CUSTOMER CHURN PREDICTION

CAPSTONE PROJECT

FINAL REPORT

-Keerthiga Sekar

Contents

List of Figures

1.INTRODUCTION OF THE BUSINESS PROBLEM

A) DEFINING PROBLEM STATEMENT

BACKGROUND:

A Direct-to-Home (DTH) provider is experiencing increased competition in the market, which is making customer retention a significant challenge. Customer churn, where accounts are closed, results in the loss of multiple customers per account. This is particularly critical as each account can host multiple customers. The company wants to proactively identify potential churners and provide them with segmented offers to prevent them from leaving.

OBJECTIVE:

Our objective is to develop a churn prediction model that accurately identifies accounts at risk of churning. Based on the model's predictions, create targeted campaign recommendations to retain these at-risk accounts without significantly impacting the company's revenue.

NEED OF THE STUDY/PROJECT

This study/project is very essential for the client to plan for future in terms of product designing, sales or in rolling out different offers for different segment of clients. The outcome of this project will give a clear understanding where the firm stands now and what's the capacity it holds in terms for taking risk. It will also denote what's the future prospective of the organization and how they can make it even better and can plan better for the same and can help them retaining customers in a longer run.

UNDERSTANDING BUSINESS/SOCIAL OPPORTUNITY

This a case study of a DTH company where in they have customers assigned with unique account ID and a single account ID can hold many customers (like family plan) across gender and marital status, customers get flexibility in terms of mode of payment they want to opt for. Customers are again segmented across various types of plans they opt for as per their usage which also based on the device they use (computer or mobile) moreover they have cashbacks on bill payment.

The overall business runs in customers loyalty and stickiness which in-turn comes from providing quality and value-added services. Also, running various promotional and festivals offers may help organization in getting new customers and also retaining the old one.

We can conclude that a customer retained is a regular income for organization, a customer added is a new income for organization and a customers lost will be a negative impact as a single account ID holds multiple number of customers i.e.; closure of one account ID means loosing multiple customers.

It's a great opportunity for the company as it's a need of almost every individual of family to have a DTH connection which in-turn also leads to increase and competition.

It costs about **five times more to acquire a new customer** than to keep an existing one.So, investing in keeping current customers happy is not just good practice, it's also a smart financial move.

Just a **5% increase in customer retention** can boost profits by **25-95%**. That's a huge potential gain just by focusing on keeping our current customers satisfied.

When we're selling to customers we already have, our success rate is **60-70%**. In contrast, selling to new customers has a success rate of just **5-20%**. So, it's much easier to sell to someone who already trusts us.

U.S. companies lose **$136.8 billion each year** because of avoidable consumer switching. That's a massive amount we could be saving if we focused on retaining our customers.

In a nutshell, keeping current customers happy and engaged isn't just nice, it's a smart business strategy that can save money and increase profits significantly.

DATA REPORT

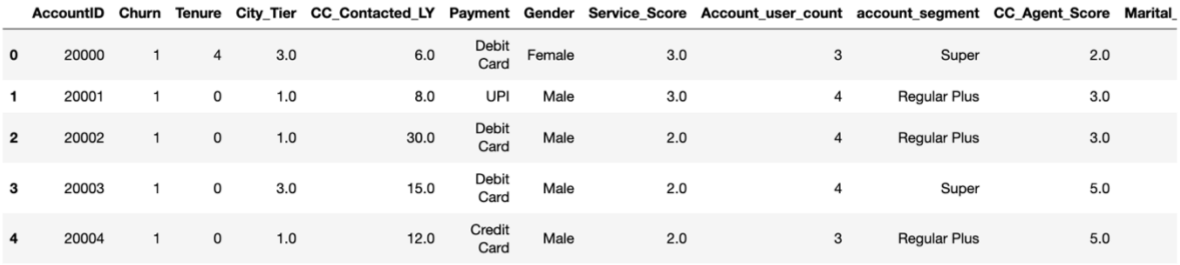
Dataset of problem: - Customer Churn Data Dictionary: -

- AccountID -- account unique identifier
- Churn -- account churn flag (Target Variable)
- 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 months)
- coupon_used_for_payment -- 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

DATA INGESTION

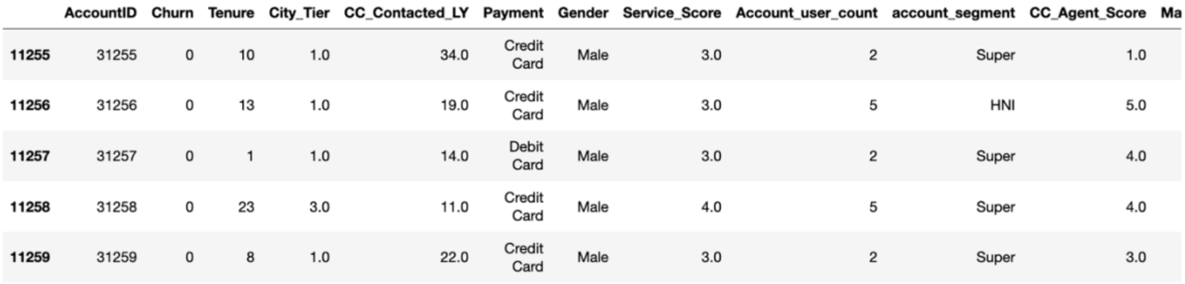
- Loaded the required packages, set the work directory and load the datafile.
- Data set has 11,260 number of records and 19 features (18 independent and 1 dependent or target variable).

STRUCTURE OF THE DATA



	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_
0	20000	1	4	3.0	6.0	Debit Card	Female	3.0	3	Super	2.0	
1	20001	1	0	1.0	8.0	UPI	Male	3.0	4	Regular Plus	3.0	
2	20002	1	0	1.0	30.0	Debit Card	Male	2.0	4	Regular Plus	3.0	
3	20003	1	0	3.0	15.0	Debit Card	Male	2.0	4	Super	5.0	
4	20004	1	0	1.0	12.0	Credit Card	Male	2.0	3	Regular Plus	5.0	

Table 1: Top 5 rows of the dataset

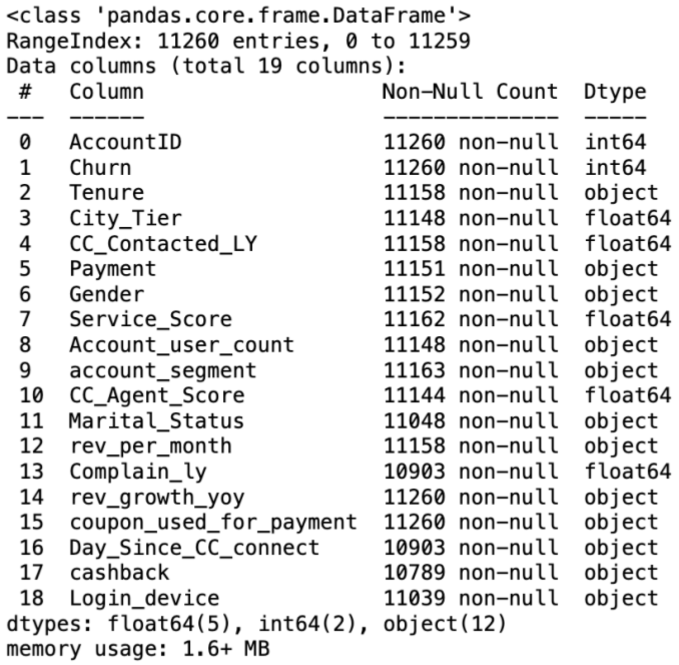


	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Ma
11255	31255	0	10	1.0	34.0	Credit Card	Male	3.0	2	Super	1.0	
11256	31256	0	13	1.0	19.0	Credit Card	Male	3.0	5	HNI	5.0	
11257	31257	0	1	1.0	14.0	Debit Card	Male	3.0	2	Super	4.0	
11258	31258	0	23	3.0	11.0	Credit Card	Male	4.0	5	Super	4.0	
11259	31259	0	8	1.0	22.0	Credit Card	Male	3.0	2	Super	3.0	

