Capstone Project Customer Churn – Final Notes

PGP in Data Science and Business Analytics

Great Learning

NOVEMBER 30,2022

Authored by:

Dheepika Mai Ganesh Babu

PGPDSBA ONLINE DEC_A 2021

Contents

Topic **Page No.**

List of Tables

List of Figures

Problem:

Customer Churn Data Analysis

Problem Statement:

An E Commerce customer churn 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 customer churn 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 customer churn, account churn is a major thing because 1 account can have multiple customers. hence by losing one account the customer churn might be losing more than one customer.

You have been assigned to develop a churn prediction model for this customer churn 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 customer churn; they are not going to approve your recommendation. Hence be very careful while providing campaign recommendation.

Sample Dataset:

	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Mar
0	20000	1	4.0	3.0	6.0	Debit Card	Female	3.0	3.0	Super	2.0	
1	20001	1	0.0	1.0	8.0	UPI	Male	3.0	4.0	Regular Plus	3.0	
2	20002	1	0.0	1.0	30.0	Debit Card	Male	2.0	4.0	Regular Plus	3.0	
3	20003	1	0.0	3.0	15.0	Debit Card	Male	2.0	4.0	Super	5.0	
4	20004	1	0.0	1.0	12.0	Credit Card	Male	2.0	3.0	Regular Plus	5.0	

Table 1 – Sample of the customer churn dataset

Data Dictionary of Customer churn Dataset:

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_L12m	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_l12m	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_l12m	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

Table 2 - Data Dictionary of Customer churn Dataset

Problem 1:

Problem 1. Introduction.

Brief introduction about the problem statement and the need of solving it.

Defining problem statement:

An E commerce company was performing well in the market, but due to heavy competition in market it started facing customer churn problem. Now the company wants to analyze the dataset to understand the customer churn behavior and find causes as well as indicators of customer churn and fix it. Hence, the company wants to study the customer behavior and pattern to increase the customer traffic on their website.

Need of the study/project:

Customer retention is one of the primary growth pillars for products with a subscription-based business model. Customer relationship management is fundamentally an investment decision for the long run. It is the business, as well as the customer base, that must be handled to get the most out of the client portfolio. Customer churn is an important metric for business professionals who use aggressive business policies in their business.

In 2022, the Indian e-commerce market is predicted to increase by 21.5%, reaching US$ 74.8 billion. India's e-commerce market is expected to reach US$ 350 billion by 2030. These numbers can significantly change due to introduction of 5G network, smartphone and internet users increasing every day. Flipkart, Amazon, Myntra, Jio Mart, Nykaa, Ajio, Snapdeal, Meesho, Snapdeal, FirstCry, PharmEasy and many other e-commerce websites are widely preferred by online purchasers in India.

Understanding business/social opportunity:

As a part of social opportunity, e-commerce business is helping lot of small-scale industries to expand their customer market geographically. We see plenty of start-ups emerging in the past few years and e-commerce has helped them thrive and generate revenue. This in turn has opened-up the job market for many and in turn contributed to the country's economy. It has become quite easier for people to start their dream of own business with lesser capital.

Problem 2 EDA and Business Implication.

Uni-variate / Bi-variate / Multi-variate analysis to understand relationship b/w variables. How your analysis is impacting the business?

Data understanding in terms of time, frequency and methodology:

No information about the data understanding in terms of time, frequency and methodology is available in the dataset and hence so insight can be drawn in this regard.

Both visual and non-visual understanding of the data:

```
There are total 11260 rows and 18 columns in the dataset
```

Table 3 – Shape of the customer churn dataset

```
<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  float64
 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  float64
 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            10469 non-null  float64
 12  Complain_ly              10903 non-null  float64
 13  rev_growth_yoy           11260 non-null  int64
 14  coupon_used_for_payment  11260 non-null  int64
 15  Day_Since_CC_connect     10903 non-null  float64
 16  cashback                 10789 non-null  float64
 17  Login_device             11039 non-null  object
dtypes: float64(10), int64(3), object(5)
memory usage: 1.5+ MB
```

Table 4 - Structure of the customer churn dataset

	count	mean	std	min	25%	50%	75%	max
Churn	11260.0	0.168384	0.374223	0.0	0.0	0.00	0.00	1.0
Tenure	11158.0	10.910468	12.861364	0.0	2.0	8.00	16.00	99.0
City_Tier	11148.0	1.653929	0.915015	1.0	1.0	1.00	3.00	3.0
CC_Contacted_LY	11158.0	17.867091	8.853269	4.0	11.0	16.00	23.00	132.0
Service_Score	11162.0	2.902526	0.725584	0.0	2.0	3.00	3.00	5.0
Account_user_count	11148.0	3.582885	1.187175	0.0	3.0	4.00	4.00	6.0
CC_Agent_Score	11144.0	3.066493	1.379772	1.0	2.0	3.00	4.00	5.0
rev_per_month	10469.0	6.362594	11.909686	1.0	3.0	5.00	7.00	140.0
Complain_ly	10903.0	0.285334	0.451594	0.0	0.0	0.00	1.00	1.0
rev_growth_yoy	11260.0	16.189076	3.766505	0.0	13.0	15.00	19.00	28.0
coupon_used_for_payment	11260.0	1.790142	1.969505	0.0	1.0	1.00	2.00	16.0
Day_Since_CC_connect	10903.0	4.632762	3.697733	0.0	2.0	3.00	8.00	47.0
cashback	10789.0	196.199993	178.663928	0.0	147.2	165.24	199.98	1997.0

Table 5(a) - Summary of numerical variables in customer churn dataset

	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	NaN	NaN	NaN	10.910468	12.861364	0.0	2.0	8.0	16.0	99.0
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	2	Male	6704	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	NaN	NaN	NaN	3.582885	1.187175	0.0	3.0	4.0	4.0	6.0
account_segment	11163	5	Regular Plus	4124	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	10469.0	NaN	NaN	NaN	6.362594	11.909686	1.0	3.0	5.0	7.0	140.0
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	NaN	NaN	NaN	16.189076	3.766505	0.0	13.0	15.0	19.0	28.0
coupon_used_for_payment	11260.0	NaN	NaN	NaN	1.790142	1.969505	0.0	1.0	1.0	2.0	16.0
Day_Since_CC_connect	10903.0	NaN	NaN	NaN	4.632762	3.697733	0.0	2.0	3.0	8.0	47.0
cashback	10789.0	NaN	NaN	NaN	196.199993	178.663928	0.0	147.2	165.24	199.98	1997.0
Login_device	11039	3	Mobile	7482	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 5(b) - Summary of all variables in customer churn dataset

```
Churn                      int64
Tenure                   float64
City_Tier                float64
CC_Contacted_LY          float64
Payment                   object
Gender                    object
Service_Score            float64
Account_user_count       float64
account_segment           object
CC_Agent_Score           float64
Marital_Status            object
rev_per_month            float64
Complain_ly              float64
rev_growth_yoy             int64
coupon_used_for_payment    int64
Day_Since_CC_connect     float64
cashback                 float64
Login_device              object
dtype: object
```

Table 6 – Data type of the variables in customer churn dataset

```
Churn                      1.772606
Tenure                     3.891973
City_Tier                  0.737107
CC_Contacted_LY            1.422977
Service_Score              0.003891
Account_user_count        -0.861867
CC_Agent_Score            -0.142149
rev_per_month              9.093909
Complain_ly                0.950876
rev_growth_yoy             0.729281
coupon_used_for_payment    2.575218
Day_Since_CC_connect       1.272969
cashback                   8.769003
dtype: float64
```

```
Churn                      1.142336
Tenure                    23.392592
City_Tier                 -1.398498
CC_Contacted_LY            8.226080
Service_Score             -0.668069
Account_user_count         1.336320
CC_Agent_Score            -1.124834
rev_per_month             86.963130
Complain_ly               -1.096036
rev_growth_yoy            -0.151865
coupon_used_for_payment    9.101885
Day_Since_CC_connect       5.328472
cashback                  81.099684
dtype: float64
```

Table 7 – Skewness(left) and Kurtosis(right) of the customer churn dataset

	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status
0	1	4.0	3.0	6.0	Debit Card	Female	3.0	3.0	Super	2.0	Single
1	1	0.0	1.0	8.0	UPI	Male	3.0	4.0	Regular Plus	3.0	Single
2	1	0.0	1.0	30.0	Debit Card	Male	2.0	4.0	Regular Plus	3.0	Single
3	1	0.0	3.0	15.0	Debit Card	Male	2.0	4.0	Super	5.0	Single
4	1	0.0	1.0	12.0	Credit Card	Male	2.0	3.0	Regular Plus	5.0	Single

Table 8 – Revised sample of the customer churn dataset – after dropping 'AccountID'

```
Churn                      2
Tenure                    37
City_Tier                  3
CC_Contacted_LY           44
Service_Score              6
Account_user_count         7
CC_Agent_Score             5
Payment                    5
Gender                     2
rev_per_month             58
coupon_used_for_payment   17
Day_Since_CC_connect      23
cashback                5692
Marital_Status             3
Complain_ly                2
rev_growth_yoy            20
dtype: int64
```

Table 9 – Unique values count in customer churn dataset

Understanding of attributes:

- Table 6 shows the variable info of all columns in the dataset given.
- Two types of datatypes are available that is float and object types.
- We have given total 18 columns as it is of no use for model building.
- We have total of 11260 records for analytics.
- Table 1 shows the sample records of the customer churn dataset.
- Originally there are 19 columns and 11260 rows in the customer churn dataset. However, after dropping 'AccountID' we have 18 columns.
- Table 2 shows the data dictionary with the column names and their description in the dataset.
- There are 5 float columns, 1 integer column, and 12 object columns.
- The target churn column has been flagged '1' for the at-risk customers. The distribution for churn count is displayed below.
- We have fixed the data discrepancies in excel sheet before reading it. All the special characters found in integer columns have been replaced with blank considering those as typo-errors at the time of data entry.

