Final Report

Customer Churn Prediction

Marripally Ravikumar

PGP – DSBA Online

May 2021

Date: 05/06/2022

Table of contents

List of Figures

List of Tables

Q1. Introduction

Problem statement

A DTH provider is facing a lot of competition in the current market and it has become a challenge to retain the existing customers in the current situation. Hence, the company wants to develop a model through which they can do churn prediction of the accounts and provide segmented offers to the potential churners. In this company, account churn is a major thing because one account can have multiple customers. hence by losing one account the company might be losing more than one customer.

Need for the study

- Account churning is a major problem in the DTH industry.
- Retaining existing customers has become a big challenge.
- Churn rate of the given DTH company is 16.8% which is higher than the mean churn rate (10%) in the DTH industry.
- Cost of acquiring a new customer is almost 5 to 6 times higher than retaining the customer. Hence, it is better to invest in retaining new customers than attracting new customers.
- Every company has to focus to reduce the churn rate to improve profits.
- The main goal of this project is to help the DTH company by developing a model to predict churn by using Machine Learning Algorithms.
- Then, we need to suggest customer-specific offers based on the revenue generated by them.

Q2. EDA and Business Implication

Visual and Non-Visual Understanding of the Data

Sample of the Dataset

	0	1	2	3	4	5	6	7	8	9
AccountID	20000	20001	20002	20003	20004	20005	20006	20007	20008	20009
Churn	1	1	1	1	1	1	1	1	1	1
Tenure	4	0	0	0	0	0	2	0	13	0
City_Tier	3.0	1.0	1.0	3.0	1.0	1.0	3.0	1.0	3.0	1.0
CC_Contacted_LY	6.0	8.0	30.0	15.0	12.0	22.0	11.0	6.0	9.0	31.0
Payment	Debit Card	UPI	Debit Card	Debit Card	Credit Card	Debit Card	Cash on Delivery	Credit Card	E wallet	Debit Card
Gender	Female	Male	Male	Male	Male	Female	Male	Male	Male	Male
Service_Score	3.0	3.0	2.0	2.0	2.0	3.0	2.0	3.0	2.0	2.0
Account_user_count	3	4	4	4	3	NaN	3	3	4	5
account_segment	Super	Regular Plus	Regular Plus	Super	Regular Plus	Regular Plus	Super	Regular Plus	Regular Plus	Regular Plus
CC_Agent_Score	2.0	3.0	3.0	5.0	5.0	5.0	2.0	2.0	3.0	3.0
Marital_Status	Single	Single	Single	Single	Single	Single	Divorced	Divorced	Divorced	Single
rev_per_month	9	7	6	8	3	2	4	3	2	2
Complain_ly	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0
rev_growth_yoy	11	15	14	23	11	22	14	16	14	12
coupon_used_for_payment	1	0	0	0	1	4	0	2	0	1
Day_Since_CC_connect	5	0	3	3	3	7	0	0	2	1
cashback	159.93	120.9	NaN	134.07	129.6	139.19	120.86	122.93	126.83	122.93
Login_device	Mobile	Mobile	Mobile	Mobile	Mobile	Computer	Mobile	Mobile	Mobile	Mobile

Table 1. Sample of the Dataset.

Insights

1. There are 18 features (columns) with 11260 observations (rows) in the dataset.
2. The dataset has both numerical variables and categorical variables.
3. The **target variable** in this dataset is **Churn**.
4. There are no duplicate observations in the dataset.

Let us Understand Every Feature in detailed

Variable	Description
AccountID	Account unique identification number.
Churn	Account churn flag. This is the target variable to be predicted.
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_L12m	How many times all the customers of the account have contacted customer care in the last 12 months
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 the 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 the 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 have been raised by account in the last 12 months
rev_growth_yoy	revenue growth percentage of the account (last 12 months vs last 24 to 13 months)
coupon_used_l12m	How many times customers have used coupons to do the payment in the last 12 months
Day_Since_CC_connect	Number of days since no customers in the account have contacted the customer care
cashback_l12m	Monthly average cashback generated by account in the last 12 months
Login device	Preferred login device of the customers in the account