Table 2: Last 5 rows of the dataset

UNDERSTANDING HOW DATA WAS COLLECTED IN TERMS OF TIME, FREQUENCY AND METHODOLOGY

- Data has been collected for random 11,260 unique account ID, across gender and marital status.
- Looking at variables “CC_Contacted_LY”, “rev_per_month”, “Complain_ly”, “rev_growth_yoy”, “c oupon_used_for_payment”, “Day_Since_CC_connect” and “cashback” ,we can conclude that the data has been collected for last 12 month.
- Data has 19 variables, 18 independent and 1 dependent or the target variable, which shows if customer churned or not.
- The data is the combination of services ,customers who are using along with their payment option and also basic individuals details as well.
- Data is mixed of categorical as well as continuous variables.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11260 entries, 0 to 11259
Data columns (total 19 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   AccountID                11260 non-null  int64  
 1   Churn                    11260 non-null  int64  
 2   Tenure                   11158 non-null  object 
 3   City_Tier                11148 non-null  float64
 4   CC_Contacted_LY          11158 non-null  float64
 5   Payment                  11151 non-null  object 
 6   Gender                   11152 non-null  object 
 7   Service_Score            11162 non-null  float64
 8   Account_user_count       11148 non-null  object 
 9   account_segment          11163 non-null  object 
 10  CC_Agent_Score           11144 non-null  float64
 11  Marital_Status           11048 non-null  object 
 12  rev_per_month            11158 non-null  object 
 13  Complain_ly              10903 non-null  float64
 14  rev_growth_yoy           11260 non-null  object 
 15  coupon_used_for_payment  11260 non-null  object 
 16  Day_Since_CC_connect     10903 non-null  object 
 17  cashback                 10789 non-null  object 
 18  Login_device             11039 non-null  object 
dtypes: float64(5), int64(2), object(12)
memory usage: 1.6+ MB
```

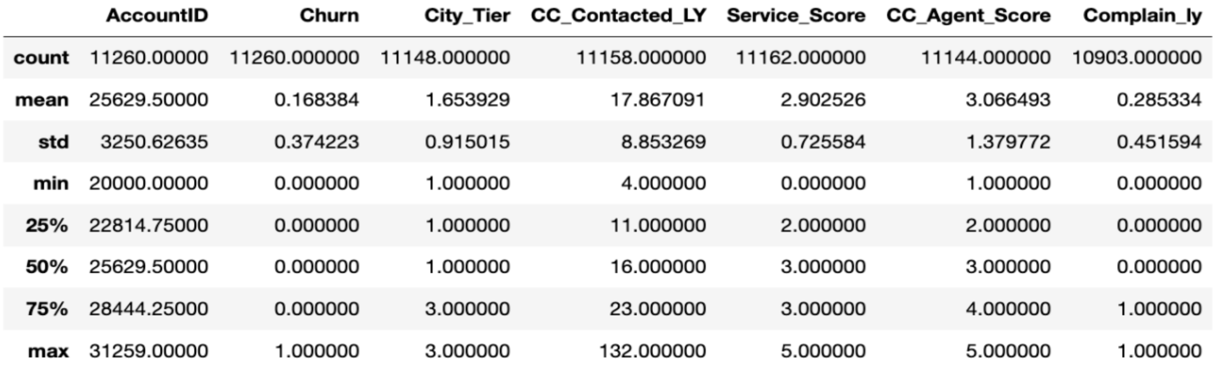
Table 3: Basic information of the dataset

OBSERVATION:

- The data frame has 11260 rows and 19 columns.
- There is a total of 11260 non-null observations in each of the columns except some columns.
- The missing values are treated.
- In these 19 columns, 2 are of integer type, 5 are of float type and 12 are of object type.

VISUAL INSPECTION OF DATA

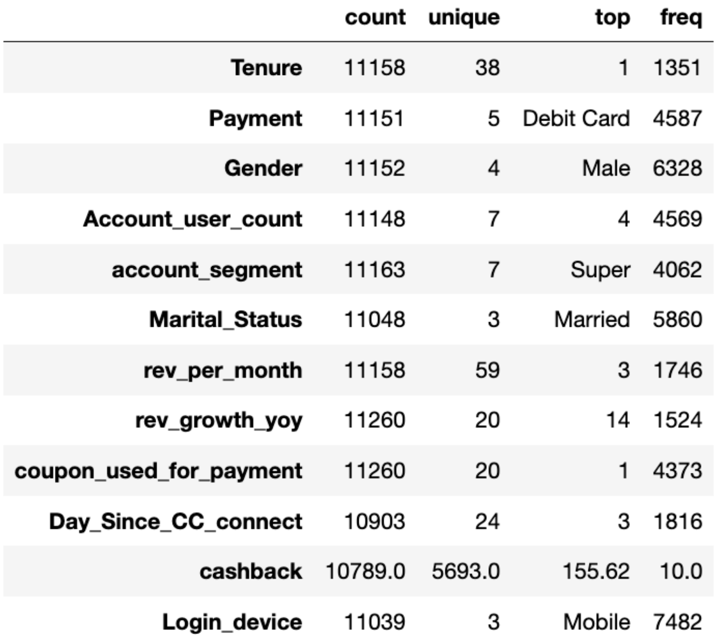
THE STATISTICAL SUMMARY FOR THE NUMERICAL VARIABLES



	AccountID	Churn	City_Tier	CC_Contacted_LY	Service_Score	CC_Agent_Score	Complain_ly
count	11260.00000	11260.000000	11148.000000	11158.000000	11162.000000	11144.000000	10903.000000
mean	25629.50000	0.168384	1.653929	17.867091	2.902526	3.066493	0.285334
std	3250.62635	0.374223	0.915015	8.853269	0.725584	1.379772	0.451594
min	20000.00000	0.000000	1.000000	4.000000	0.000000	1.000000	0.000000
25%	22814.75000	0.000000	1.000000	11.000000	2.000000	2.000000	0.000000
50%	25629.50000	0.000000	1.000000	16.000000	3.000000	3.000000	0.000000
75%	28444.25000	0.000000	3.000000	23.000000	3.000000	4.000000	1.000000
max	31259.00000	1.000000	3.000000	132.000000	5.000000	5.000000	1.000000

Table 4: Statistical Summary for The Numerical Variables

THE STATISTICAL SUMMARY FOR THE CATEGORICAL VARIABLES



