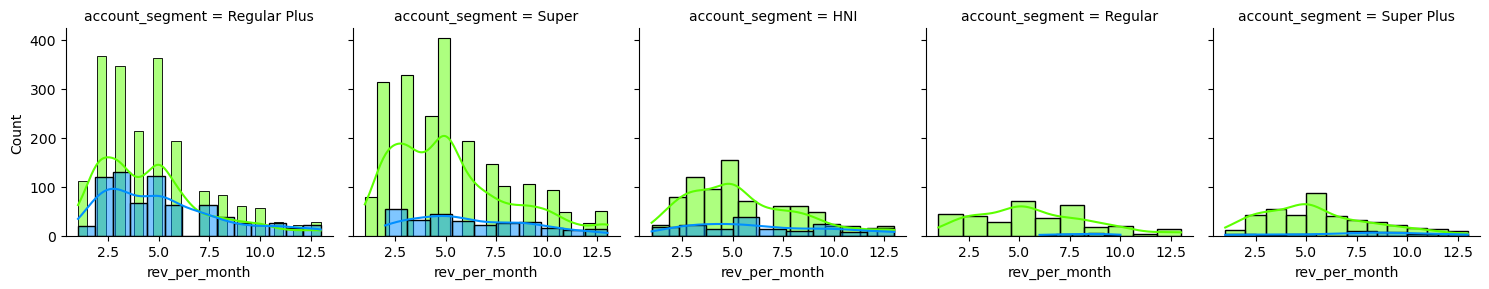
**Fig no-15: City Tier Vs Payment mode**

**Insights:**

* **As we can see, E wallet is used only by the Tier 1 & Tier 3 cities.**
* **Tier-2 Customers don’t prefer using E-wallet as payment mode.**
* **Most of the customers prefer using E wallet and Debit Card in tier 3 cities.**

## **Segment on the Basis of Gender**

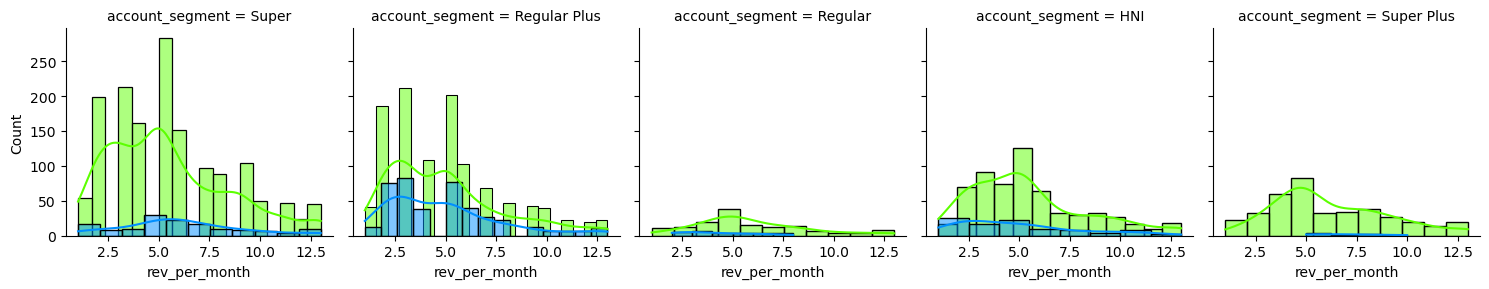
* **Count of Gender:**
* **Male Count: 6812**
* **Female Count: 4448**
* **Revenue Generated based on Spend (Male Customers):**