- We drop column ‘AccountID’ before we perform the analysis on the dataset as it is irrelevant for model building.
- We see that not even a single column is normally distributed, and data is skewed.
- Table 7 shows that the minimum value range and maximum value range is close at some columns and at higher level at some places.
- On first look we can say that count of all columns is not same.
- There are unique values present in the data.
- Integer type variables are "Tenure", "Account_user_count", "rev_per_month", "rev_growth_yoy", "coupon_used_for_payment", "Day_Since_CC_connect" and "cashback".
- In last 12 months, maximum and minimum number of times all the customers of the accounts have contacted customer care is 132 and 84 times respectively. On an average 16 customers has contacted customer care.
- There are missing values or anomalies present in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10995 entries, 0 to 11259
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Churn                    10995 non-null  float64
 1   Tenure                   10995 non-null  float64
 2   City_Tier                10995 non-null  float64
 3   CC_Contacted_LY          10995 non-null  float64
 4   Payment                  10995 non-null  object
 5   Gender                   10995 non-null  object
 6   Service_Score            10995 non-null  float64
 7   Account_user_count       10995 non-null  float64
 8   account_segment          10995 non-null  object
 9   CC_Agent_Score           10995 non-null  float64
 10  Marital_Status           10995 non-null  object
 11  rev_per_month            10995 non-null  float64
 12  Complain_ly              10995 non-null  float64
 13  rev_growth_yoy           10995 non-null  float64
 14  coupon_used_for_payment  10995 non-null  float64
 15  Day_Since_CC_connect     10995 non-null  float64
 16  cashback                 10995 non-null  float64
 17  Login_device             10995 non-null  object
dtypes: float64(13), object(5)
memory usage: 1.6+ MB
```

Table 10 – Structure of the revised customer churn dataset

Univariate analysis:

	Churn	City_Tier	CC_Contacted_LY
count	11260.000000	11148.000000	11158.000000
mean	0.168384	1.653929	17.867091
std	0.374223	0.915015	8.853269
min	0.000000	1.000000	4.000000
25%	0.000000	1.000000	11.000000
50%	0.000000	1.000000	16.000000
75%	0.000000	3.000000	23.000000
max	1.000000	3.000000	132.000000

Table 11 – Description of the variables Churn, City_Tier and CC_Contacted_LY in customer churn dataset

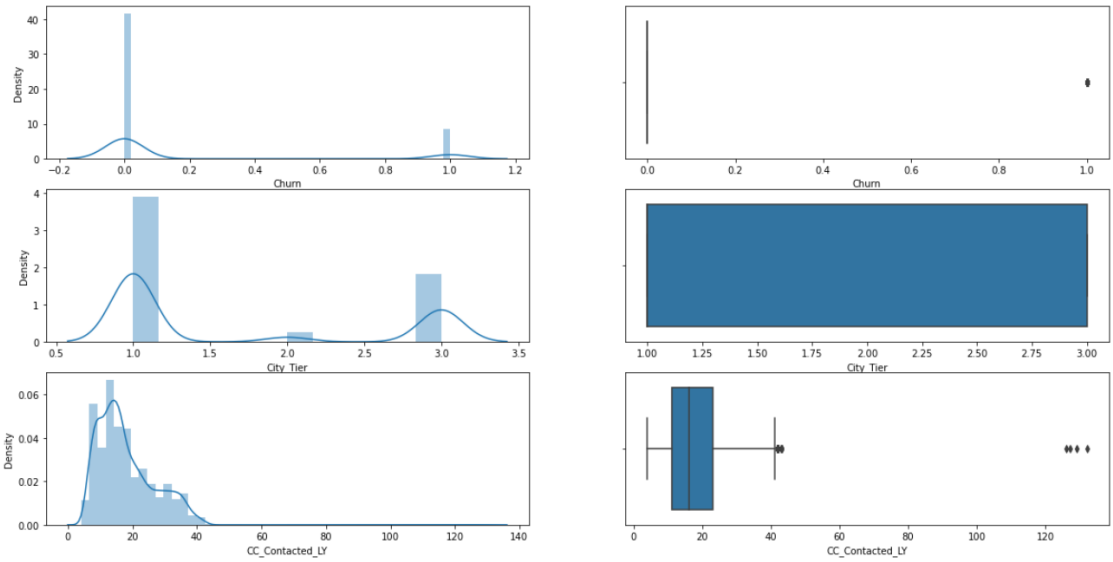

Fig 1 - Distplot and boxplot to visualize the univariate analysis of variables Churn, City_Tier and CC_Contacted_LY in customer churn dataset