	count	unique	top	freq
Tenure	11158	38	1	1351
Payment	11151	5	Debit Card	4587
Gender	11152	4	Male	6328
Account_user_count	11148	7	4	4569
account_segment	11163	7	Super	4062
Marital_Status	11048	3	Married	5860
rev_per_month	11158	59	3	1746
rev_growth_yoy	11260	20	14	1524
coupon_used_for_payment	11260	20	1	4373
Day_Since_CC_connect	10903	24	3	1816
cashback	10789.0	5693.0	155.62	10.0
Login_device	11039	3	Mobile	7482

Table 5: Statistical Summary for The Categorical Variables

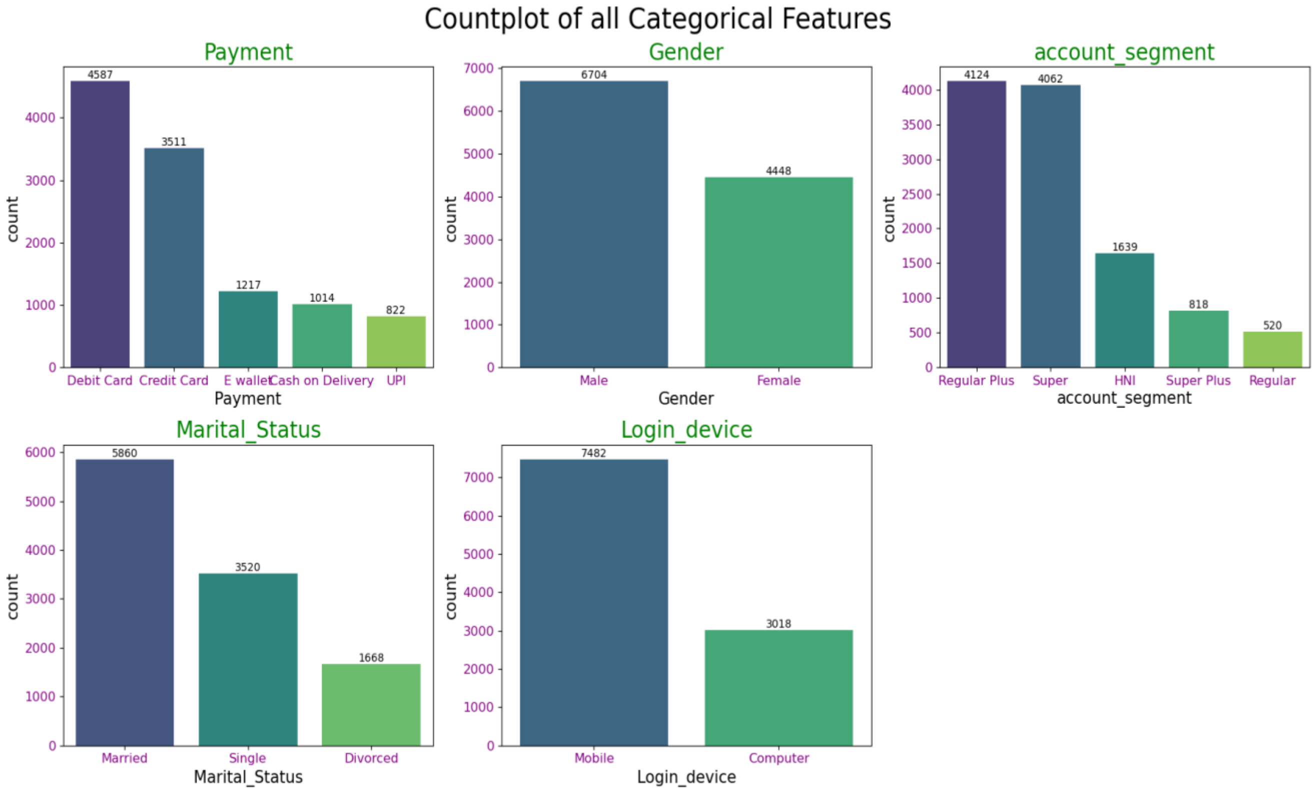
OBSERVATION:

1. This shows description of variation in various statistical measurements across variables which denotes that each variable is unique and different.
2. The Service_Score ranges from 0 to 5, where:
   - **0**: Represents the lowest level of satisfaction, indicating very poor service.
   - **5**: Represents the highest level of satisfaction, indicating excellent service.
   - Scores closer to 5 are assumed to indicate higher customer satisfaction with the service provided by the company.
   - Scores closer to 0 indicate dissatisfaction.
   - A score of 4 is better than 3 but not as good as 5.

2.EXPLORATORY DATA ANALYSIS

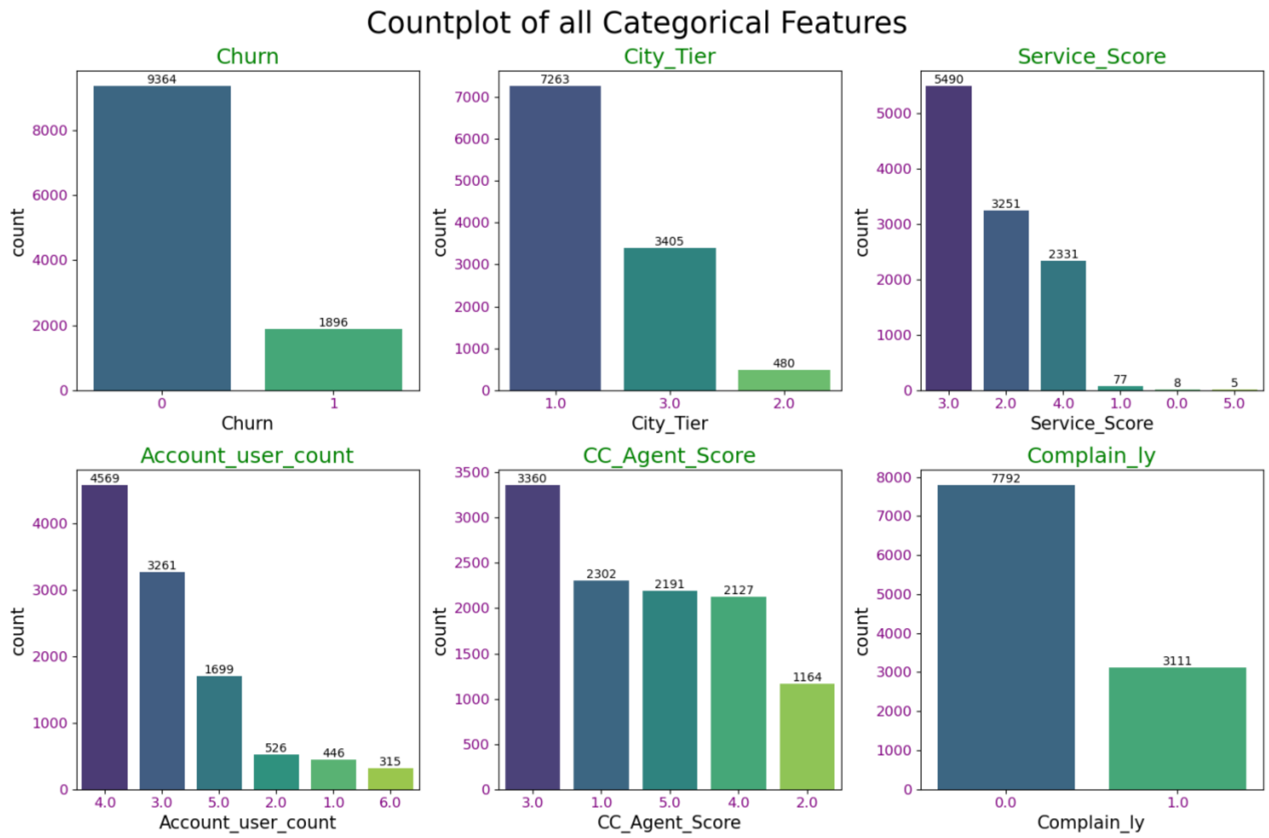
UNIVARIATE ANALYSIS (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

Univariate Analysis-Categorical Variable:



OBSERVATION:

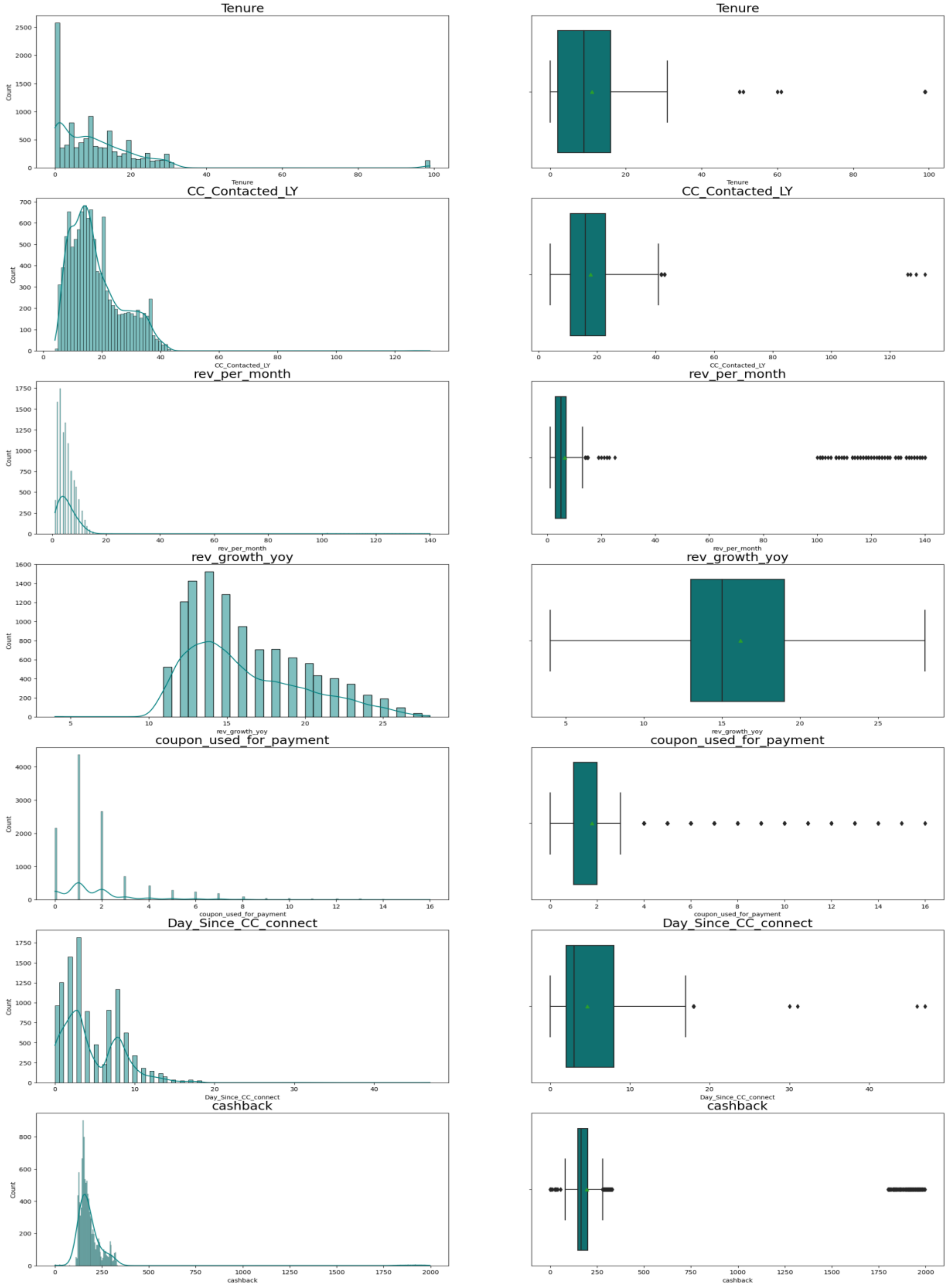
- Most Customers Preferred Payment mode are Debit card and Credit Card.
- Very few customers make payment through UPI.
- Male customers are more as compared to female.
- There are more customers that belong to account segment- Regular plus and Super.
- Very less customers belongs to Regular segment based 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.



OBSERVATION:

- 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.

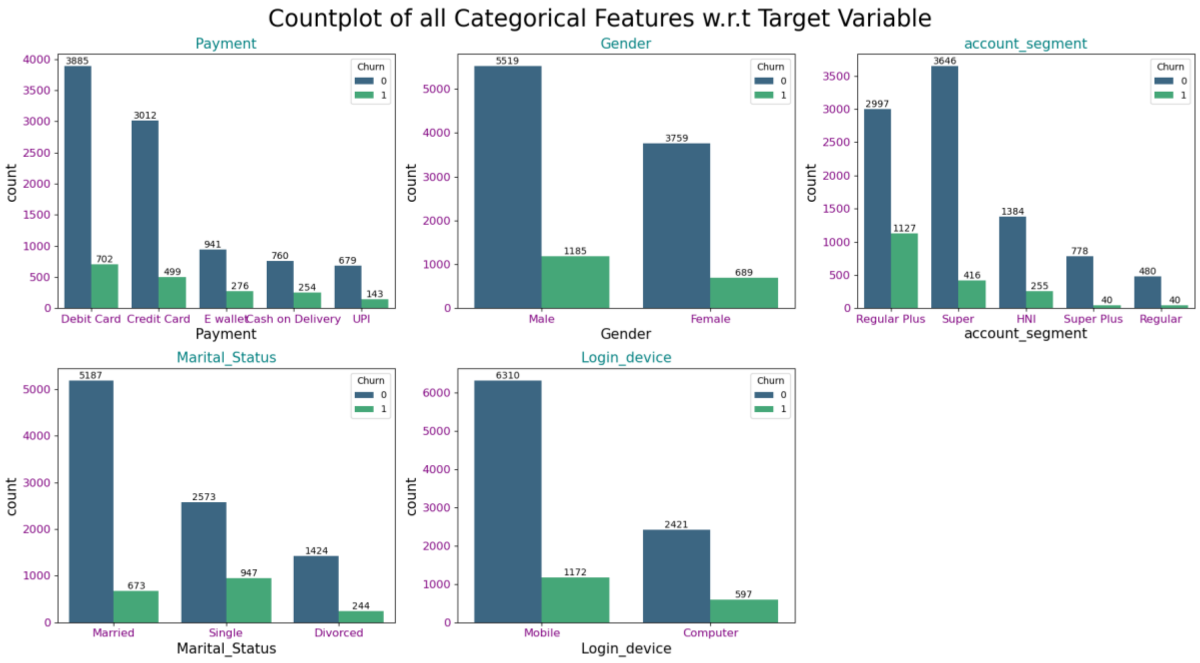
Histogram and Boxplot of all continuous variables



OBSERVATION:

- 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.

Bivariate analysis



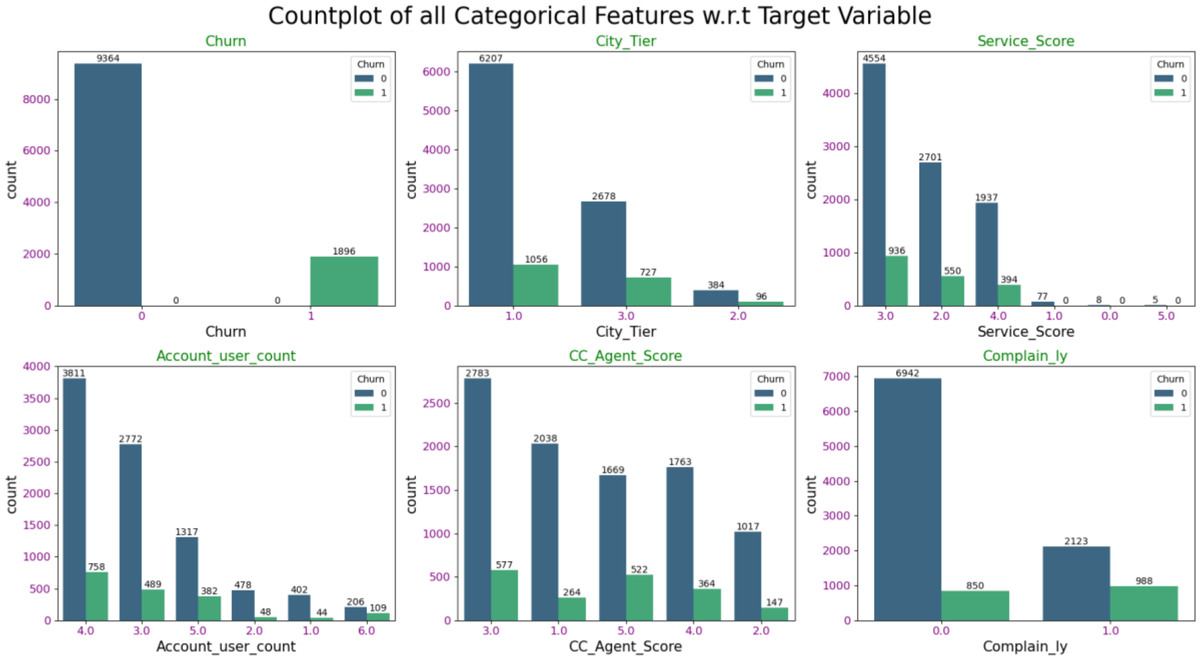
OBSERVATIONS:

- The proportions of Churners are more of Male customers as compared to Female.

- Most of the churners make payment via Debit Card. Maximum number of customers prefer debit and credit card as their preferred mode of payment. The churning rate is high for cash on delivery customers.

- The ratio of male customers are higher when compared to female and the churning rate is high for male customers

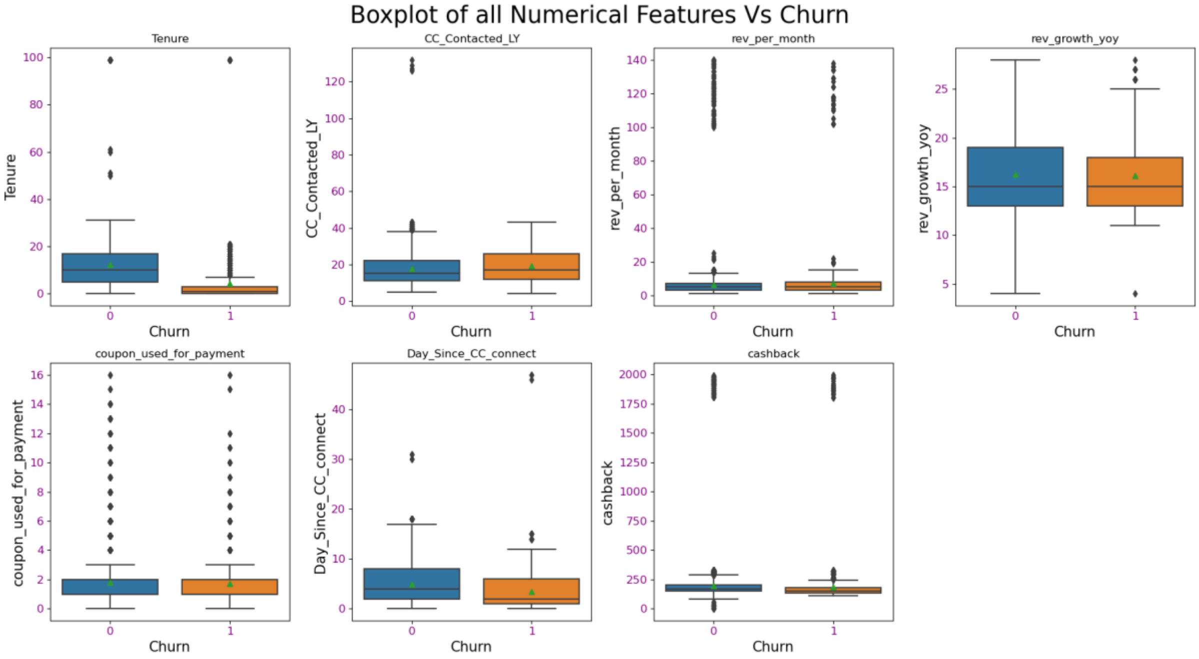
- Non-churners preferable mode of payment are Debit cards and Credit Cards. Most of the churners belongs to Regular plus and Super account segment.
- Maximum non-Churners belongs to Super Segment followed by Regular plus.
- Most of the churners are single.
- Non-Churners are Maximum Married Couples.
- Most Preferred Login Device for both Churners and Non-Churners is Mobile since it's handy.



OBSERVATION

- Maximum customers are form tier 1 city. Tier 1 city is the metro city, which indicates the high number of population density in this city type.
- 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.

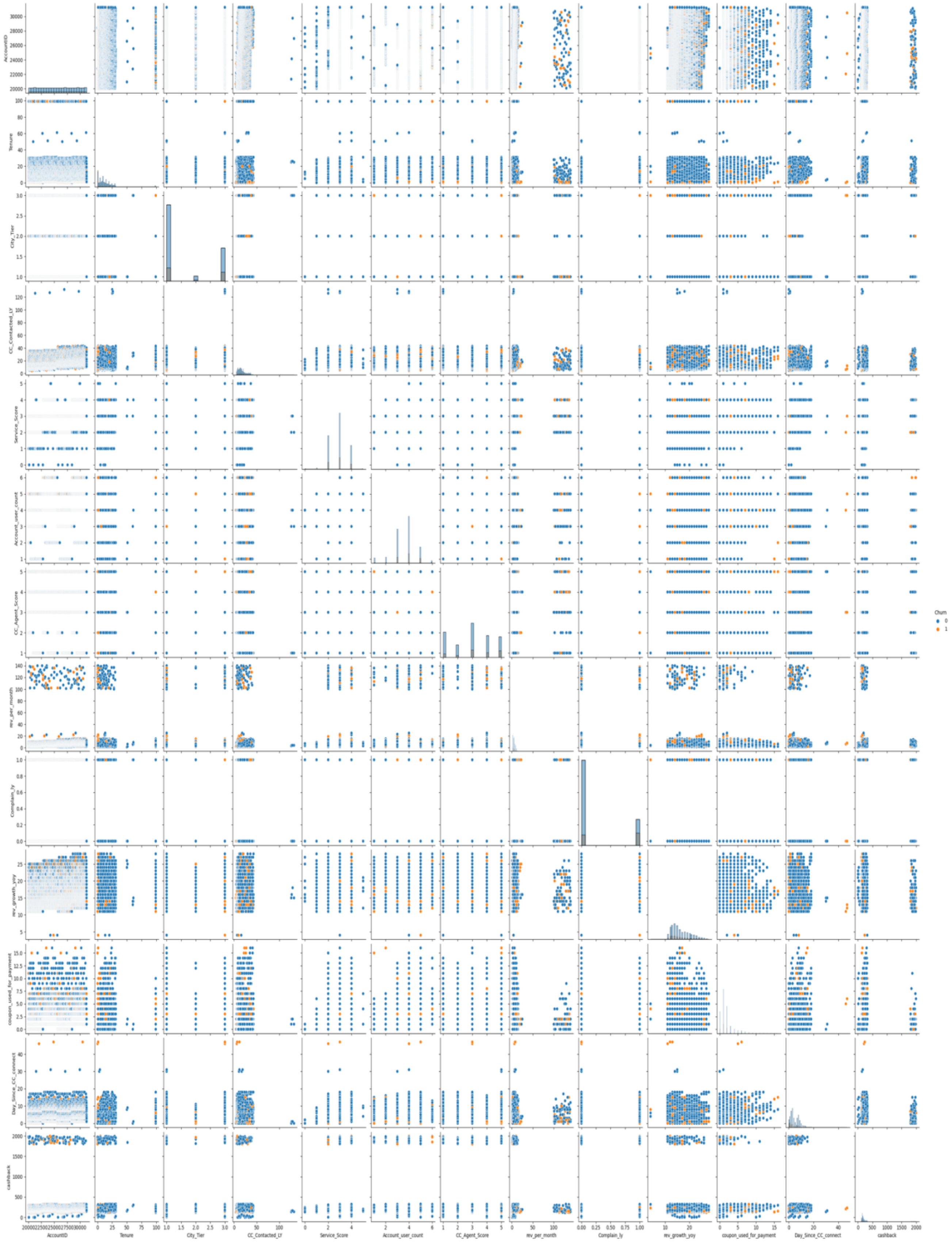
- It is evident from the fact that Customers who have raised the complaint maximum no of times are mostly non-churners. This shows that raising complaint have solved their issue and hence made them retain the use of service and decreasing churn Rate.