**Fig no-16: Facet grid male customer**

**Insights:**

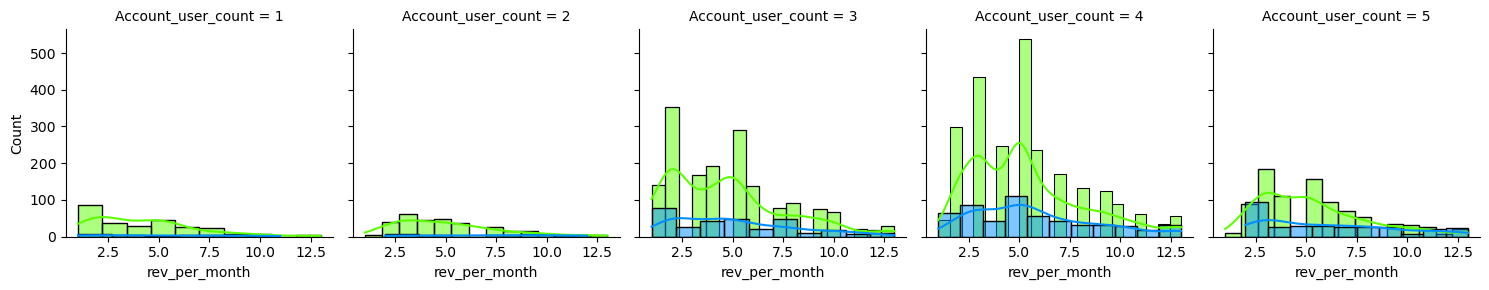
* **Maximum Revenue is generated by Regular plus and Super account holder by male customers.**
* **Revenue Generated based on Spend (Female Customers)**



**Fig no-17: Facet grid Female customer**

**Insights:**

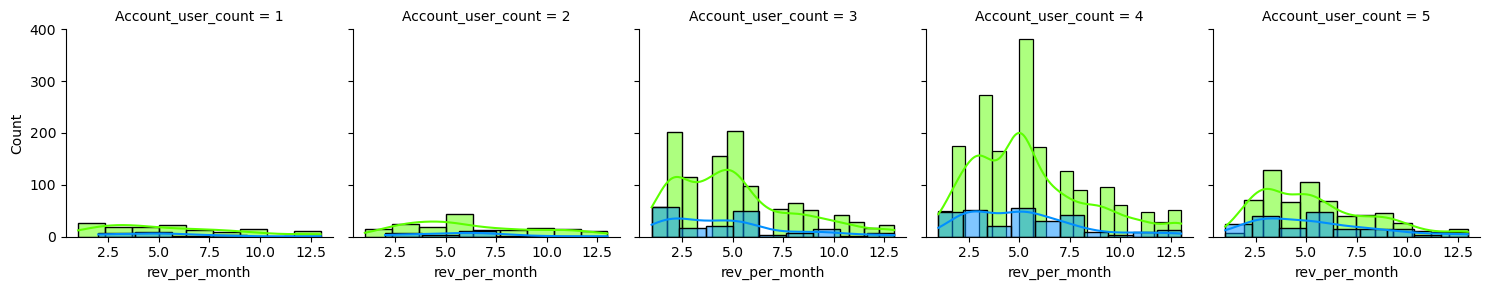
* **Maximum revenue is generated by Super account holder followed by Regular Plus Female customers.**
* **Revenue Generated based on Account\_user\_count (Male Customers):**



**Fig no-18: Facet grid Account user count Male**

**Insights:**

* **Account tagged with more users generate more Revenue by Male customers.**
* **Revenue Generated based on Account\_user\_count (Female Customers):**