	Service_Score	CC_Agent_Score	Complain_ly
count	11162.000000	11144.000000	10903.000000
mean	2.902526	3.066493	0.285334
std	0.725584	1.379772	0.451594
min	0.000000	1.000000	0.000000
25%	2.000000	2.000000	0.000000
50%	3.000000	3.000000	0.000000
75%	3.000000	4.000000	1.000000
max	5.000000	5.000000	1.000000

Table 12 – Description of the variables Service_Score, CC_Agent_Score and Complaint_ly in customer churn dataset

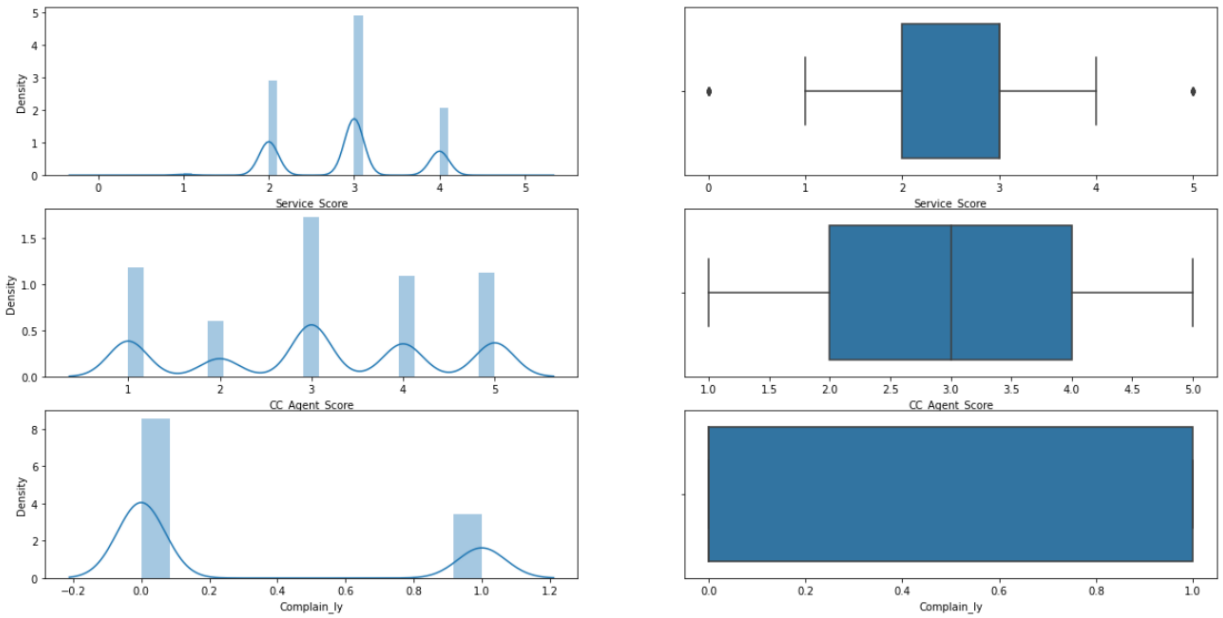

Fig 2 - Distplot and boxplot to visualize the univariate analysis of variables Service_Score, CC_Agent_Score and Complaint_ly in customer churn dataset

	Account_user_count	rev_per_month	rev_growth_yoy
count	11148.000000	10469.000000	11260.000000
mean	3.582885	6.362594	16.189076
std	1.187175	11.909686	3.766505
min	0.000000	1.000000	0.000000
25%	3.000000	3.000000	13.000000
50%	4.000000	5.000000	15.000000
75%	4.000000	7.000000	19.000000
max	6.000000	140.000000	28.000000

Table 13 – Description of the variables Account_user_count, rev_per_month and rev_growth_yoy in customer churn dataset

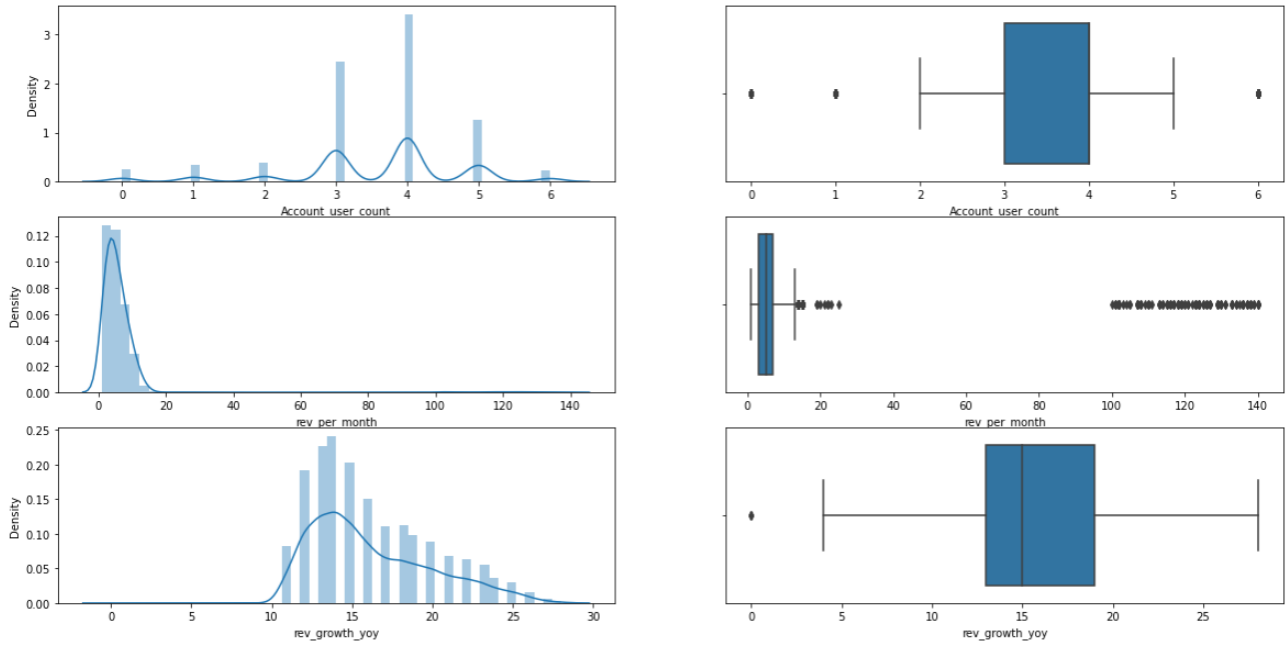

Fig 3 - Distplot and boxplot to visualize the univariate analysis of variables Account_user_count, rev_per_month and rev_growth_yoy in customer churn dataset

	coupon_used_for_payment	Day_Since_CC_connect	cashback
count	11260.000000	10903.000000	10789.000000
mean	1.790142	4.632762	196.199993
std	1.969505	3.697733	178.663928
min	0.000000	0.000000	0.000000
25%	1.000000	2.000000	147.200000
50%	1.000000	3.000000	165.240000
75%	2.000000	8.000000	199.980000
max	16.000000	47.000000	1997.000000

Table 14 – Description of the variables coupon_used_for_payment, Day_Since_CC_connect and cashback in customer churn dataset

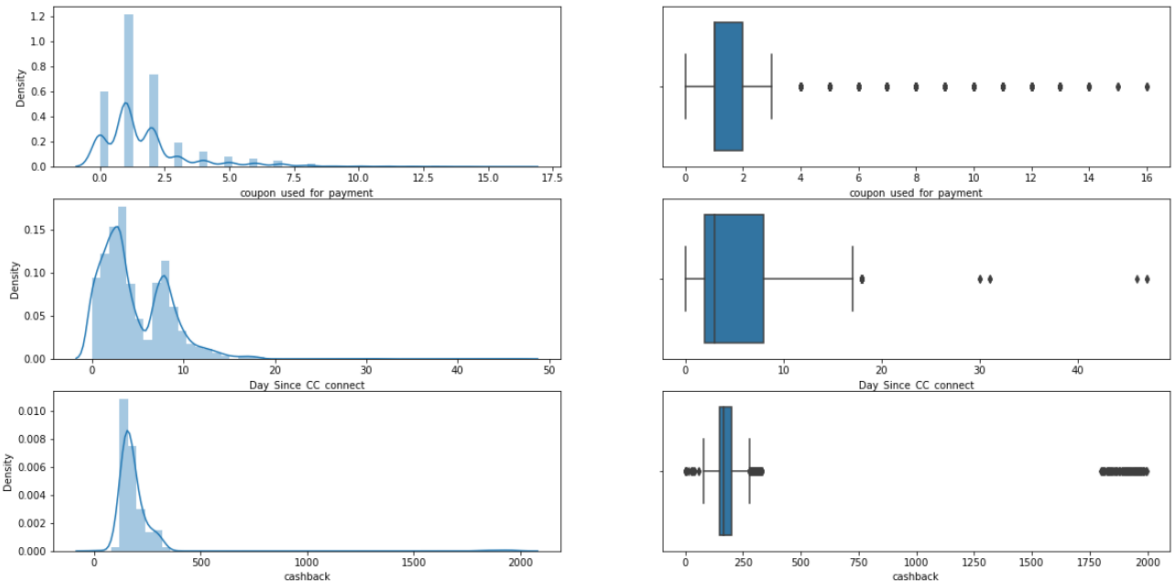

Fig 4 - Distplot and boxplot to visualize the univariate analysis of variables coupon_used_for_payment, Day_Since_CC_connect and cashback in customer churn dataset

Inferences:

- From Fig 1, Churn column is almost left skewed and city tier column does not show any skewness.
- The data of contacted column is left skewed mostly and for service score its almost normally distributed.
- Fig 2 shows that CC_Agent_Score column is almost normally distributed complain doesn't show skewness.
- From Fig 3 & Fig 4 we see that the distribution is skewed to right tailed for all the variable except for variable rev_growth_yoy.
- Since skewness is more than +1 for Variables Day_Since_CC_connect, CC_Contacted_LY, Churn, coupon_used_for_payment, Tenure, cashback and rev_per_month these variables are highly skewed.