OBSERVATION:

- 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.

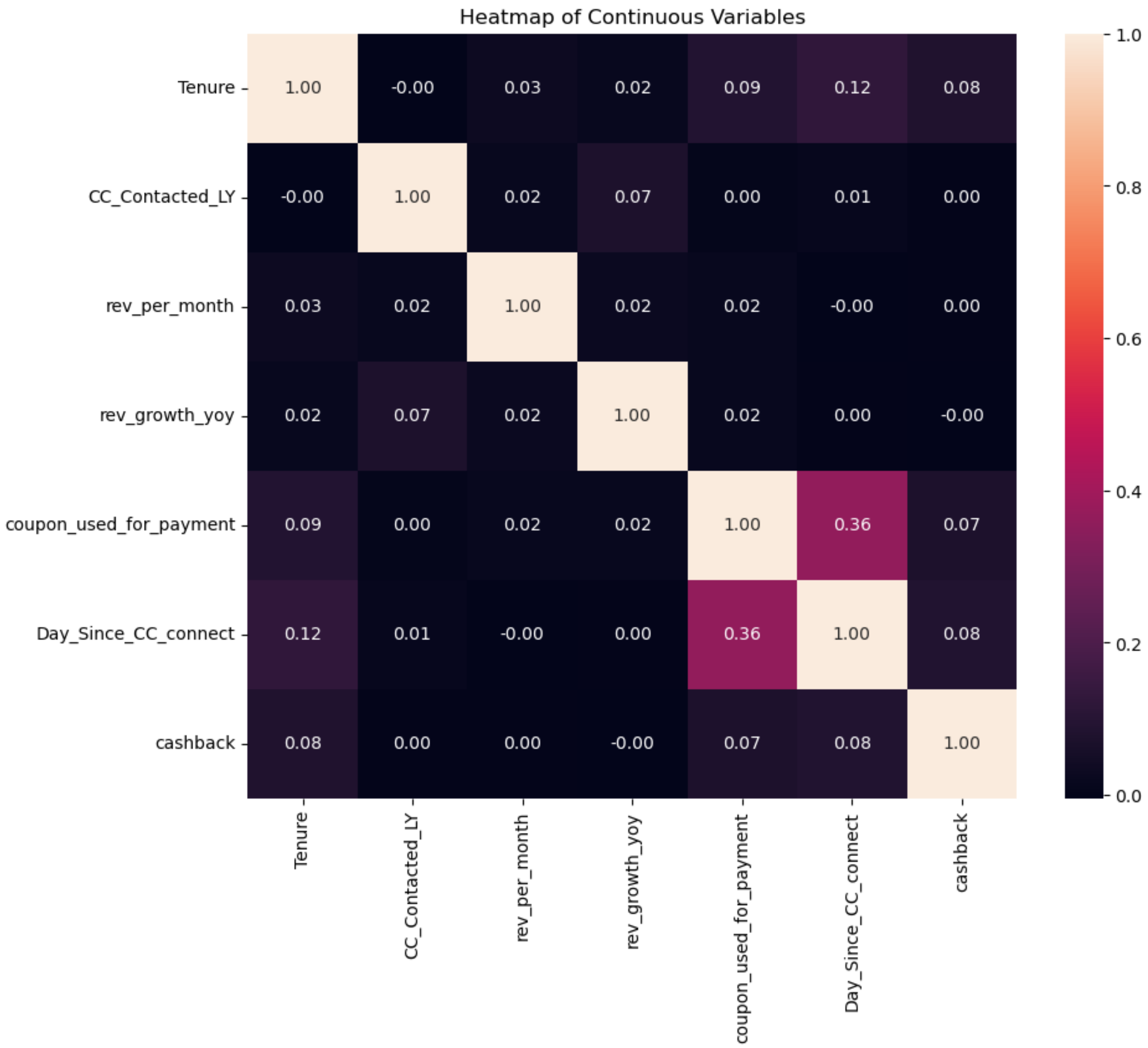
MULTIVARIATE ANALYSIS



OBSERVATION:

- 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.

Correlation Heatmap:



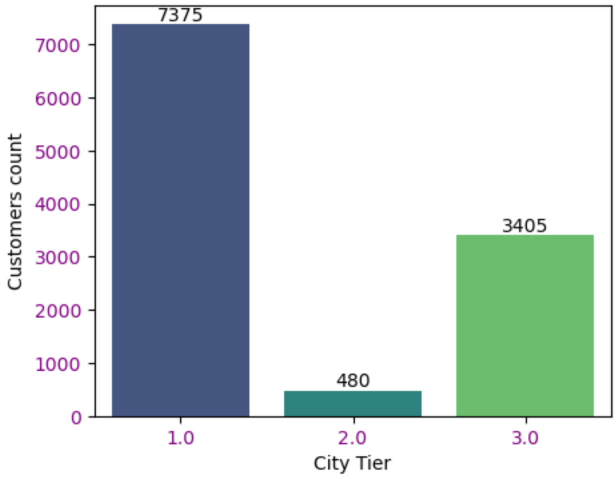
OBSERVATION:

- The heatmap suggests that the continuous variables are generally independent of one another, except for a moderate relationship between **Day_Since_CC_connect** and **coupon_used_for_payment**.
- No pair of variables appears to have a strong linear relationship, as indicated by most of the correlation values being small.

Data Visualization using Segmentation

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

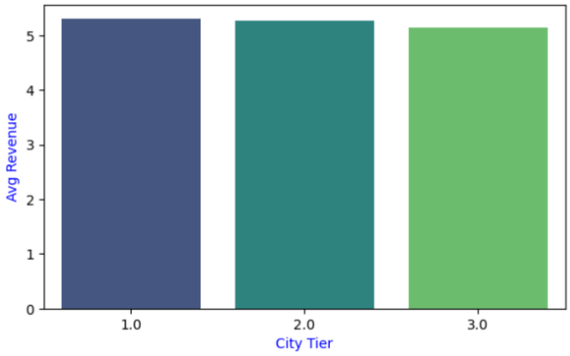
There are 3 city tiers mentioned in the dataset. Generally, tier 1 cities are considered as the major metro cities where the people tend to use more DTH Services. So, accordingly, can we say that tier 1 city customers tend to generate more avg revenue as compared to tier 2 and tier 3 city customers.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

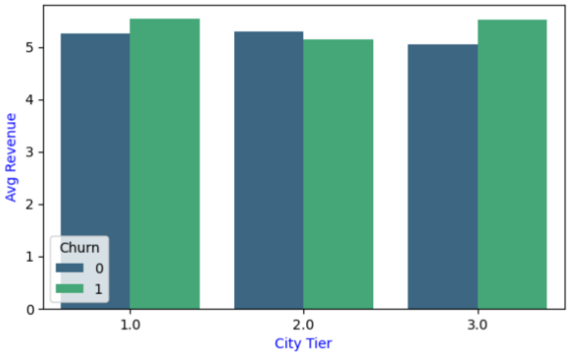


Insights:

- Count of customers are more in Tier-1 followed by Tier-3.
- Very less customers belong to Tier-2**.**

City Tier Vs Monthly Average Revenue:





Insights:

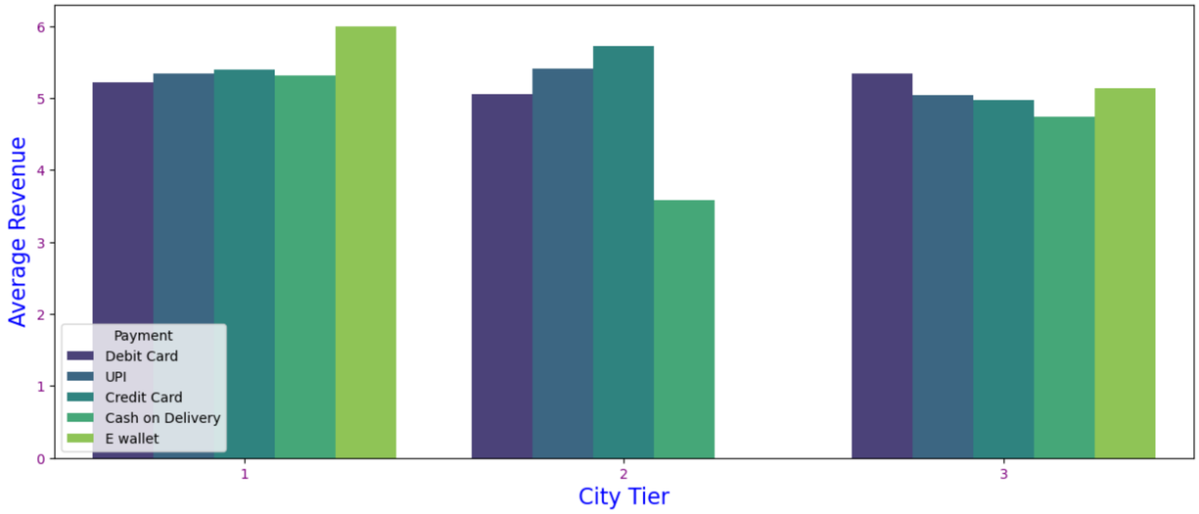
- We can see that Average Revenue Generation is same across all the City Tiers.
- Though the Customers using the Services in Tier-2 is very less as compared to Tier-1 & 3, But the Average Revenue generated per month by the account is same across all the Tiers.

- 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

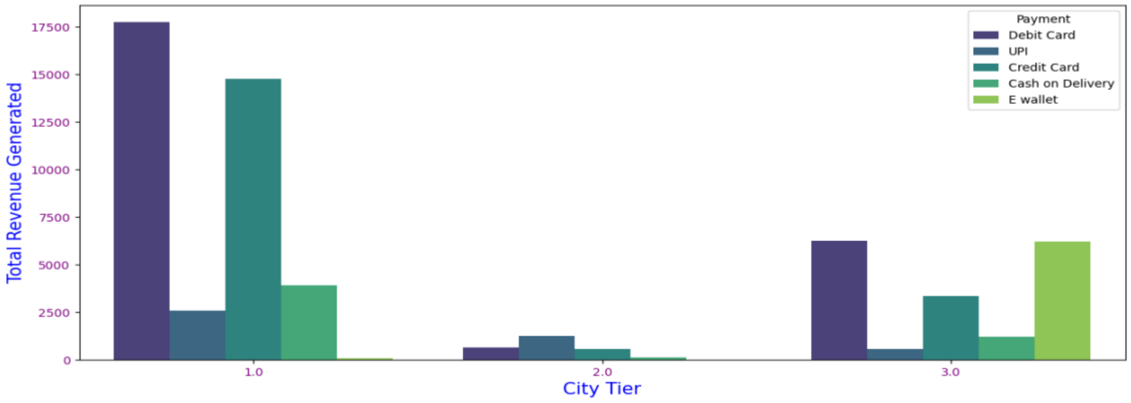
There are different payment modes (CC, DC, COD, E-wallet & UPI). Depending on the city tiers, Let's Visualize and find the preferred payment modes used by the customers**.**

- 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.



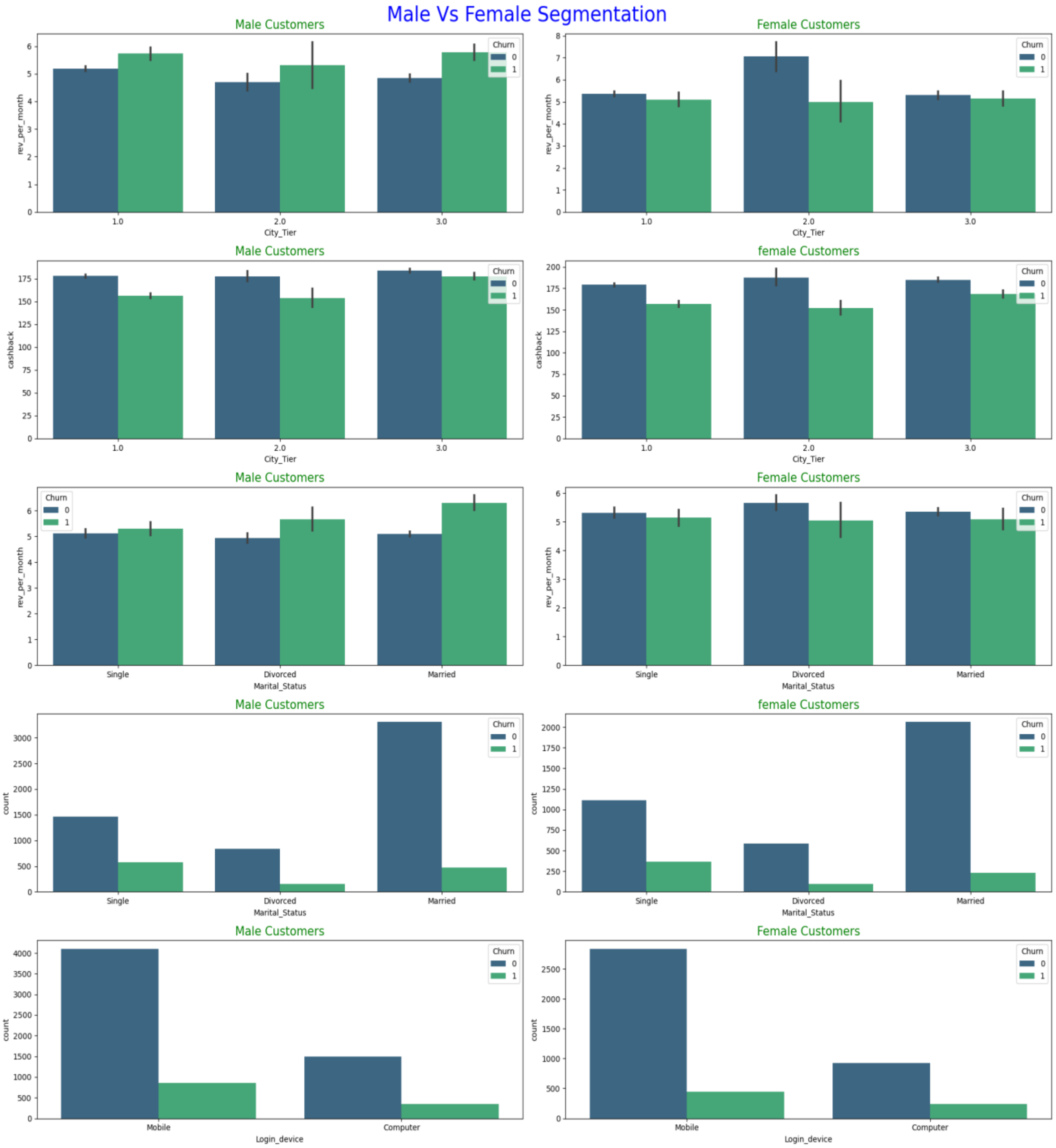
Insights:

- 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.



- 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.
- Almost 35% of the tier 3 city customers prefers E wallet as their payment mode, most revenue will be generated from such customers from tier 3 cities.
- Almost 46% of the tier 1 city customers prefers Debit Card as the payment mode, so most revenue will be generated from such customers from tier 1 cities.
- Almost 47% of the tier 2 city customers prefers UPI as the payment mode, so most revenue will be generated from such customers from tier 2 cities.

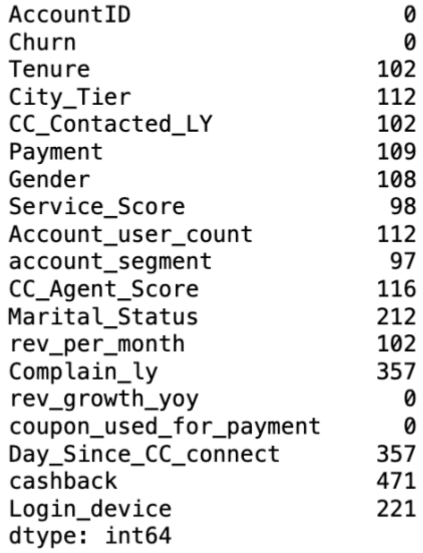
GENDER



- In Tier-1 & Tier-2, avg monthly revenue is more generated by Male customer than Females and in Tier-3 by Females.
- Cashback ratio is same for male & female in all Tier and across Marital status.
- Avg revenue generated by Singles Male and Females are almost same in all Tiers.
- Revenue generated by Divorced Females are slightly more than male.
- Count of Marital status is same across Genders.
- Mobile is the most preferred Login device among Male and Females.

3.Data Cleaning and Pre-processing

CHECKING NULL VALUES

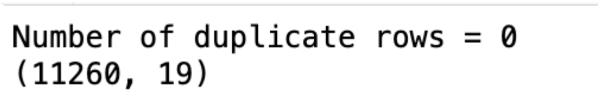