**Fig no-19: Facet grid Account user count Female**

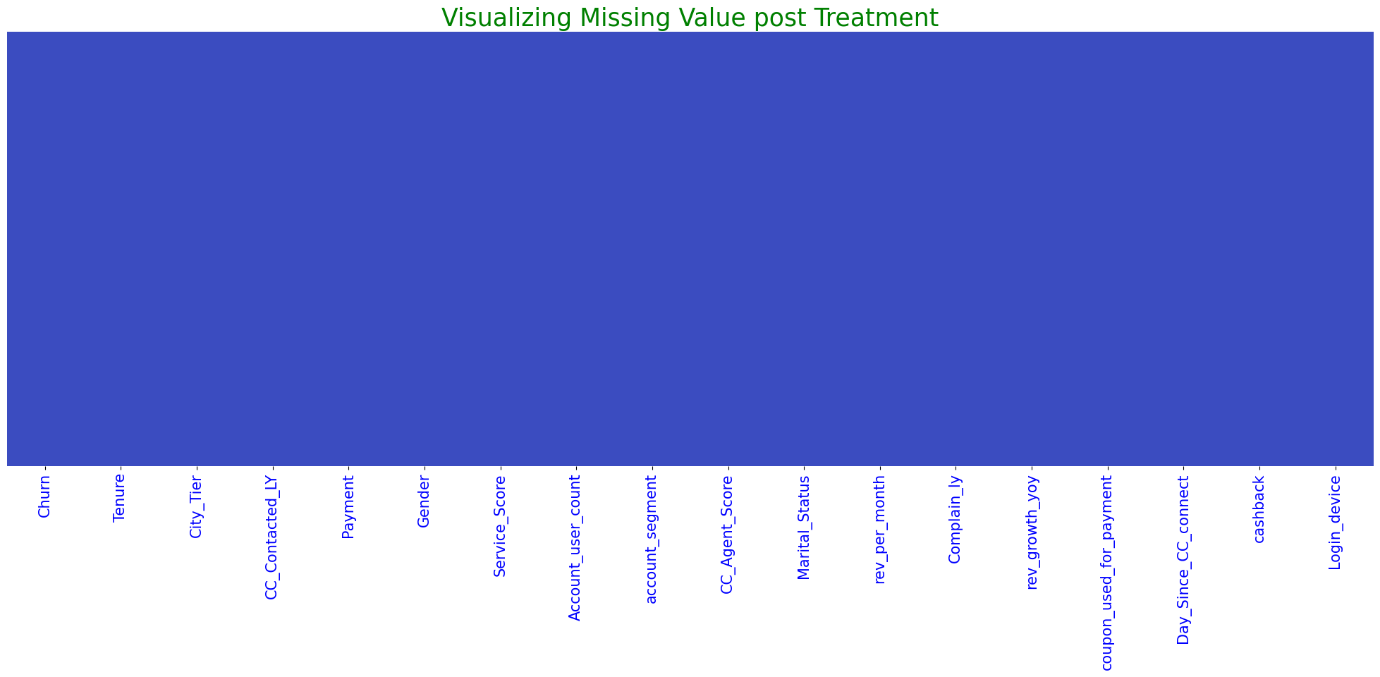
**Insights:**

* **Account tagged with more users generate more Revenue by Female customers.**

**Data Cleaning and Pre-processing**

**Missing Value Treatment**

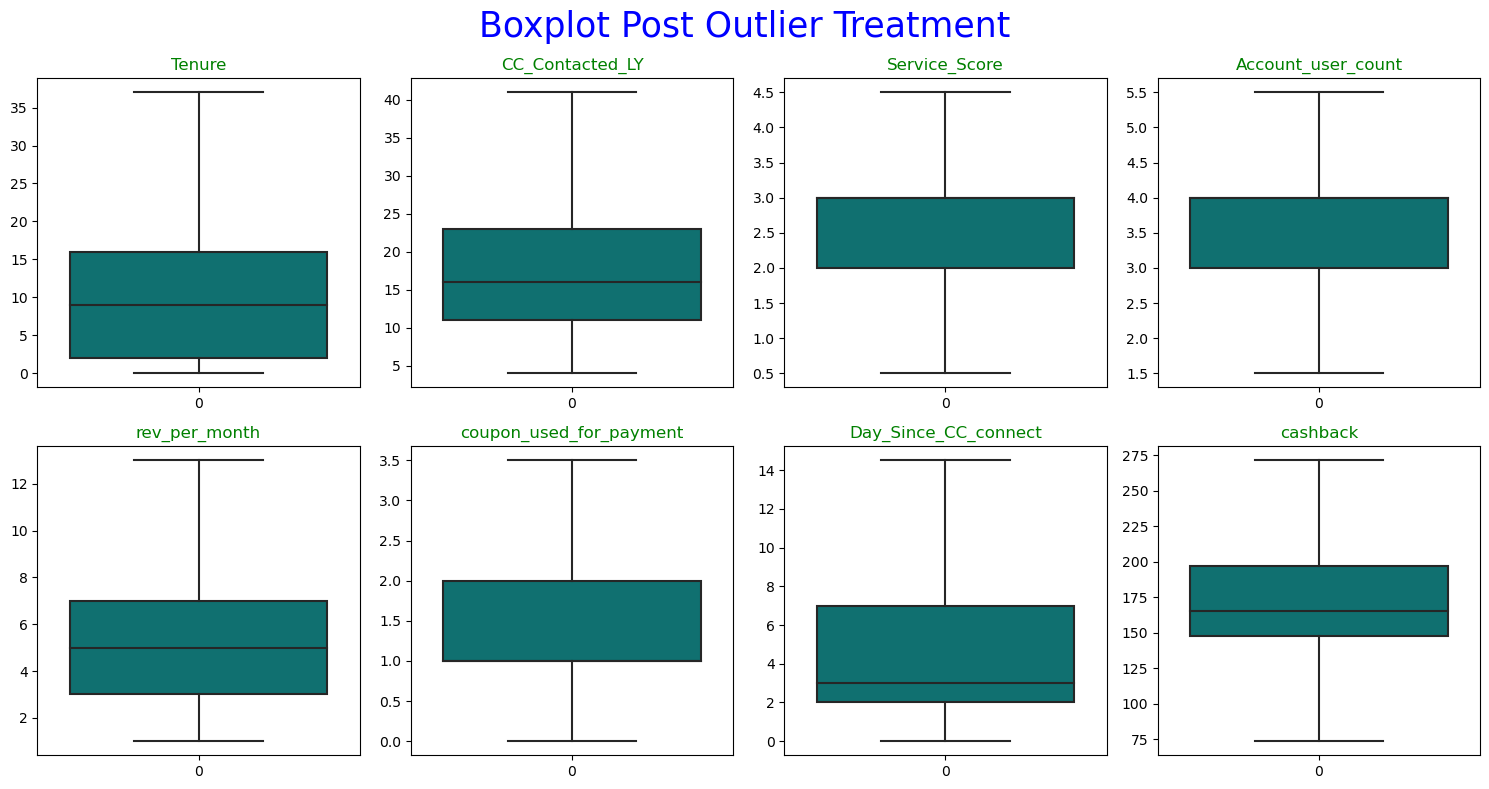
* **Missing value is find Using isnull () function.**
* **Since the proportion of Missing Values was less than 30%, so we imputed the missing values rather than dropping them.**
* **Imputed the Missing Values using Simple Imputer.**



**Fig no-20: Missing Value Visualization**

**Outlier Treatment**

**Outlier is treated using Quantile method by finding Upper and Lower range using the formulae,** **Lower\_range = Q1 - (1.5 \* IQR) and upper\_range = Q3 + (1.5 \* IQR)**



**Fig no-21: Boxplot post Outlier Treatment**

**Data Encoding:**

* **Data Encoding is done using Label Encoder for Categorical Variable.**

## **Feature Transformation:**

* **Predictors are Scaled using Min Max Scalar.**

## **Variables removed or added and why (if any):**

* **Addition of New variables is not required in this problem.**
* **Dropped column AccountID as it is not required for Analysis.**