- Also, since the skewness is ranging between -0.5 and 0.5 for variables we can say that data is moderately skewed.
- Negative skew refers to a longer or fatter tail on the left side of the distribution, while positive skew refers to a longer or fatter tail on the right. The mean of positively skewed data will be greater than the median.
- Variable rev_growth_yoy is symmetrical since skewness is between 0.5 to +1.

Bivariate analysis:

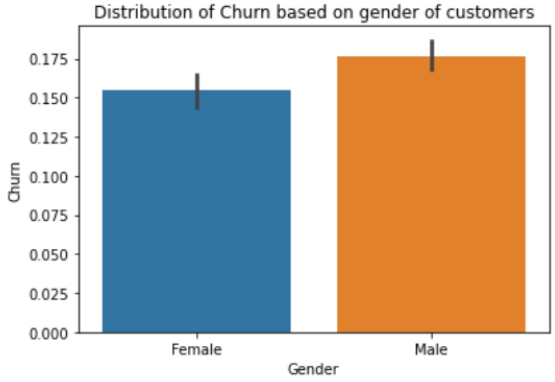

Fig 5 – Barplot of customer churn based on Gender

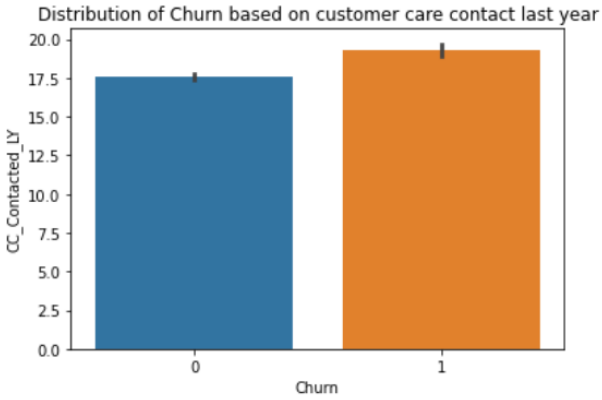

Fig 6 – Barplot of customer churn based on CC_Contacted_LY

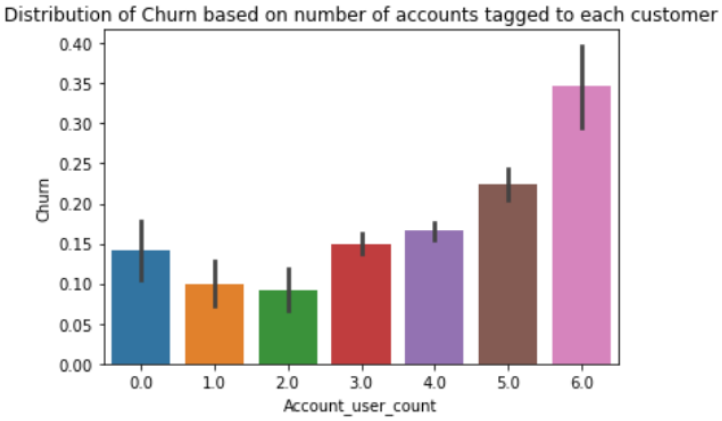

Fig 7 – Barplot of customer churn based on Account_user_count

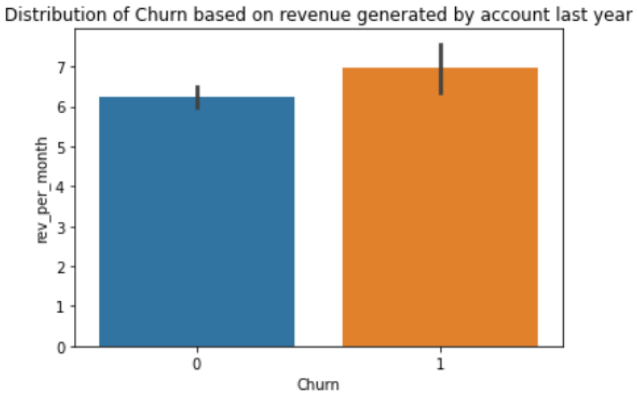

Fig 8(a) – Barplot of customer churn based on rev_per_month

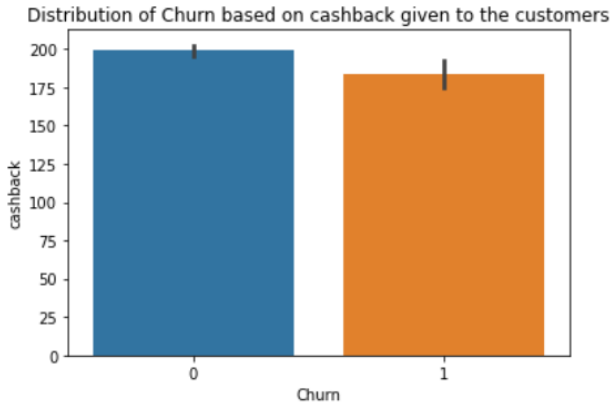

Fig 8(b) – Barplot of customer churn based on cashback

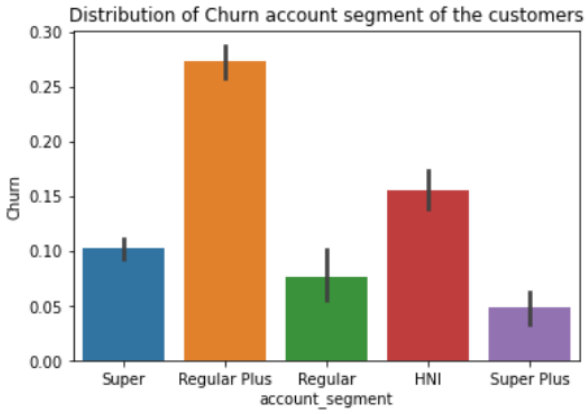

Fig 9 – Barplot of customer churn based on account segment of the customers

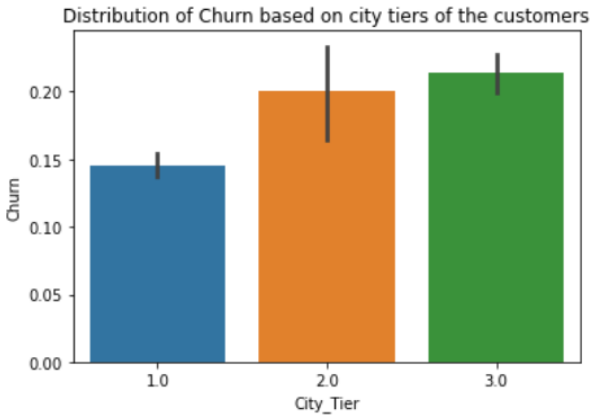

Fig 10 – Barplot of customer churn based on city tier of the customers

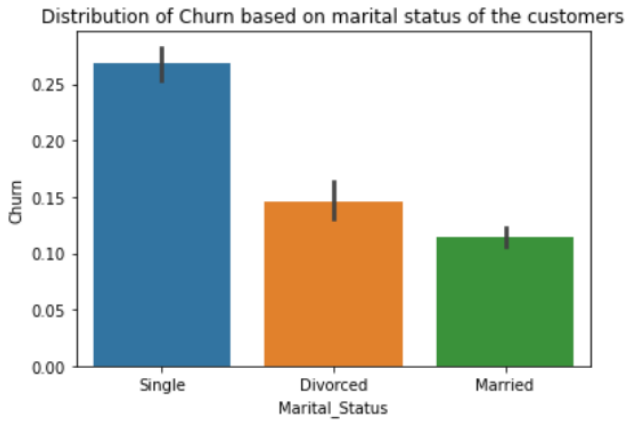

Fig 11 – Barplot of customer churn based on marital status of the customers

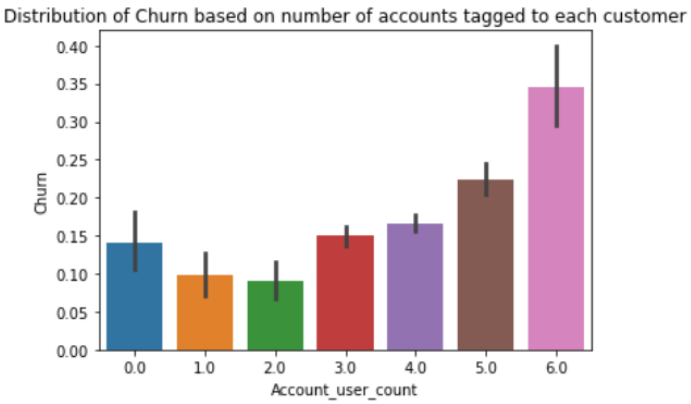

Fig 12 – Barplot of customer churn based on number of users per account

Inferences:

We get below insights from bar plots from Fig 5 till Fig 12.

- Male customers are more prone to churn when compared to female customers.
- We see most of the peoples like 50% are using debit card and credit card to do payment.
- We can see account_segment univariate as below. It has seven different types of segments.
- Customers who are single show more signs of churning, divorced customers show comparatively lesser churning rate and married customers show the least.
- Most of the customers who are in Regular Plus and super have higher churning rate.
- Customers living in City_Tier 3 are in the verge of churning.
- Customers who do not have coupon for recharge are churning more than the customers with coupon.
- Customers who have not contacted the customer care have a higher rate of churning than the one who have contacted.
- Likewise, customers who get cashback have lesser churn rate than the customer without cashback offers.
- Customers with maximum no of device are churning more.
- We need to focus on elongating the tenure of the customer as it significantly decreases the churn rate.
- Customers using mobiles are more in number and show the least churn rate.