```
AccountID                    0
Churn                        0
Tenure                     102
City_Tier                  112
CC_Contacted_LY            102
Payment                    109
Gender                     108
Service_Score               98
Account_user_count         112
account_segment             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
```

OBSERVATION:

Except variables “AccountID”, “Churn”, “rev_growth_yoy” and “coupon_used_for_payment” all other variables have null values present.

NULL VALUE CHECK



```
Number of duplicate rows = 0
(11260, 19)
```

OBSERVATION:

There is no duplicate observations.

UNDERSTANDING OF ATTRIBUTES (VARIABLE INFO, RENAMING IF REQUIRED)

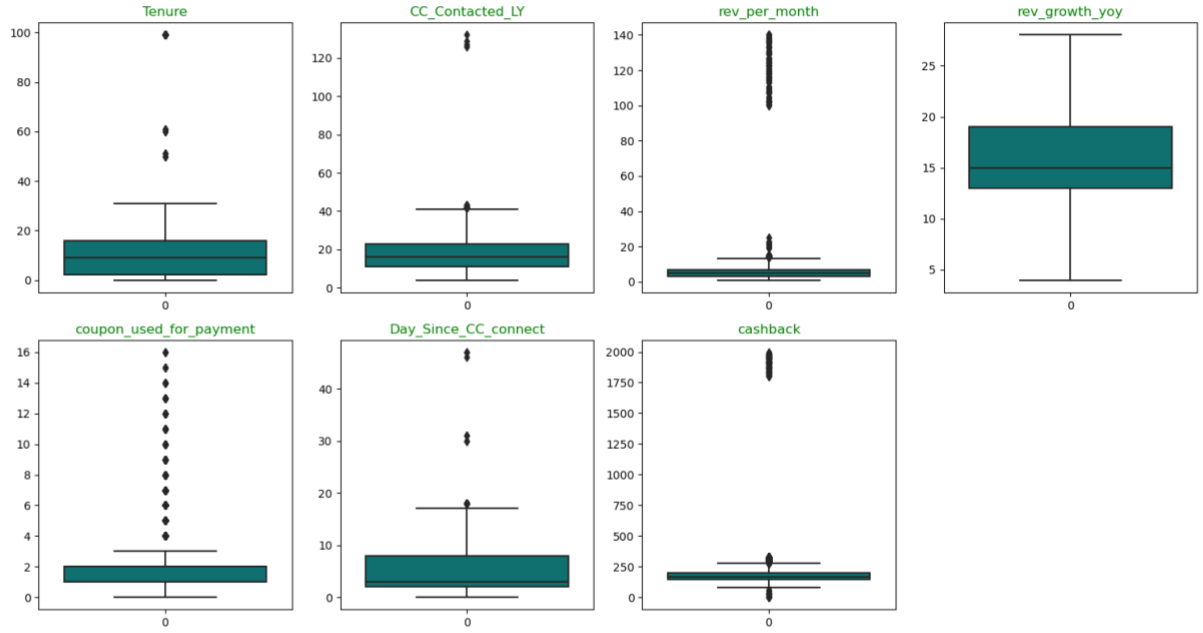
This dataset has 18 attributes contributing towards the target variable. Let’s discuss about these variables one after another.

1. AccountID – This variable represents a unique ID which represents a unique customer. This is of Integer data type and there is no null values present in this.
2. Churn – This is our target variable, which represents if customer has churned or not. This is categorical in nature will no null values. “0” represents “NO” and ”1” represents “YES”.
3. Tenure – This represents the total tenure of the account since opened. This is a continuous variable with 102 null values.
4. City_Tier – These variable segregates customer into 3 parts based on city the primary customer resides. This variable is categorical in nature and have 112 null values.
5. CC_Contacted_LY – This variable represents the number of times all the customers of the account has contacted customer care in last 12months. This variable is continuous in nature and have 102 null values.
6. Payment – This variable represents the preferable mode of bill payment opted by customer. This is categorical in nature and have 109 null values.
7. Gender – This variable represents the gender of the primary account holder. This is categorical in nature and 108 null values.
8. Service_Score – Scores provided by the customer basis the service provided by the company. This variable is categorical in nature and have 98 null values.
9. Account_user_count – This variable gives the number of customers attached with an accountID. This is continuous in nature and have 112 null values.
10. account_segment – These variable segregates customers into different segment basis their spend and revenue generation. This is categorical in nature and have 97 null values.
11. CC_Agent_Score -- Scores provided by the customer basis the service provided by the customer care representative of the company. This variable is categorical in nature and have 116 null values.
12. Marital_Status – This represents marital status of the primary account holder. This is categorical in nature and have 212 null values.
13. rev_per_month – This represents average revenue generated per account ID in last 12 months. This variable is continuous in nature and have 102 null values.
14. Complain_ly – This denotes if customer have raised any complaints in last 12 months. This is categorical in nature and have 357 null values.
15. rev_growth_yoy – This variable shows revenue growth in percentage of account for 12 months Vs 24 to 13 months. This is continuous in nature and doesn’t have any null values.
16. coupon_used_lY – This represents the number of times customers have used discount coupons for bill payment. This is continuous in nature and doesn’t have any null values.
17. Day_Since_CC_connect – This represents the number of days since customer have contacted the customer care. Higher the number of days denotes better the service. This is continuous in nature and have 357 null values.
18. cashback– This variable represents the amount of cash back earned by the customer during bill payment. This is continuous in nature and have 471 null values. 19. Login_device – This variable represents in which device customer is availing the services if it’s on phone or on computer. This is categorical in nature and have 221 null values.

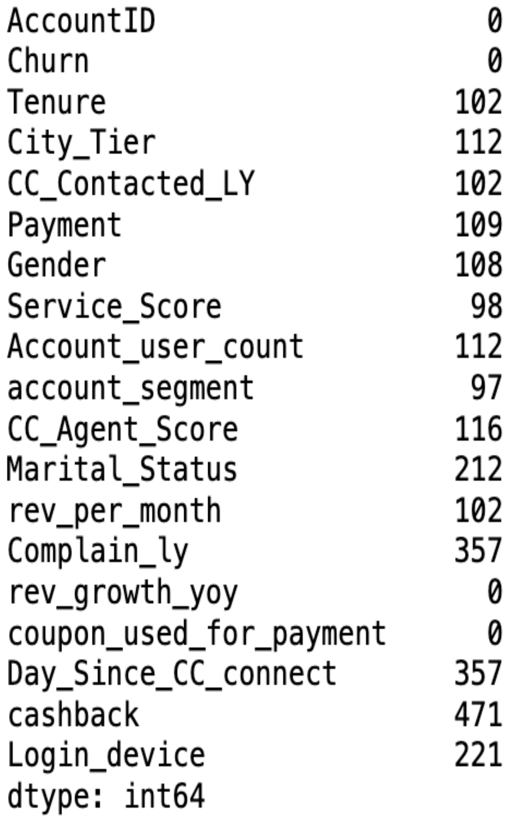
OBSERVATION:

- With above understanding of data, renaming of any of the variable is not required.
- We can move towards the EDA part where in we will understand the data little better along with treating bad data, null values and outliers.

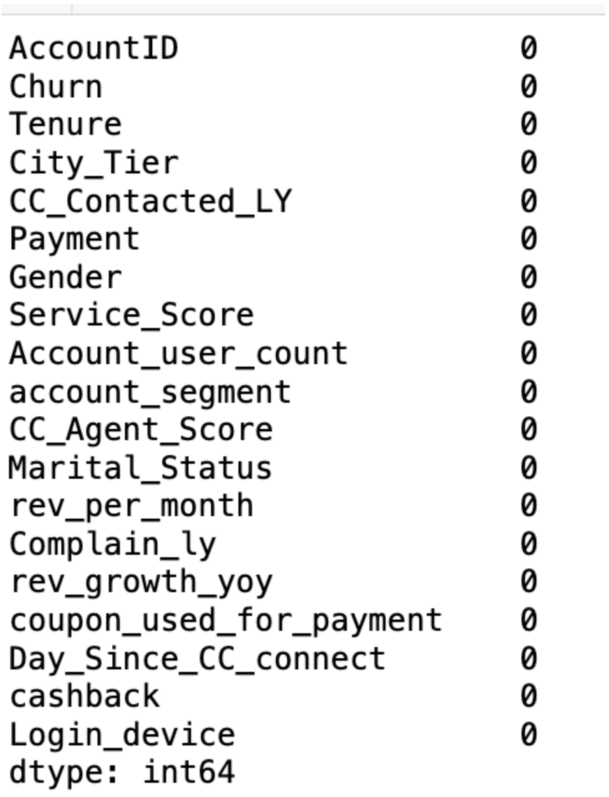
Checking for Outliers



HANDLING MISSING VALUES



```
AccountID                  0
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
```