**Feature Selection**

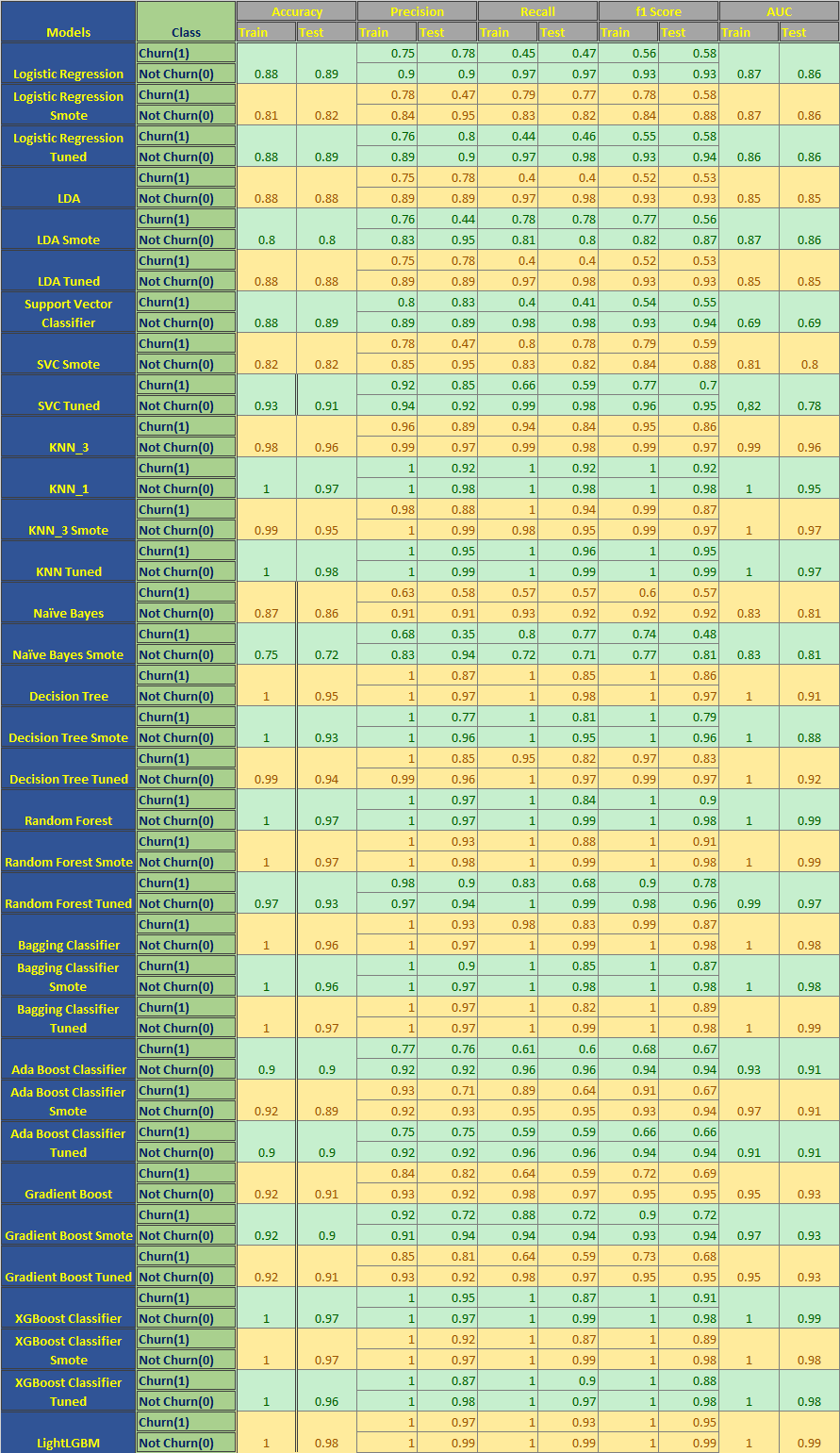
**Variance Inflation Factor is used to Select the Important features.**

* + - * **We will consider features having VIF < 5**
      * **Features having VIF > 5 are removed from the datasets.**
      * **We removed 4 Features Considering VIF > 5.**
      * **Below are the important features.**

### 

### ****Model building.****

* **Build various models as part of the project.**
* **Build models on both balanced and unbalanced datasets and checked the Accuracy of each model against the Test data.**
* **Models are optimized through Hyperparameters using GridSearchCv to increase the Accuracy and remove Overfitting of the model.**
* **The following table below shows the various models built along with their performance metrices like Accuracy, Precision, Recall, f1 score and AUC value.**



**Fig No: 22 Model Comparison Table**

* **Interpretations:**
* **Among all Built models, LightGBM model is giving Highest Accuracy of 98% on Test data and Highest AUC score of 99%.**
* **So, LightGBM is the most Optimal model in my case.**