Basic Information of the 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  object
 2   city_tier                11148 non-null  float64
 3   cc_contacted_ly          11158 non-null  float64
 4   payment                  11151 non-null  object
 5   gender                   11152 non-null  object
 6   service_score            11162 non-null  float64
 7   account_user_count       11148 non-null  object
 8   account_segment          11163 non-null  object
 9   cc_agent_score           11144 non-null  float64
 10  marital_status           11048 non-null  object
 11  rev_per_month            11158 non-null  object
 12  complain_ly              10903 non-null  float64
 13  rev_growth_yoy           11260 non-null  object
 14  coupon_used_for_payment  11260 non-null  object
 15  day_since_cc_connect     10903 non-null  object
 16  cashback                 10789 non-null  object
 17  login_device             11039 non-null  object
dtypes: float64(5), int64(1), object(12)
memory usage: 1.5+ MB
```

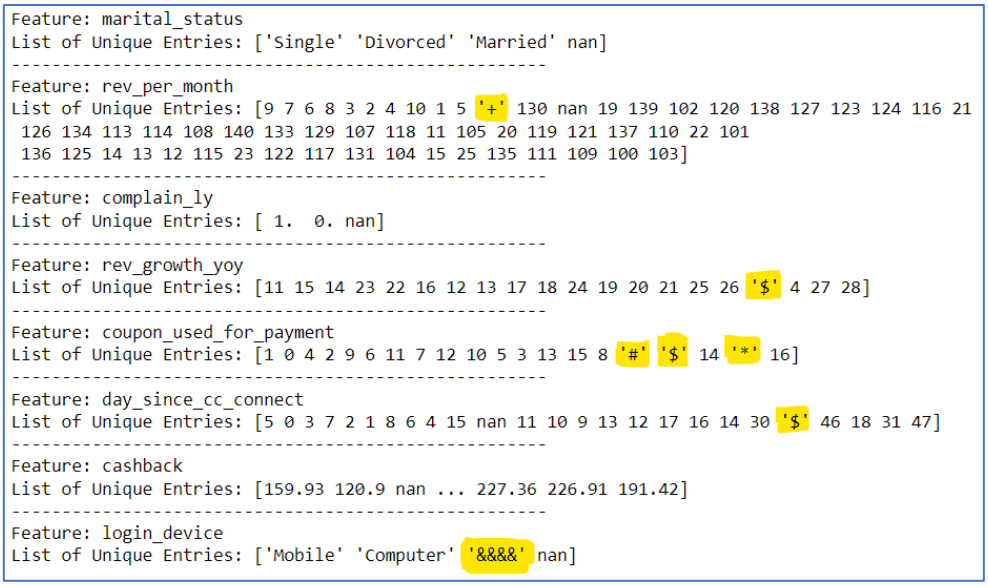
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11260 entries, 0 to 11259
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   churn                    11260 non-null  int64
 1   tenure                   11042 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       10816 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           11257 non-null  float64
 14  coupon_used_for_payment  11257 non-null  float64
 15  day_since_cc_connect     10902 non-null  float64
 16  cashback                 10787 non-null  float64
 17  login_device             10500 non-null  object
dtypes: float64(12), int64(1), object(5)
memory usage: 1.5+ MB
```

Table 2. Data Types of All Features in the Dataset before and after cleaning.

- From the table, we can notice that features like **tenure, rev_per_month, rev_growth_yoy, coupon_used_for_payment, day_since_cc_connect and cashback** are expected to have **numerical data** types but they have an **object data** type.
- We need to find out why these features have object data types before proceeding to further analysis.

Anomalies in the dataset

```
Feature: churn
List of Unique Entries: [1 0]
----------------------------------------------------
Feature: tenure
List of Unique Entries: [4 0 2 13 11 '#' 9 99 19 20 14 8 26 18 5 30 7 1 23 3 29 6 28 24 25 16 10
 15 22 nan 27 12 21 17 50 60 31 51 61]
----------------------------------------------------
Feature: city_tier
List of Unique Entries: [ 3.  1. nan  2.]
----------------------------------------------------
Feature: cc_contacted_ly
List of Unique Entries: [  6.   8.  30.  15.  12.  22.  11.   9.  31.  18.  13.  20.  29.  28.
  26.  14.  10.  25.  27.  17.  23.  33.  19.  35.  24.  16.  32.  21.
  nan  34.   5.   4. 126.   7.  36. 127.  42.  38.  37.  39.  40.  41.
 132.  43. 129.]
----------------------------------------------------
Feature: payment
List of Unique Entries: ['Debit Card' 'UPI' 'Credit Card' 'Cash on Delivery' 'E wallet' nan]
----------------------------------------------------
Feature: gender
List of Unique Entries: ['Female' 'Male' 'F' nan 'M']
----------------------------------------------------
Feature: service_score
List of Unique Entries: [ 3.  2.  1. nan  0.  4.  5.]
----------------------------------------------------
Feature: account_user_count
List of Unique Entries: [3 4 nan 5 2 '@' 1 6]
----------------------------------------------------
Feature: account_segment
List of Unique Entries: ['Super' 'Regular Plus' 'Regular' 'HNI' 'Regular +' nan 'Super Plus'
 'Super +']
----------------------------------------------------
Feature: cc_agent_score
List of Unique Entries: [ 2.  3.  5.  4. nan  1.]
```