```
AccountID                  0
Churn                      0
Tenure                     0
City_Tier                  0
CC_Contacted_LY            0
Payment                    0
Gender                     0
Service_Score              0
Account_user_count         0
account_segment            0
CC_Agent_Score             0
Marital_Status             0
rev_per_month              0
Complain_ly                0
rev_growth_yoy             0
coupon_used_for_payment    0
Day_Since_CC_connect       0
cashback                   0
Login_device               0
dtype: int64
```

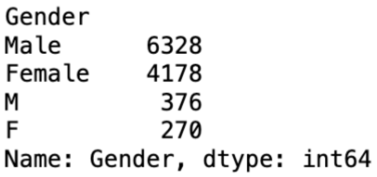
Before missing value treatment ⟶ **After missing value treatment**

Imputed null values with median for numerical variables and mode for categorical variables. Mean is impacted by extreme points. That is why mode is used for imputation. The data is clean with no null values.

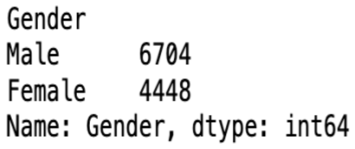
VARIABLE TRANSFORMATION

- The 'Gender' column has some inconsistencies with the representation of gender

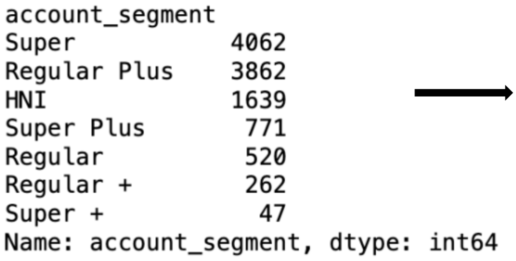
- The 'account_segment' column has similar categories that need to be standardized.
- The 'Login_device' column has an invalid entry '&&&&' that should be addressed.



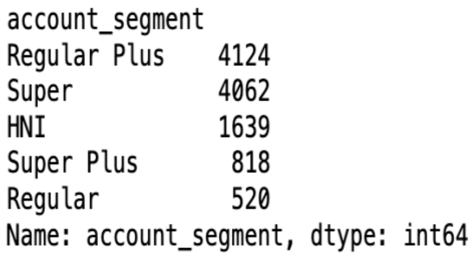
```
Gender
Male      6328
Female    4178
M          376
F          270
Name: Gender, dtype: int64
```



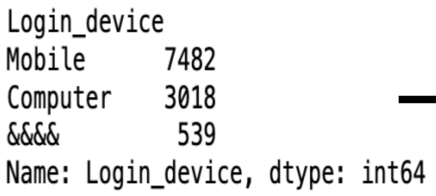
```
Gender
Male      6704
Female    4448
Name: Gender, dtype: int64
```



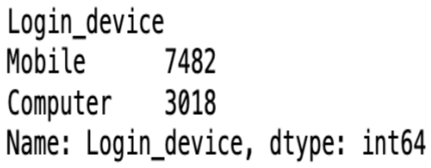
```
account_segment
Super           4062
Regular Plus    3862
HNI             1639
Super Plus       771
Regular          520
Regular +        262
Super +           47
Name: account_segment, dtype: int64
```



```
account_segment
Regular Plus    4124
Super           4062
HNI             1639
Super Plus       818
Regular          520
Name: account_segment, dtype: int64
```



```
Login_device
Mobile      7482
Computer    3018
&&&&         539
Name: Login_device, dtype: int64
```