```
<seaborn.axisgrid.PairGrid at 0x26b489aeb80>
```

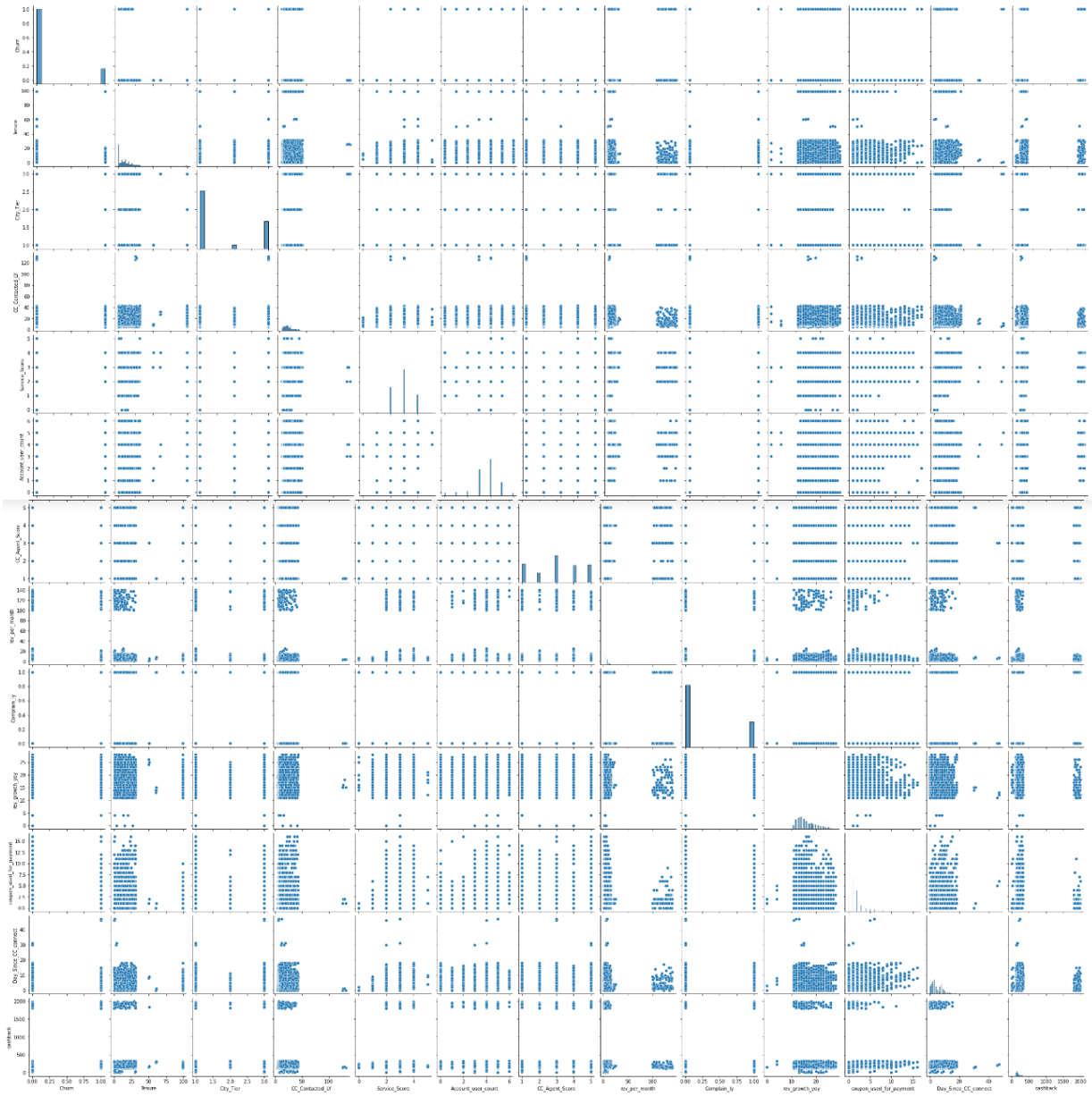

Fig 13 – Pairplot to visualize the bivariate analysis of all variables in the customer churn dataset

Multivariate Analysis:

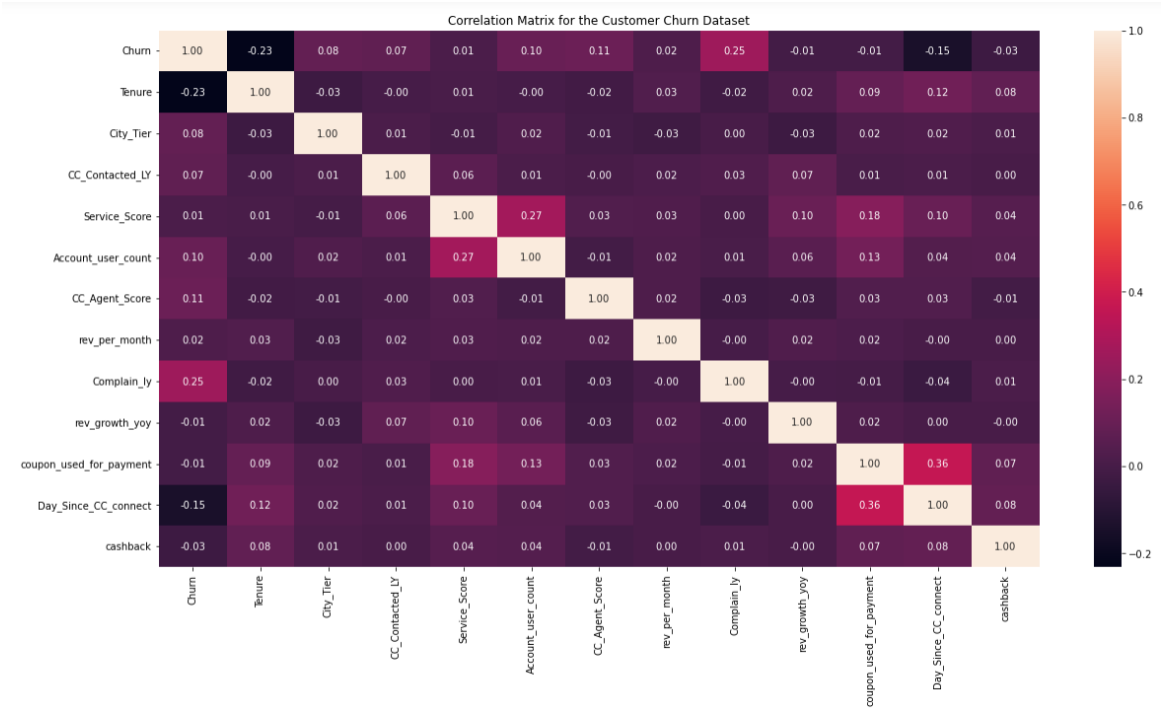

Fig 14 – Correlation Heatmap to visualize the bivariate analysis of all variables in the customer churn dataset

Inferences:

From the correlation heat map (Fig 9), we can notice the following:

- Order count and coupon used show a negative correlation with the average cashback per order, which indicated that as the number of coupons increase, the cashback amount decreases.
- Order count and coupon used have a very high correlation.
- Coupon used and days since the last order have very high correlation which could mean that people take a lot of time to come back after placing large orders.
- Tenure and cashback amount show a high correlation.

Problem 3 Data Cleaning and Pre-processing.

Approach used for identifying and treating missing values and outlier treatment (and why).

Fixing Data Anomalies:

- We see some bad Values or typo-errors in the dataset, like #, $, * etc. Firstly, we will first treat these in Numerical data in the excel file and then import them. Then we must perform missing value imputation i.e., replacing these values with NaN values.

- Also, variable "Login_Device" has character name &&&& which we are replacing with "Both" as the details of the device logged in can be from both laptop and mobile.
- Variable, account_segment has unique values but it has discrepancy with the values present in it, "Super Plus" and "Super + ", "Regular +" and "Regular Plus" indicate the same account variants. We replace 'Super+' as 'Super Plus' and 'Regular +' and 'Regular Plus'.
- Gender variable has values as 'F', 'Female', 'M' and 'Male'. We replace the values 'F' and 'M' as Female and male respectively.

Missing Value treatment:

Table 15 shows that there are missing values present in the dataset.

Churn	False
Tenure	True
City_Tier	True
CC_Contacted_LY	True
Payment	True
Gender	True
Service_Score	True
Account_user_count	True
account_segment	True
CC_Agent_Score	True
Marital_Status	True
rev_per_month	True
Complain_ly	True
rev_growth_yoy	False
coupon_used_for_payment	False
Day_Since_CC_connect	True
cashback	True
Login_device	True
dtype: bool	

Table 15 – Missing values in the variables in customer churn dataset

We need to treat the missing values. We have removed missing values by treating categorical and continuous columns differently.

- For numerical columns as the dataset has extreme values/outliers, we will replace these missing values with median values.
- For categorical column, we will replace the missing values with 'No_info'.

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

Table 16 – Missing values in the variables in customer churn dataset – after treatment

Fixing Duplicate Values:

From below Table 17, we see that there are duplicate values in the dataset.