- Few columns having their entries as **special symbols like #, @, +, $, *, &.** This might be the reason for identifying numerical data types as an object.
- Let us **replace these special symbols with null values** and later null values will be imputed with an appropriate method.
- Few features have duplicate sublevels like Gender having ‘Male’ and ‘M’.
- We need to clean these anomalies before proceeding to further analysis.

Description of the Dataset

Feature/Measure	Tenure	cc contacted ly	Rev per month	Rev growth yoy	Coupon used for payment	Day since cc connect	Cashback
count	11042.0	11158.0	10469.0	11257.0	11257.0	10902.0	10787.0
mean	11.0	17.9	6.4	16.2	1.8	4.6	196.2
std	12.9	8.9	11.9	3.8	2.0	3.7	178.7
min	0.0	4.0	1.0	4.0	0.0	0.0	0.0
25%	2.0	11.0	3.0	13.0	1.0	2.0	147.2
50%	9.0	16.0	5.0	15.0	1.0	3.0	165.3
75%	16.0	23.0	7.0	19.0	2.0	8.0	200.0
max	99.0	132.0	140.0	28.0	16.0	47.0	1997.0

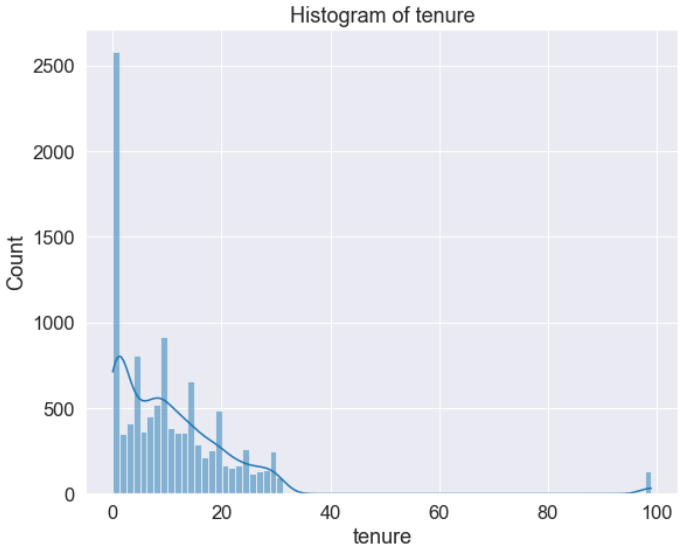
Table 3. Description of Numerical Features in Dataset.

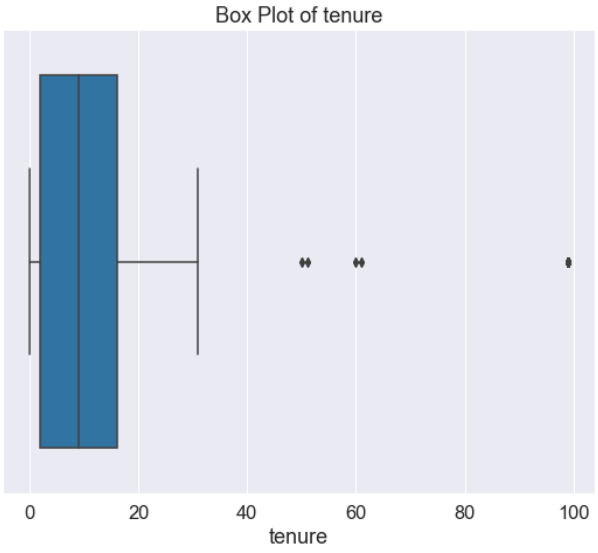
Feature	Count	Unique	Top	Frequency
City tier	11148	3	1	7263
Payment	11151	5	Debit Card	4587
Gender	11152	2	Male	6704
Service score	11162	6	3	5490
Account user count	10816	6	4	4569
Account segment	11163	5	Regular Plus	4124
cc agent score	11144	5	3	3360
Marital status	11048	3	Married	5860
Complain ly	10903	2	0	7792
Login device	10500	2	Mobile	7482