```
Login_device
Mobile      7482
Computer    3018
Name: Login_device, dtype: int64
```

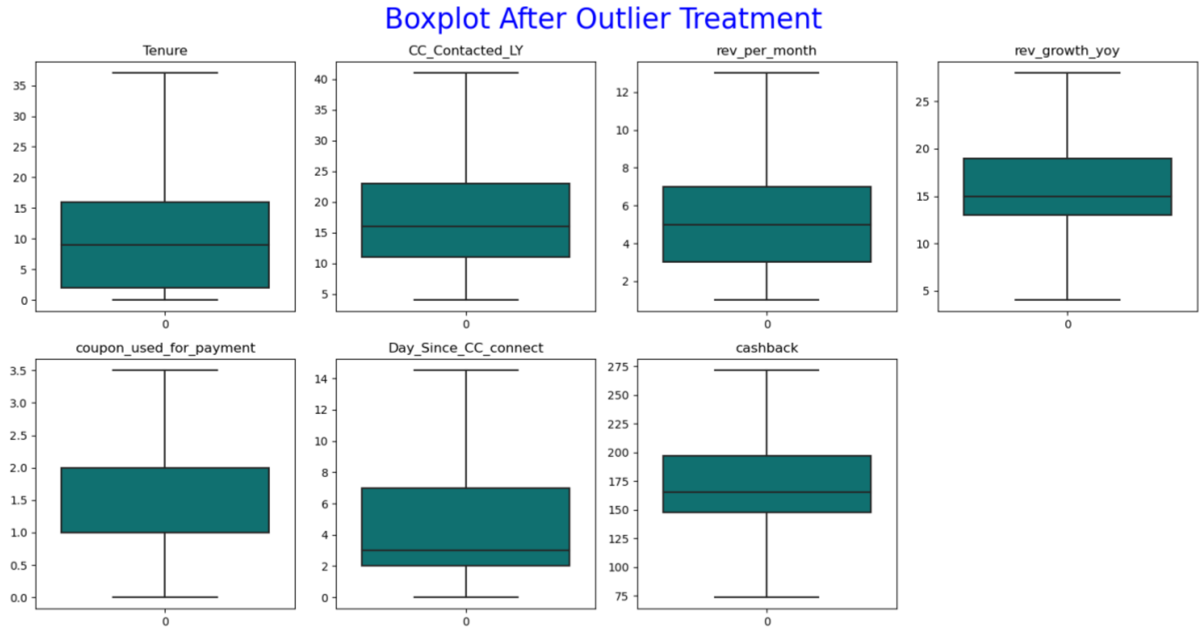
- Some of the columns like ‘Tenure’ is tranformed into Categorical column as ‘Tenure_Cat’ with buckets named low tenure,medium tenure, high tenure and very high tenure based on the duration.
- Column ‘CC_Contacted_LY’ is transformed as ‘CC_Contacted_LY_cat’ and its values are binned as Low contact,medium contach,high contact and special cases.
- Column ‘Revenue_Cat’ is feature engineered as ‘rev_per_month’ with buckets low revenue,medium revenue and high revenue.
- Column ‘Day_Since_CC_connect’ is transformed as ‘CC_connect_category’ with buckets very recent,recent, moderate,old.
- ‘Cashback’ column is featured as ‘cashback_category’ and its values are bucketed as low,medium and high.

REMOVAL OF UNWANTED VARIABLES

'AccountID','Tenure','CC_Contacted_LY','rev_per_month','Day_Since_CC_connect','cashback'.These variables are dropped from the dataset. AccoutID is dropped since it contains unique values, which will not be used for analytica; point of view. The other columns are dropped to reduce redundancy in the dataset.

OUTLIER TREATMENT

The Outliers are treated in this dataset. The columns are feature engineered and bifurcated into buckets.



ADDITION OF NEW VARIABLES

As of now,no new additional variables are created.Already existed variables are feature engineered based on some logic.The Categorical values in the data set in scaled using one hot encoding method in order to perform Clustering.

FEATURE SELECTION

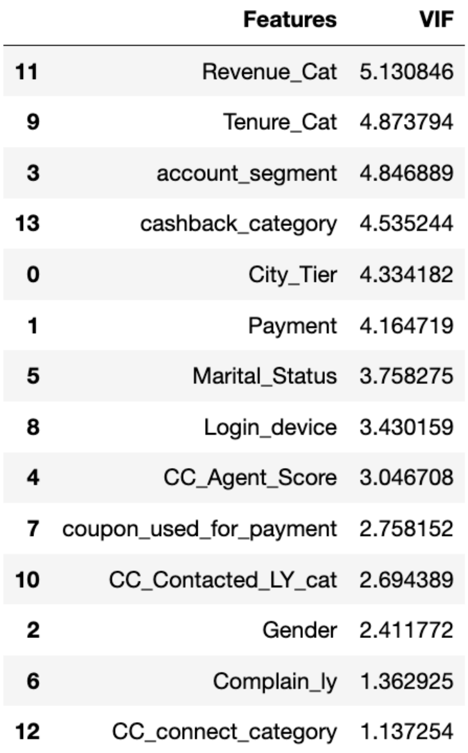
Used Variance Inflation Factor to select the important features using the VIF value. There are different ways of detecting (or testing) multicollinearity. One such way is Variation Inflation Factor.

Variance Inflation factor: Variance inflation factors measure the inflation in the variances of the regression coefficients estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient βk is "inflated" by the existence of correlation among the predictor variables in the model.

General Rule of Thumb:

If VIF is 1, then there is no correlation among the k th predictor and the remaining predictor variables, and hence, the variance of βk is not inflated at all. If VIF exceeds 5, we say there is moderate VIF, and if it is 10 or exceeding 10, it shows signs of high multi-collinearity. The purpose of the analysis should dictate which threshold to use.

Service_Score, rev_growth_yoy, Account_user_count, these are the variables dropped. The following are the features used after dropping some variables with VIF >5.



	Features	VIF
11	Revenue_Cat	5.130846
9	Tenure_Cat	4.873794
3	account_segment	4.846889
13	cashback_category	4.535244
0	City_Tier	4.334182
1	Payment	4.164719
5	Marital_Status	3.758275
8	Login_device	3.430159
4	CC_Agent_Score	3.046708
7	coupon_used_for_payment	2.758152
10	CC_Contacted_LY_cat	2.694389
2	Gender	2.411772
6	Complain_ly	1.362925
12	CC_connect_category	1.137254

TRAIN/TEST SPLIT

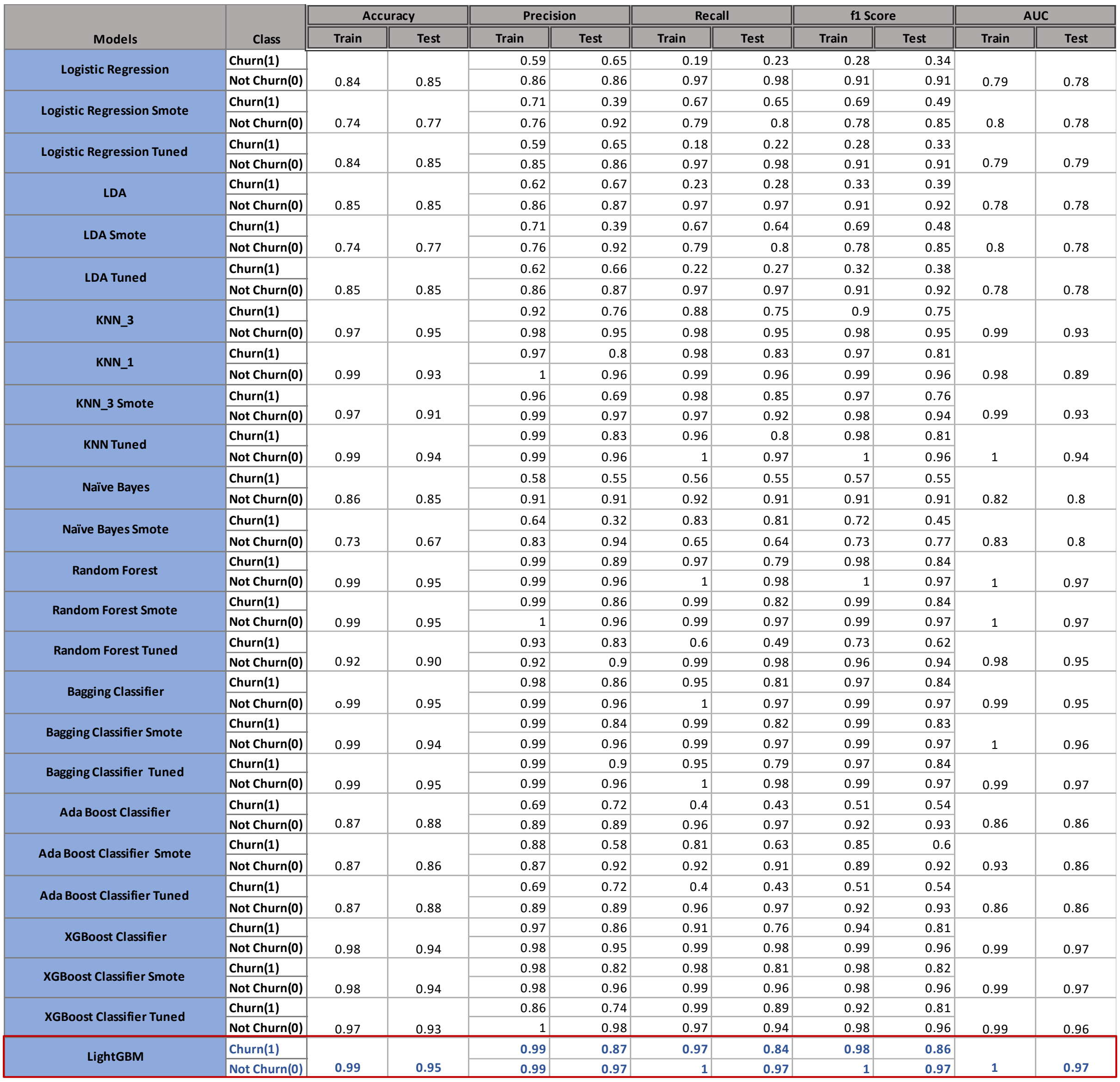
The data is split into training and testing dataset in 70:30 ratio

4. Model building

- Built various models as part of the project.
- The models are built on both balanced and unbalanced datasets and checked the Recall values of different models. Based on **recall values**, optimum model is finalized.
- Models are optimized through Hyperparameters using GridSearchCv to increase the Accuracy and remove Overfitting of the model.
- **Models Tried:**
  - A bunch of models had been put to the test, including **Logistic Regression**, **LDA**, **KNN**, **Naïve Bayes**, **Random Forest**, **Bagging Classifier**, **Ada Boost**, **XG Boost**, and **Light GBM**. Each has its strengths and weaknesses, so we wanted to see how they stacked up.

The following table below shows the various models built along with their performance metrices like Accuracy, Precision, Recall, f1 score and AUC value.

Model Comparison Table

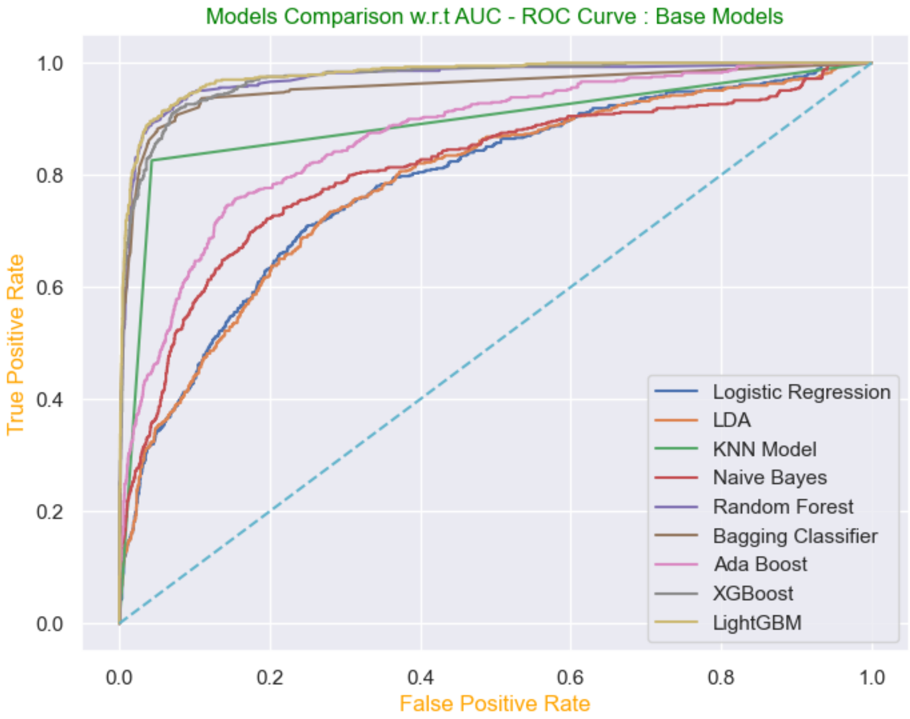


Models	Class	Accuracy		Precision		Recall		f1 Score		AUC	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	Churn(1)	0.84	0.85	0.59	0.65	0.19	0.23	0.28	0.34	0.79	0.78
	Not Churn(0)			0.86	0.86	0.97	0.98	0.91	0.91		
Logistic Regression Smote	Churn(1)	0.74	0.77	0.71	0.39	0.67	0.65	0.69	0.49	0.8	0.78
	Not Churn(0)			0.76	0.92	0.79	0.8	0.78	0.85		
Logistic Regression Tuned	Churn(1)	0.84	0.85	0.59	0.65	0.18	0.22	0.28	0.33	0.79	0.79
	Not Churn(0)			0.85	0.86	0.97	0.98	0.91	0.91		
LDA	Churn(1)	0.85	0.85	0.62	0.67	0.23	0.28	0.33	0.39	0.78	0.78
	Not Churn(0)			0.86	0.87	0.97	0.97	0.91	0.92		
LDA Smote	Churn(1)	0.74	0.77	0.71	0.39	0.67	0.64	0.69	0.48	0.8	0.78
	Not Churn(0)			0.76	0.92	0.79	0.8	0.78	0.85		
LDA Tuned	Churn(1)	0.85	0.85	0.62	0.66	0.22	0.27	0.32	0.38	0.78	0.78
	Not Churn(0)			0.86	0.87	0.97	0.97	0.91	0.92		
KNN_3	Churn(1)	0.97	0.95	0.92	0.76	0.88	0.75	0.9	0.75	0.99	0.93
	Not Churn(0)			0.98	0.95	0.98	0.95	0.98	0.95		
KNN_1	Churn(1)	0.99	0.93	0.97	0.8	0.98	0.83	0.97	0.81	0.98	0.89
	Not Churn(0)			1	0.96	0.99	0.96	0.99	0.96		
KNN_3 Smote	Churn(1)	0.97	0.91	0.96	0.69	0.98	0.85	0.97	0.76	0.99	0.93
	Not Churn(0)			0.99	0.97	0.97	0.92	0.98	0.94		
KNN Tuned	Churn(1)	0.99	0.94	0.99	0.83	0.96	0.8	0.98	0.81	1	0.94
	Not Churn(0)			0.99	0.96	1	0.97	1	0.96		
Naïve Bayes	Churn(1)	0.86	0.85	0.58	0.55	0.56	0.55	0.57	0.55	0.82	0.8
	Not Churn(0)			0.91	0.91	0.92	0.91	0.91	0.91		
Naïve Bayes Smote	Churn(1)	0.73	0.67	0.64	0.32	0.83	0.81	0.72	0.45	0.83	0.8
	Not Churn(0)			0.83	0.94	0.65	0.64	0.73	0.77		
Random Forest	Churn(1)	0.99	0.95	0.99	0.89	0.97	0.79	0.98	0.84	1	0.97
	Not Churn(0)			0.99	0.96	1	0.98	1	0.97		
Random Forest Smote	Churn(1)	0.99	0.95	0.99	0.86	0.99	0.82	0.99	0.84	1	0.97
	Not Churn(0)			1	0.96	0.99	0.97	0.99	0.97		
Random Forest Tuned	Churn(1)	0.92	0.90	0.93	0.83	0.6	0.49	0.73	0.62	0.98	0.95
	Not Churn(0)			0.92	0.9	0.99	0.98	0.96	0.94		
Bagging Classifier	Churn(1)	o.99	0.95	0.98	0.86	0.95	0.81	0.97	0.84	0.99	0.95
	Not Churn(0)			0.99	0.96	1	0.97	0.99	0.97		
Bagging Classifier Smote	Churn(1)	0.99	0.94	0.99	0.84	0.99	0.82	0.99	0.83	1	0.96
	Not Churn(0)			0.99	0.96	0.99	0.97	0.99	0.97		
Bagging Classifier Tuned	Churn(1)	0.99	0.95	0.99	0.9	0.95	0.79	0.97	0.84	0.99	0.97
	Not Churn(0)			0.99	0.96	1	0.98	0.99	0.97		
Ada Boost Classifier	Churn(1)	0.87	0.88	0.69	0.72	0.4	0.43	0.51	0.54	0.86	0.86
	Not Churn(0)			0.89	0.89	0.96	0.97	0.92	0.93		
Ada Boost Classifier Smote	Churn(1)	0.87	0.86	0.88	0.58	0.81	0.63	0.85	0.6	0.93	0.86
	Not Churn(0)			0.87	0.92	0.92	0.91	0.89	0.92		
Ada Boost Classifier Tuned	Churn(1)	0.87	0.88	0.69	0.72	0.4	0.43	0.51	0.54	0.86	0.86
	Not Churn(0)			0.89	0.89	0.96	0.97	0.92	0.93		
XGBoost Classifier	Churn(1)	0.98	0.94	0.97	0.86	0.91	0.76	0.94	0.81	0.99	0.97
	Not Churn(0)			0.98	0.95	0.99	0.98	0.99	0.96		
XGBoost Classifier Smote	Churn(1)	0.98	0.94	0.98	0.82	0.98	0.81	0.98	0.82	0.99	0.97
	Not Churn(0)			0.98	0.96	0.99	0.96	0.98	0.96		
XGBoost Classifier Tuned	Churn(1)	0.97	0.93	0.86	0.74	0.99	0.89	0.92	0.81	0.99	0.96
	Not Churn(0)			1	0.98	0.97	0.94	0.98	0.96		
LightGBM	Churn(1)	0.99	0.95	0.99	0.87	0.97	0.84	0.98	0.86	1	0.97
	Not Churn(0)			0.99	0.97	1	0.97	1	0.97		

why was a particular model(s) chosen?

1. **Models Comparison**:

- Different models such as **Logistic Regression**, **LDA**, **KNN**, **Naive Bayes**, **Random Forest**, **Bagging Classifier**, **Ada Boost**, **XGBoost**, and **LightGBM** have been tested and tuned.
- Some models have been applied with **SMOTE** (Synthetic Minority Oversampling Technique) to handle imbalanced data.
- Comparing the AUC Score we can say that Random Forest and LightGBM Models are performing Better having a High AUC score of 97% 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.



- Among all Built models, LightGBM model is giving Highest recall on Test data and Highest AUC score of 97%.

2. **Overfitting Issues**:

- Some models, like **KNN_3**, show high training performance but much lower test performance for certain metrics (e.g., recall for Churn(1) in the test set is much lower than in the training set). This suggests possible **overfitting**, where the model is too tailored to the training data and does not generalize well.
- Similarly, **Random Forest** and **Bagging Classifier** show excellent performance on training data but a noticeable drop in test performance, indicating overfitting.

3. **Best Performing Models**:

- **LightGBM** appears to be the best model across most metrics, especially on the test data:
  - **Accuracy**: 0.95 on the test set.
  - **Precision** for both classes is high (0.95 for Not Churn(0), 0.87 for Churn(1)).
  - **Recall**: Excellent recall for both classes (0.99 for Not Churn, 0.87 for Churn), which indicates it is identifying both churners and non-churners well.
  - **F1 Score**: High values for both classes.
  - **AUC**: 0.97, which indicates strong discriminatory power between churners and non-churners.
- **XGBoost Tuned** also performs well, especially in recall and F1 score for the Churn class.

4. **Class Imbalance Handling**:

- Models trained with **SMOTE** (like Logistic Regression Smote, KNN_3 Smote, Random Forest Smote) generally show better recall for Churn(1) than their non-SMOTE counterparts. This is because SMOTE helps in dealing with the imbalance between churners and non-churners.
- However, using SMOTE can sometimes lower precision, as seen in **Logistic Regression Smote** and **LDA Smote**.

5. **Evaluation Metrics**:

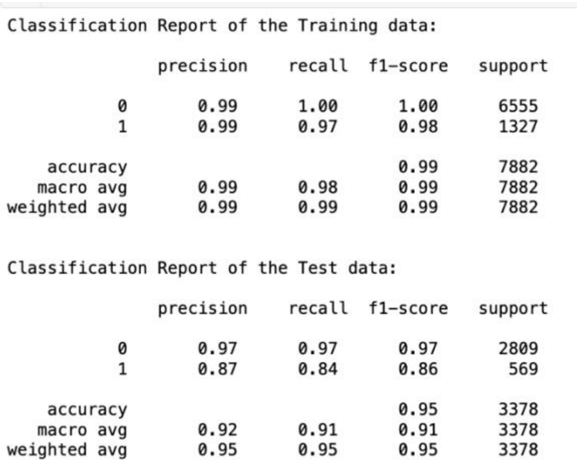
- **Accuracy** alone may not be a sufficient metric, especially in imbalanced datasets. Hence, more weight should be given to **Recall** and **F1 Score**, especially for Churn(1), to ensure the model is correctly identifying churners.
- **Recall for Churn(1)** is particularly important because predicting churners accurately is crucial for taking action.

6. **General Observations**:

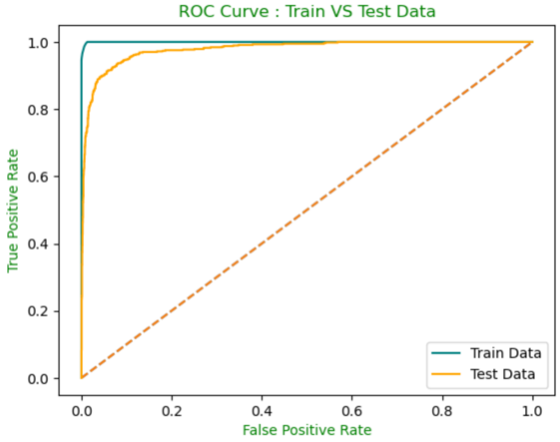
- **Random Forest Tuned** and **Bagging Classifier Tuned** also perform well on the test set but with slightly lower AUC and precision compared to **LightGBM** and **XGBoost**.
- **Logistic Regression** models perform decently but are outperformed by more complex models like **XGBoost** and **LightGBM**, especially in recall and F1 score for Churn(1).

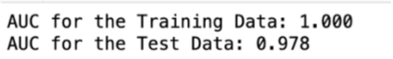
7. **Model Chosen**:

- **LightGBM** is the best performing model in this comparison based on the combination of accuracy, precision, recall, F1 score, and AUC, particularly for the Churn(1) class. It balances between minimizing false negatives (recall) and false positives (precision) effectively.

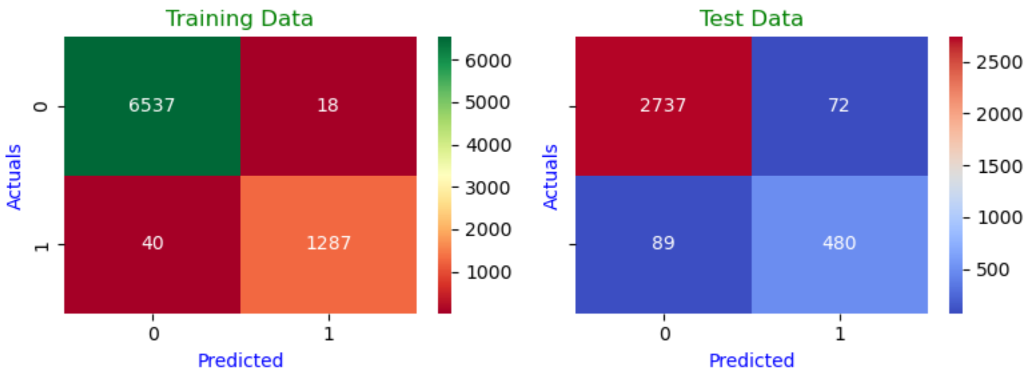