```
There are total 265 duplicate records in the dataset
```

Table 17 – Checking for duplicate values in customer churn dataset

We need to remove duplicates before modelling the data.

```
There are total 0 duplicate records in the dataset
```

Table 18 – Duplicate values removed from customer churn dataset

Table 18 shows there are no duplicates after cleaning the data.

Outlier treatment:

To confirm our analysis, we will further detect outliers and decide how these outliers should be treated.

```
lower range of churn: 0.0 upper range of churn: 0.0
lower range of City_Tier: -2.0 upper range of City_Tier: 6.0
lower range of CC_Contacted_LY: -7.0 upper range of CC_Contacted_LY: 41.0
lower range of Service_Score: 0.5 upper range of Service_Score: 4.5
lower range of CC_Agent_Score: -1.0 upper range of CC_Agent_Score: 7.0
lower range of Complain_ly: -1.5 upper range of Complain_ly: 2.5
```

Table 19 – Outlier range in customer churn dataset

```
Outlier proportion: Churn: 16.83%
Outlier proportion: Tenure: 1.26%
Outlier proportion: City_Tier: 0.0%
Outlier proportion: CC_Contacted_LY: 0.38%
Outlier proportion: Service_Score: 0.12%
Outlier proportion: Account_user_count: 9.75%
Outlier proportion: CC_Agent_Score: 0.0%
Outlier proportion: rev_per_month: 1.68%
Outlier proportion: Complain_ly: 0.0%
Outlier proportion: rev_growth_yoy: 0.03%
Outlier proportion: coupon_used_for_payment: 12.42%
Outlier proportion: Day_Since_CC_connect: 1.16%
Outlier proportion: cashback: 8.61%
```

Table 20 – Outlier proportion of customer churn dataset

Fig 15 shows that all the outliers are removed from the dataset.

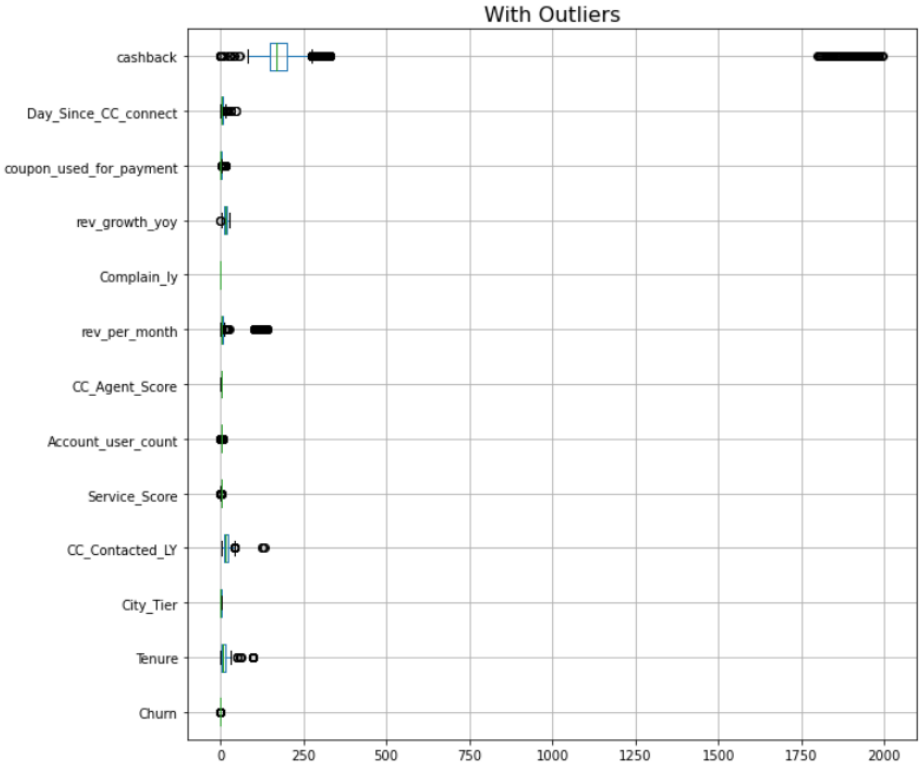

Fig 15 – Outliers in customer churn dataset

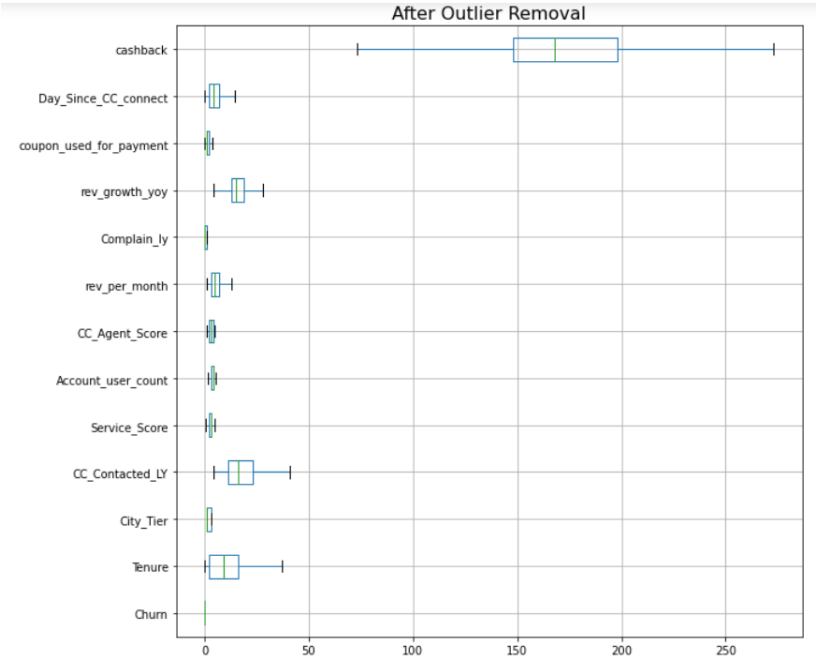

Fig 16 – Outliers in customer churn dataset after removal

Need for variable transformation (if any).

Variable transformation:

Variable transformation makes the data work better during model building. Listing down below three most common approaches for data encoding:

- Dummy variable Encoding
- Label Encoding

Dummy variable Encoding:

We perform dummy variable encoding for the columns Marital_Status and Payment. Table 16 shows the modified dataset.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10995 entries, 0 to 11259
Data columns (total 25 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Churn                      10995 non-null  float64
 1   Tenure                     10995 non-null  int64
 2   City_Tier                  10995 non-null  object
 3   CC_Contacted_LY            10995 non-null  float64
 4   Gender                     10995 non-null  object
 5   Service_Score              10995 non-null  float64
 6   Account_user_count         10995 non-null  float64
 7   account_segment            10995 non-null  object
 8   CC_Agent_Score             10995 non-null  float64
 9   rev_per_month              10995 non-null  float64
 10  Complain_ly                10995 non-null  float64
 11  rev_growth_yoy             10995 non-null  float64
 12  coupon_used_for_payment    10995 non-null  float64
 13  Day_Since_CC_connect       10995 non-null  float64
 14  cashback                   10995 non-null  float64
 15  Login_device               10995 non-null  object
 16  Marital_Status_Married     10995 non-null  uint8
 17  Marital_Status_No_info     10995 non-null  uint8
 18  Marital_Status_Single      10995 non-null  uint8
 19  Payment_Cash on Delivery   10995 non-null  uint8
 20  Payment_Credit Card        10995 non-null  uint8
 21  Payment_Debit Card         10995 non-null  uint8
 22  Payment_E wallet           10995 non-null  uint8
 23  Payment_No_info            10995 non-null  uint8
 24  Payment_UPI                10995 non-null  uint8
dtypes: float64(11), int64(1), object(4), uint8(9)
memory usage: 1.5+ MB
```

Table 21 – Redefined customer churn dataset – after dummy encoding

Label Encoding:

We perform label encoding for the columns account_segment and Gender. We drop the columns "Marital_Status_No_info" and "Payment_No_info" as it is column created for missing values and has no point in creating model.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10995 entries, 0 to 11259
Data columns (total 23 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Churn                      10995 non-null  float64
 1   Tenure                     10995 non-null  int64
 2   City_Tier                  10995 non-null  object
 3   CC_Contacted_LY            10995 non-null  float64
 4   Gender                     10995 non-null  int32
 5   Service_Score              10995 non-null  float64
 6   Account_user_count         10995 non-null  float64
 7   account_segment            10995 non-null  int32
 8   CC_Agent_Score             10995 non-null  float64
 9   rev_per_month              10995 non-null  float64
 10  Complain_ly                10995 non-null  float64
 11  rev_growth_yoy             10995 non-null  float64
 12  coupon_used_for_payment    10995 non-null  float64
 13  Day_Since_CC_connect       10995 non-null  float64
 14  cashback                   10995 non-null  float64
 15  Login_device               10995 non-null  object
 16  Marital_Status_Married     10995 non-null  uint8
 17  Marital_Status_Single      10995 non-null  uint8
 18  Payment_Cash on Delivery   10995 non-null  uint8
 19  Payment_Credit Card        10995 non-null  uint8
 20  Payment_Debit Card         10995 non-null  uint8
 21  Payment_E wallet           10995 non-null  uint8
 22  Payment_UPI                10995 non-null  uint8
dtypes: float64(11), int32(2), int64(1), object(2), uint8(7)
memory usage: 1.4+ MB
```

Table 22 – Structure of the new customer churn dataset

Variables removed or added and why (if any).

Removal of unwanted variables:

- We have removed all unwanted variables and special characters as well as symbols like *, $, # from all columns.
- Also, we have removed unwanted column 'AccountID'.

Addition of new variables:

We don't have to add any new variable in given data set as of now. We can add it later when building the model.

Problem 4 Model building.

Clear on why was a particular model(s) chosen.

We have built 3 models for prediction.

- CART/Decision Tree Classifier Model
- Logistic Regression
- Random Forest

- The above models were built as the problem was that of classification.
- Apart from prediction of the customers who would churn, we also wanted to know which of the features are more important in predicting the customer churn and hence which of features should be focused upon to reduce the churn.
- All the models above provide the customer prediction as well as feature importance. Further the models should be easy to understand.
- Hence Logistic Regression, Random Forest and Decision trees are the **easiest to understand and transparent.**
- As a part of ensembling technique we have built Random Forest model.

Splitting data into Train and Test Dataset (70:30):

- For building the models we will now have to split the dataset into training and testing data with the ratio of 70:30. These two datasets are stored in 'X_train' and 'X_test' with their corresponding dimensions as follows. Before which we must split the dataset into dependent and independent variables. The dimensions of the training and test data are below, the samples are almost equally distributed between the train and test datasets:

```
X_train (7696, 25)
X_test (3299, 25)
y_train (7696,)
y_test (3299,)
```

Table 23 – Train and Test dataset of customer churn dataset

Effort to improve model performance.

Since the data was imbalanced, SMOTE technique is applied to balance the data and thereafter it was observed that the accuracy as well as specificity of the models increased considerably. Once the data is balanced, we apply GridSearchCV on all three models and tune the hyperparameters.

```
{'criterion': 'gini', 'max_depth': 8, 'min_samples_leaf': 40, 'min_samples_split': 250}
```

```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=8, min_samples_leaf=40, min_samples_split=250,
                       random_state=1)
```

Table 24 – Best Parameter – GridSearchCV - DT model tuned

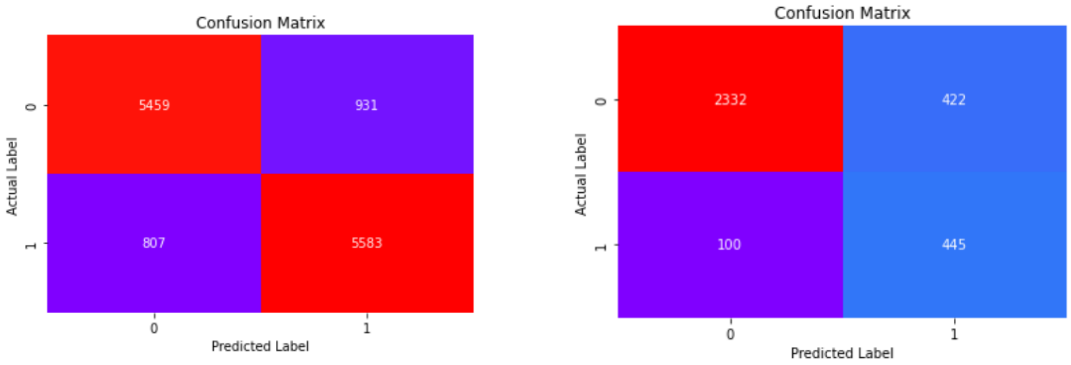

Fig 17 - Confusion Matrix graph train(left) & test(right) dataset – DT model tuned