## **LightGBM Classifier**

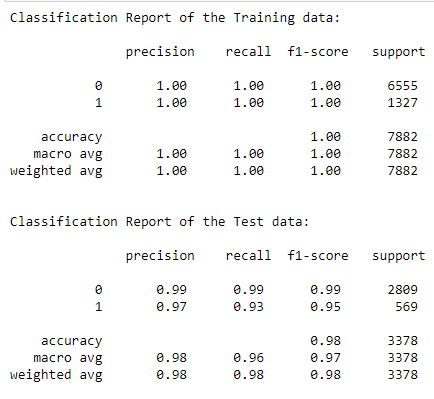
**LightGBM models was chosen to increase Accuracy and Reduce Overfitting of the model.**

* **Accuracy of Train and Test set:** 
  + **Accuracy on Train Data: 1.0**
  + **Accuracy on Test Data: 0.98**

**Model validation.**

**Model is validated using Accuracy, Precision, Recall and AUC score.**

* **Classification Report:**

****

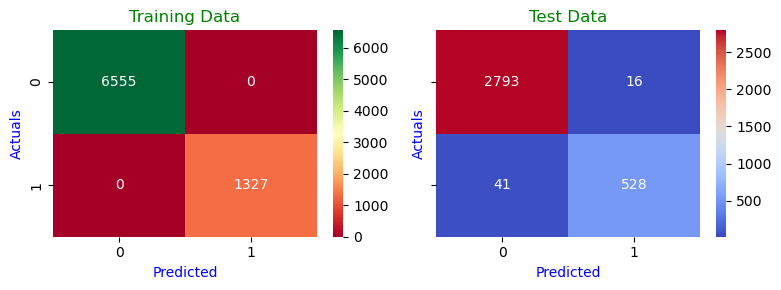
* **Interpretations:**

**---------------------------------------------Test set-------------------------------------------**

* **Accuracy:**
* **The Accuracy of model is 98%.**
* **98% of the Data points are identified correctly.**
* **Recall: Represents the correctness of the Model in a particular class.**
* **Recall for class 1(Churn) is 93% and recall for class 0(non-churn) is 99%.**
* **93% of those Churned were correctly identified as Churned by the model.**
* **Out of all the customers who are Churned,93% of them are identified.**

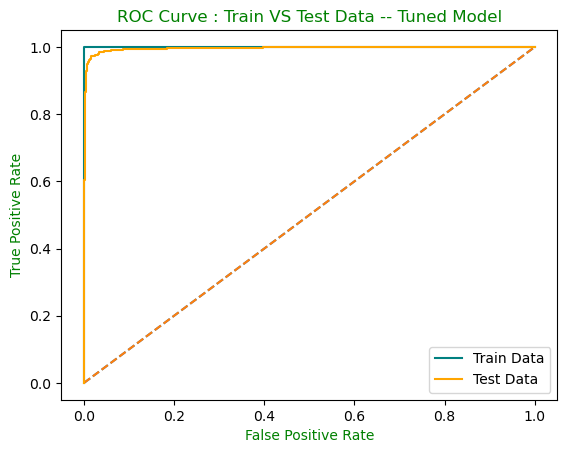
**correctly as Churned.**

* **Out of all the customers who are not Churned, 99% of them are identified correctly as Not Churned.**
* **Precision:**
* **Precision for class 1(Churn) is 97% and for class 0(non-churn) is 99%.**
* **97% of the customers predicted are actually Churned Out of all the Customers that are predicted as Churned.**
* **99% of the Customers predicted are actually Not Churned out of all the customers predicted as not Churned.**
* **Confusion Matrix:**



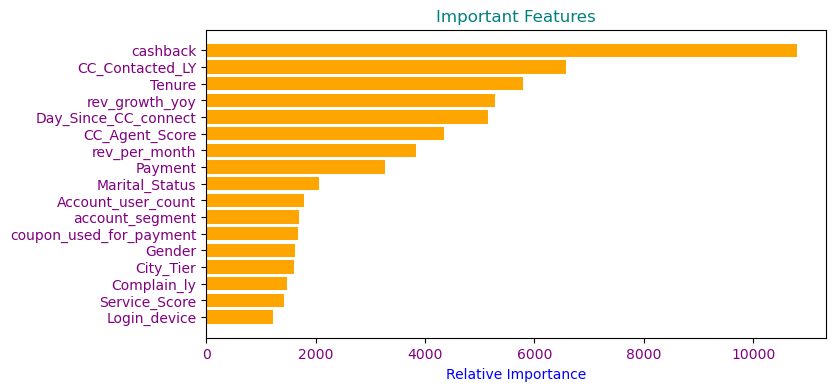
**Fig no - 23: CM LightGBM**

* **AUC - ROC Curve:**
* AUC for the Training Data: 1.000
* AUC for the Test Data: 0.996