```
Classification Report of the Training data:

              precision    recall  f1-score   support

           0       0.99      1.00      1.00      6555
           1       0.99      0.97      0.98      1327

    accuracy                           0.99      7882
   macro avg       0.99      0.98      0.99      7882
weighted avg       0.99      0.99      0.99      7882

Classification Report of the Test data:

              precision    recall  f1-score   support

           0       0.97      0.97      0.97      2809
           1       0.87      0.84      0.86       569

    accuracy                           0.95      3378
   macro avg       0.92      0.91      0.91      3378
weighted avg       0.95      0.95      0.95      3378
```





```
AUC for the Training Data: 1.000
AUC for the Test Data: 0.978
```



Efforts taken to improve model performance:

1. Model Tuning:

- Hyperparameter Tuning: Hyperparameters were optimized using techniques like GridSearchCV to improve model performance.
- Tuning hyperparameters helps improve the model's ability to generalize, enhances its predictive performance, and reduces overfitting or underfitting.

2. Use of SMOTE (Synthetic Minority Oversampling Technique):

- Handling Imbalanced Data: SMOTE was applied in some models (e.g., Logistic Regression SMOTE, LDA SMOTE, KNN_3 SMOTE, Random Forest SMOTE, XGBoost Classifier SMOTE). This technique oversamples the minority class (Churn(1)) to balance the dataset.
- When the dataset is imbalanced (more non-churners than churners), models may tend to favour the majority class, leading to poor recall for the minority class. By using SMOTE, the models are exposed to more churner data, which improves their ability to predict churners (as seen in the improved recall values for Churn(1)).

3. Ensemble Techniques:

- Bagging and Boosting: Techniques like Bagging Classifier, Random Forest, Ada Boost, and XGBoost were used. These are ensemble methods that improve model performance by combining predictions from multiple models:
  - Bagging (Bootstrap Aggregating): Reduces variance by averaging predictions from multiple base learners, leading to more robust predictions.
  - Boosting: Sequentially improves the model by correcting errors from previous iterations, often leading to better performance for more complex patterns.
  - Ensemble methods generally outperform individual models by reducing bias and variance, leading to higher accuracy, recall, and F1 scores, as seen in the tuned versions of these models.

4. Different Classifiers Tested:

- A wide range of models have been tested, from simpler models like Logistic Regression and Naive Bayes to more complex ones like Random Forest, XGBoost, and LightGBM.

- Trying different algorithms allows finding the best performing model for the given data. Complex models like XGBoost and LightGBM tend to capture non-linear relationships in the data better than simpler models like Logistic Regression.

5. Cross-Validation and Model Selection:

- Models were evaluated using both training and test datasets, and multiple performance metrics were reported (Accuracy, Precision, Recall, F1 Score, and AUC).
- Using various metrics helps to choose the model that balances precision and recall, especially for the minority class (Churn(1)). Cross-validation ensures the model generalizes well on unseen data, avoiding overfitting.

6. Reducing Overfitting:

- Tuned versions of complex models (e.g., Random Forest, XGBoost, Bagging Classifier) and techniques like SMOTE helped reduce overfitting:
  - Overfitting: Some models performed very well on training data but dropped in performance on test data. This was addressed through tuning and regularization (e.g., using hyperparameters like the number of trees, depth, and learning rates in Random Forest and XGBoost).
  - Reducing overfitting allows the model to perform well on unseen test data, ensuring that the model isn't simply memorizing the training data.

7. Focusing on Recall and Precision:

- Given that predicting Churn(1) (minority class) is the main objective, significant efforts were put into improving recall and F1 score for this class, even if it meant a trade-off in precision.
  - Purpose: In a churn prediction problem, recall (correctly identifying churners) is crucial since missing a churner (false negative) can lead to lost business. The high recall values for some models (especially LightGBM and XGBoost) demonstrate efforts to capture most churners.

Summary of Efforts:

- Hyperparameter Tuning: Improved performance by optimizing model parameters.
- SMOTE: Addressed class imbalance and improved recall for churn prediction.
- Ensemble Methods: Boosted performance by combining predictions of multiple models.
- Cross-Validation & Metric Reporting: Ensured the model generalizes well and is evaluated comprehensively across different metrics.
- Reducing Overfitting: Improved generalization by tuning complex models and applying regularization.

5. Model validation

The models in the churn prediction analysis were validated using multiple metrics beyond just accuracy.

1. Accuracy:

- It is the ratio of correctly predicted observations (both churners and non-churners) to the total observations. It gives an overall effectiveness of the model.
- In imbalanced datasets, like typical churn data where the number of non-churners might significantly outnumber churners, accuracy might not be the best measure as it could be misleadingly high due to a high number of true negatives (correct predictions of the majority class).

2. Precision:

- Precision is the ratio of correctly predicted positive observations to the total predicted positives. It measures the accuracy of positive predictions.
- This metric is crucial when the cost of a false positive is high. In the context of churn prediction, high precision means that when a model predicts a customer will churn, it is very likely correct, minimizing wasted retention efforts on customers who would have stayed anyway.

3. Recall (Sensitivity):

- Recall is the ratio of correctly predicted positive observations to all actual positives. It measures the ability of the model to find all relevant cases (all actual churners).
- **High recall** is vital in scenarios where missing a positive prediction is costly. For churn prediction, a high recall implies the model is effective in identifying most churners, which is critical for targeted customer retention strategies.

4. F1 Score:

- The F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.
- It is a better measure than examining Precision and Recall individually, especially in imbalanced datasets. It helps balance the trade-off between Precision and Recall, providing a more holistic view of model performance.

5. AUC - ROC Curve:

- The area under the Receiver Operating Characteristic (ROC) curve (AUC - ROC) is a performance measurement for classification problems. It tells how much a model is capable of distinguishing between classes.
- Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. For churn prediction, a high AUC indicates that the model has a good measure of separability between churners and non-churners.

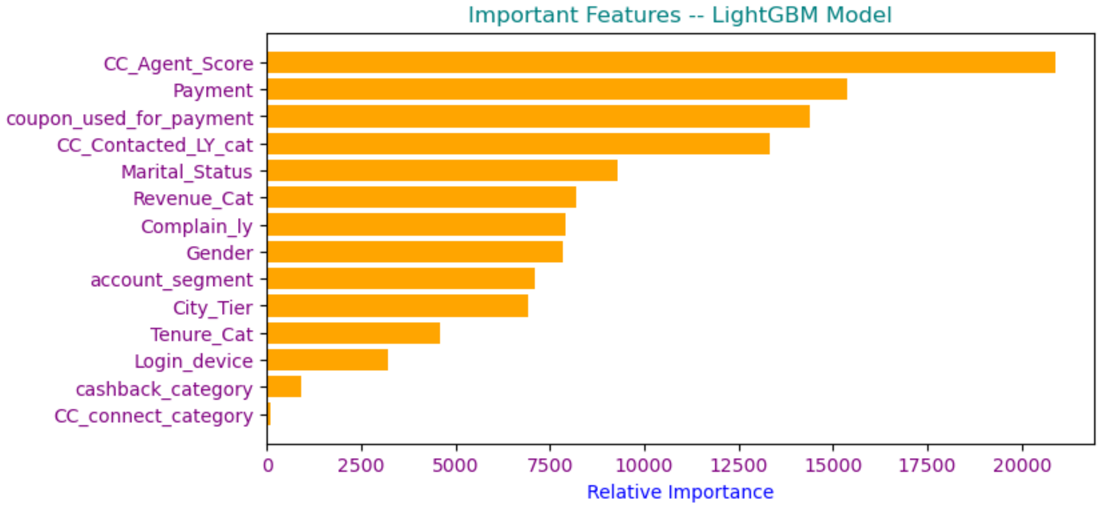
Model Validation Techniques:

- **Cross-Validation**: Often used to ensure that the model’s performance is consistent across different subsets of the dataset and that it generalizes well to unseen data.

The use of multiple metrics such as Precision, Recall, F1 Score, and AUC alongside Accuracy provides a comprehensive view of the model's performance. This multifaceted validation approach is crucial in churn prediction to ensure that the model is both effective in identifying true churners (high recall) and efficient in confirming its predictions (high precision), balancing the overall prediction capabilities (F1 score), and distinguishing between churners and non-churners effectively (high AUC). This robust validation ensures that the model's predictions are reliable and actionable.

- **Accuracy** alone may not be a sufficient metric, especially in imbalanced datasets. Hence, more weight is given to **Recall** and **F1 Score**, especially for Churn(1), to ensure the model is correctly identifying churners.
- **Recall for Churn(1)** is particularly important because predicting churners accurately is crucial for taking action.

FEATURE IMPORTANCE



- The **CC_Agent_Score** and **Payment** have significantly higher importance than other features, indicating the model heavily relies on customer satisfaction with agents and payment-related data.
- Demographic and behavioural features like **Marital_Status**, **Revenue_Cat**, **Gender** also play an important role, indicating that customer profile and interactions significantly affect outcomes.

6. Final interpretation / recommendation

Insights

- One of the biggest concerns we've noticed from the data is customer satisfaction or lack thereof. A huge **78% of customers** have rated the overall service as **3 or lower**. And when it comes to our customer care agents, things aren't much better—**61% of customers** gave them a score of **3 or lower**.
- When it comes to churn ,it turns out that people who use Cash on Delivery (COD) are more likely to churn compared to those using other payment methods.

- We could see a comparatively low churning rate in Debit/credit card . This could be due to auto debit option from the user bank account.
- It looks like churned customers tend to have lower cashback compared to those who stick around. This suggests that cashback might be a factor in keeping customers engaged.
- Interestingly, both churned and active customers are using about the same number of coupons for payments. So, it doesn't seem like coupons are making a big difference in whether customers stay or leave.
- It seems like the current retention programs might not be targeting those at higher risk of churning effectively.
- About 31% of customers who registered a complaint ended up churning, compared to only 11% of those who didn't lodge a complaint. Clearly, complaints are a strong indicator of potential
- **customers who leave are usually the ones who haven't been around for long**—specifically, those in their first 0 to 2 months. On the flip side, our long-term customers tend to stick around and seem pretty happy with us.
- We've seen that **high-revenue customers** are actually churning at a higher rate than those who spend less. Specifically, customers who bring in **$7 or more per month** are leaving us more frequently compared to those with lower revenue.
- We need to **identify and address their pain points**. Maybe they're facing issues or have unmet needs that we haven't tackled yet.

Recommendation

- The company should really invest in **training customer care agents**. They're the frontline of communication with our customers. If they're not equipped to handle issues or provide great service, it's going to reflect badly on the entire brand. So, extensive training could help improve not just their performance but also the way customers perceive the service.
- **Conduct a full audit of the service**. We need to know what's driving this dissatisfaction—whether it's response times, technical issues, or service quality. Once we have that information, we can figure out the exact pain points.
- Payment preferences can vary across different regions **.There should be a Smooth COD Experience.** Since COD is quite popular, it's important to make sure the process is super smooth and hassle-free for customers. A little improvement here could really help reduce churn.
- **Promote E-wallets and UPI.** These payment options are not just convenient but also secure. Giving them a bit more spotlight could be a win-win, making things easier for customers and potentially reducing churn.
- **Combine Cashback with Coupons:** For even better results, we could mix cashback offers with coupon usage. This combo could drive more engagement and make customers feel they're getting more value.
- Set up automated messages or reminders that are tailored to individual customers. This could help keep them engaged and make them feel valued.
- Boosting the speed and quality of customer interactions could make a big difference. Faster and more helpful responses might help reduce churn.

- Offering more self-service options like FAQ sections, chatbots, or mobile apps can empower customers to solve their own issues quickly, which could lead to higher satisfaction and lower churn.
- **Conduct Product Education Campaigns** to help customer understand the service better right from the start.
- **Automate some check-ins**. Simple messages or emails can help stay connected and address any issues before they become reasons for customers to leave.
- **Conduct exit interviews or surveys**. This will help us get direct insights into why they're leaving and what we can improve.

CONCLUSION

- To conclude , reducing customer churn has a direct impact on profitability for DTH service providers. By using machine learning models like LightGBM, we can effectively predict and mitigate churn. Moving forward, focusing on customer engagement during the critical first year could greatly improve retention and, ultimately, revenue."