```
GridSearchCV
estimator: LogisticRegression
LogisticRegression
```

```
{'penalty': 'none', 'solver': 'lbfgs', 'tol': 0.0001}

LogisticRegression(max_iter=10000, n_jobs=2, penalty='none')
```

Table 25 – Best Parameter – GridSearchCV - LR model tuned

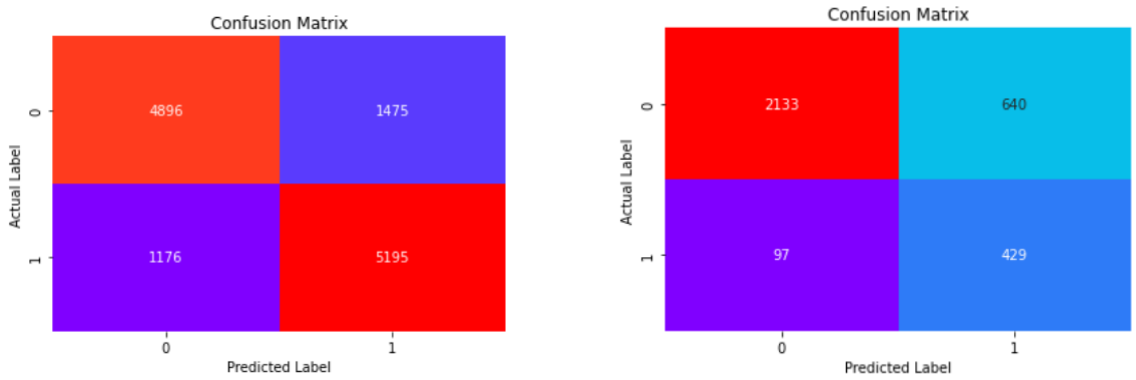

Fig 18 - Confusion Matrix graph train(left) & test(right) dataset – LR Model Tuned

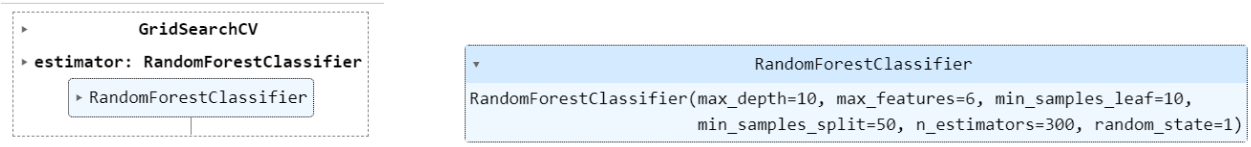

Table 26 – Best Parameter – GridSearchCV - RF model tuned

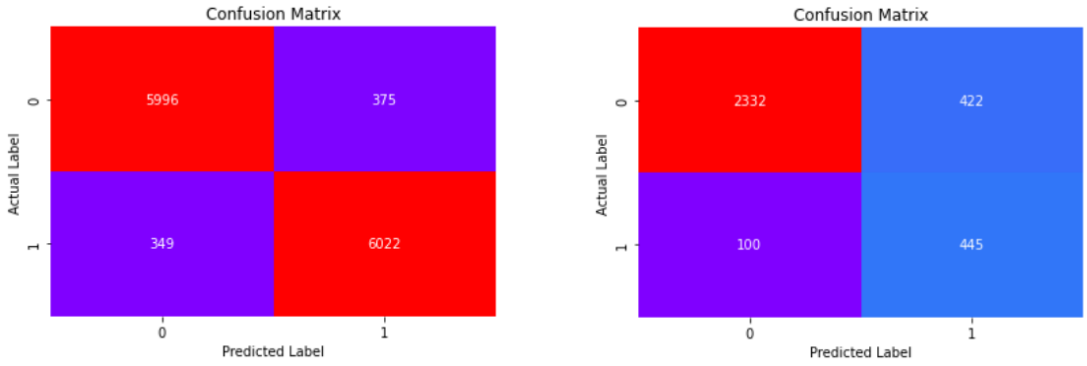

Fig 19 - Confusion Matrix graph train(left) & test(right) dataset – RF Model Tuned

Inferences from Models:

CART/Decision Tree:

- CART Model Score on Train Data is 86% and Test data is 84%.
- Here we can see that after the Data was balanced using SMOTE, the accuracy, sensitivity and specificity of the models have increased considerably**.**

- Noticeably, CART Model without Tuning was going with the baseline approach calling everything as true hence it's an extreme.
- In this case also, Recall is more important than Precision and business would like to have less False Negatives in trade off to have more False Positives. Meaning, getting a False Negative will be costlier for the companies.
- Model can be considered a good model.

Logistic Regression:

- From the confusion matrix on Train and Test data, it is clear that Train data is 87% accurate as compared to 86% accuracy shown by Test Data.
- The Area Under the Curve (AUC) score for Train Dataset is 87% whereas AUC Score for Test Dataset is 86%.
- The score on Train Data is 0.79 and that on Test Dataset is 0.78. KS values lie between 0 and 1 where 1 is termed as a good fit and 0 as not a good fit model. Out values lie in between which means that the model is good fit model.
- Noticeably, train and test set are do not show similar results and deviation is more than 10%. Hence, **fine tuning the model has not improved the model performance** much and model is overfitting**.**
- Logistic Regression Model Score with Grid search on Train Data and Test data is only 82%. Train and Test data scores are not in line and the overall performance of model on test data does not look good. Hence, it can be inferred that overall, this model cannot be considered as a good model.

Random Forest:

- The model built with Best Parameter is performing better than original model but still is overfit Model. Best parameter is identified using trial and error method for by varying various parameters in the GridSearchCV algorithm.
- We get a very good AUC score of 99% for the training data set and the ROC curve for the same is observed.
- Training and Test set results are almost similar, and with the overall measures high, the model is a good model. AUC is for train is 99 % and for test it is 95%.
- F1 score, precision and recall values are somewhat similar though recall and F1 score are closer. For both train and test we have FP's and FN's on the lower side.
- Noticeably, train and test set are giving almost similar results and deviation is less than 10% it is safe to say that model is not overfitting or underfitting.
- In this case also, Recall is more important than Precision and business would like to have less False Negatives in trade off to have more False Positives. Meaning, getting a False Negative will be costlier for the companies.

Problem 5 Model validation.

How was the model validated? Just accuracy, or anything else too?

Models were compared based on the following metrics:

- Specificity (ability to correctly predict the churn)
- Accuracy (Denoted by Model Score)
- AUC-Score and area under ROC Curves.
- Feature importance as validated in the EDA.

Below in Table 21 & 22 we can see we can see the scores of the various models built.

- Based on the Senstivity and specificity scores, the AUC - RoC score and overall scores, Random Forest model seems to be highly overfit and hence might be a unreliable. Logistic Regression, and Decision Tree seem to be doing well. However, upon looking at the feature importance, Random Forest doesn't look like a reliable one**.**
- Between Logistic Regression and Decision Tree models, Logistic is more accurate.

	Decision Tree Train	Decision Tree Test	Random Forest Train	Random Forest Test	Logistic Regression Train	Logistic Regression Test
Accuracy	0.88	1.0	0.93	1.0	0.88	0.89
AUC	0.80	1.0	0.97	1.0	0.87	0.88
Recall	1.00	1.0	0.63	1.0	0.48	0.44
Precision	1.00	1.0	0.93	1.0	0.76	0.77
F1 Score	1.00	1.0	0.75	1.0	0.59	0.56

Table 27 – Performance Metrics – Score comparison of all models unbalanced

	DT Train GRID	DT Test GRID	RF Train GRID	RF Test GRID	LR Train GRID	LR Test GRID
Accuracy	0.86	0.84	0.94	0.91	0.79	0.78
AUC	0.93	0.88	0.99	0.95	0.87	0.86
Recall	0.87	0.82	0.95	0.80	0.82	0.82
Precision	0.86	0.51	0.94	0.67	0.78	0.40
F1 Score	0.87	0.63	0.94	0.73	0.80	0.54

Table 28 – Performance Metrics – Score comparison of all models – Tuned

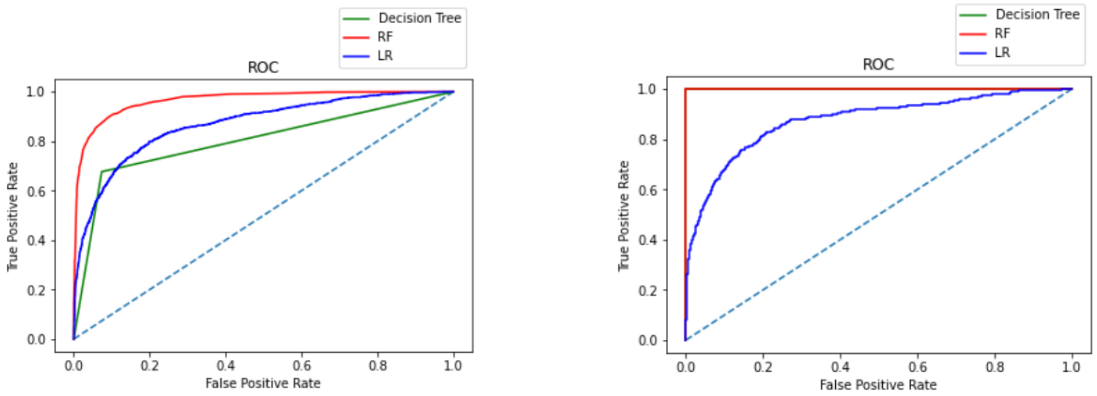

Fig 20 - ROC curve analysis train(left) and test(right) – All models unbalanced

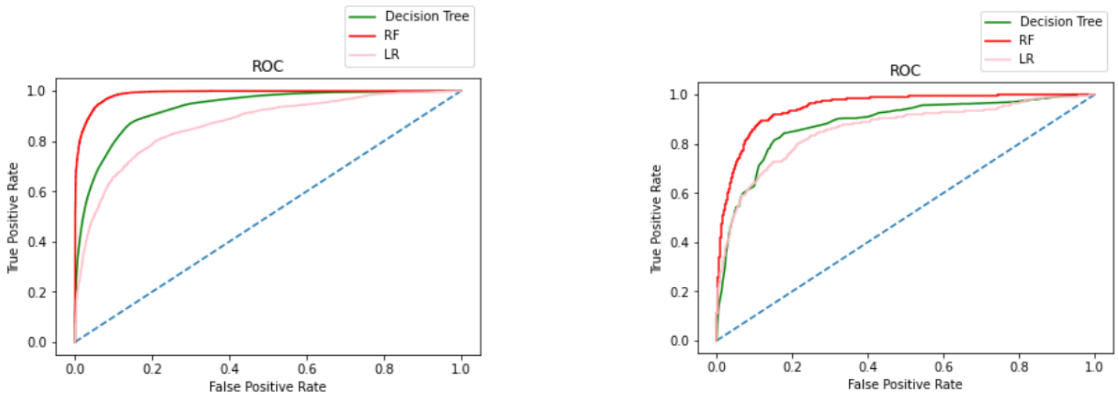

Fig 21 - ROC curve analysis train(left) and test(right) – All models tuned data

- Logistic Regression, CART and Random Forest models are all able to predict the Customer churn with more than 80% accuracy.
- SMOTE technique used to balance the data has drastically improved the accuracy of KNN model but has made the model overfit.
- GridSearchCV technique has also improved the accuracy of models.
- As per the Performance Metrics computed above for all the models it can be evidently seen that the best model for our prediction is **CART Model and Logistic Regression Model have the highest Precision, Area under the curve and Accuracy**.
- Logistic Regression turns out to be the best models owing to the following reasons:
  - High AUC score
  - Better Specificity
  - Correct prediction of the feature importance.

Problem 6 Final interpretation / recommendation.

Detailed recommendations for the management/client based on the analysis done

Insights from EDA:

- New customers with tenures less than 5 years are generally churned more often.
- Churned customers have their preferred login device as Mobile phones.
- Majority of churned customers belong to Tier 1 cities and have their preferred mode of payment as Debit card.
- Churned customers have contacted the customer care almost 10-20 times in the last year.
- Majority of churned customers are males.
- The customers that are churned majorly have given a service score of 3 and customer care agent score as 3. Majority of the churned accounts had 2-3 users per account.
- A major number of churned customers had "regular Plus" plan.
- Majority of the customers churned are single.
- Revenue generated of majority of churned accounts is between 0-5 which is quite low and revenue growth from last year is between 12-14%.
- Churned customers have used very few coupons.
- The customers that have made several complaints are mostly churned.

Business Recommendations:

- New customers should be given more benefit in terms of discounts and promotions.
- Customers who are giving satisfaction score more than 3 should be targeted to provide more satisfaction in terms of services and products.
- Complaints should be taken seriously by the company and customer service should be the top priority.
- An option to customize plans and a freedom to pause and skip subscription plans could attract more customers.
- Loyalty programs to encourage and appreciate long term subscribers could help reduce existing customer churn.
- Competitive marketing strategies should be taken into consideration. Market research should be done to provide good pricing and discounts.
- Family plans should be made cost effective so that the entire family can avail the benefits from a single account.
- Loyalty programs to encourage and appreciate long term subscribers.
- Company should focus on check out pages to make seamless payments.
- Also, check out point can include simple and consistent brand awareness designs as promotional offers.
- The industry standards for Churn rate in consumer goods industry is around 9.62%. It is determined that a churn rate of 10% will cost approx. 10% of the total income. The Churn rate in the data provided is around 18% which is high.
- Hence, reducing the Churn to 10% will increase the revenue by 8%.