**Fig no - 24: ROC Curve LightGBM**

* **Observations:**
* **AUC for Train data is 100% and AUC for Test data is 99%.**
* **The AUC & ROC curve shows that it is covering the same area for Train and Test data. Therefore, this is considered a Generalized good model.**
* **Also, the AUC score is very high and is able to distinguish between Positive and Negative classes very well.**
* **Feature Importance:**



**Fig no 25: Feature Imp LightGBM**

* **The 1st five Important features of LightGBM Model are: cashback, CC\_Contacted\_LY, Tenure, rev\_growth\_yoy and Day\_Since\_CC\_connect.**

**Model Comparison**

**Comparing the Models built by Plotting the ROC-AUC Curve for Test data on Base model and Tuned Model.**

### Base Models

### 

### ****Fig no 26: ROC Curve Comparison****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models | **Logistic Regression** | **LDA** | **SVC** | **KNN** | **Naïve Bayes** | **Decision Tree** |
|  |
| AUC | **0.86** | **0.85** | **0.69** | **0.95** | **0.81** | **0.91** |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models | **Random Forest** | **Bagging Classifier** | **Ada Boost** | **Gradient Boost** | **XGBoost** | **LightGBM** |
|  |
| AUC | **0.99** | **0.98** | **0.91** | **0.93** | **0.99** | **0.99** |  |

* **Interpretations:**
* **Comparing the AUC Score we can say that Random Forest, XGBoost and LightGBM Models are performing Better having a High AUC score of 99% on Test data.**
* **These Models are able to Separate between the Churn and non-churn Classes Very well.**
* **These models can be considered a Good Generalized model.**
* **Interpretations:**
* **Among all Built models, LightGBM model is giving Highest Accuracy of 98% on Test data and Highest AUC score of 99%.**
* **So, LightGBM is the most Optimal model in my case.**

**Business Insights, Implications and Recommendations**

* **Introducing Personalized offers or Promotions can decrease Churn rate.**
* **Tier-1 customers have high churn rate, suggesting some incentives or discounted offer can help in Retaining customers.**
* **UPI and E-wallet payment mode should be encouraged to increase customers.**
* **Tenure period should be Maximized to Retain the customers.**
* **Customer care services should be improvised so that the customers complaint can be resolved quickly and efficiently.**
* **Customized services should be offered to different account segment customers to retain customers.**
* **Launch tailoring offers to multiuser account holders to retain customers.**
* **The 1st five Important features of LightGBM Model are: cashback, CC\_Contacted\_LY, Tenure, rev\_growth\_yoy and Day\_Since\_CC\_connect.**
* **These 5 features are Contributing most towards the Accuracy of the LightGBM Model.**
* **Since these 5 features are contributing most towards the Accuracy, the business should ensure that these features should not contain any Anomalies, Null values or any kind of unwanted characters as a part of these columns in the future.**
* **Also, we should make sure these 5 features are made compulsory for model building and accurate data are captured.**
* **The business can make use of this model to identify customers who may Churn in future.**
* **Introducing Personalized offers or Promotions can decrease Churn rate.**
* **Tier-1 customers have high churn rate, suggesting some incentives or discounted offer can help in Retaining customers.**
* **UPI and E-wallet payment mode should be encouraged to increase customers.**
* **Tenure period should be Maximized to Retain the customers.**
* **Customer care services should be improvised so that the customers complaint can be resolved quickly and efficiently.**
* **Customized services should be offered to different account segment customers to retain customers.**
* **Launch tailoring offers to multiuser account holders to retain customers.**

**Thank You**