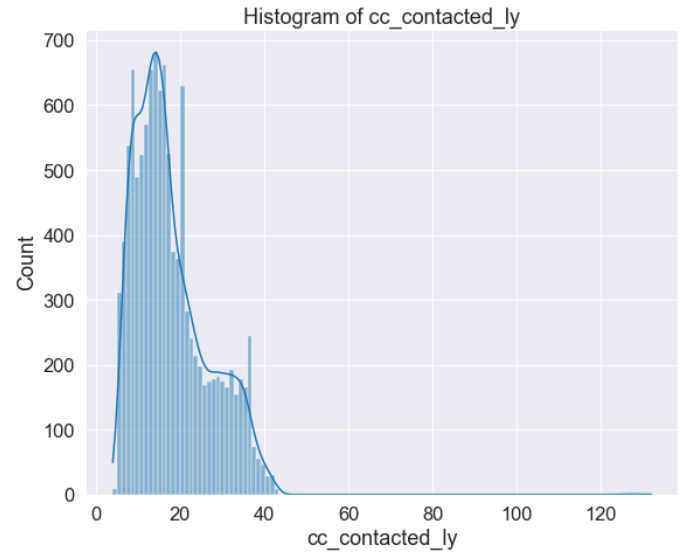
Table 4. Description of Categorical Features in the Dataset.

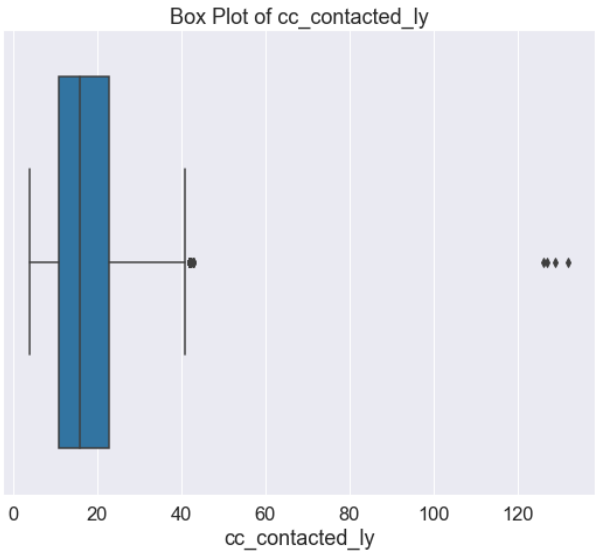
Univariate Analysis

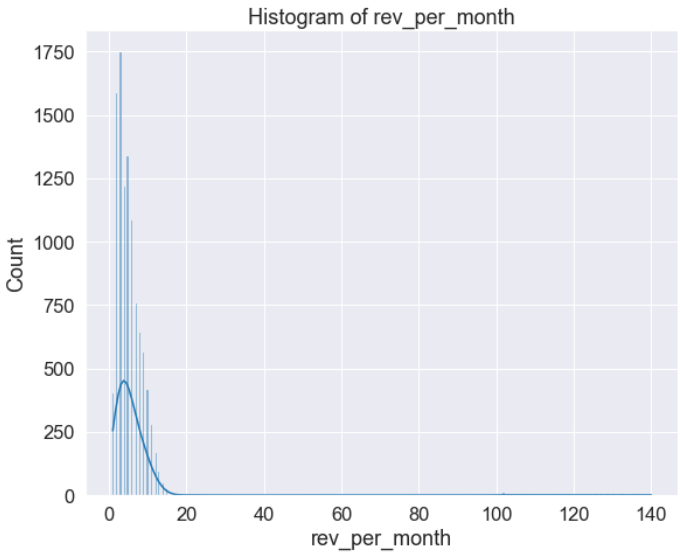
Histogram and Box Plots for Continuous Variables

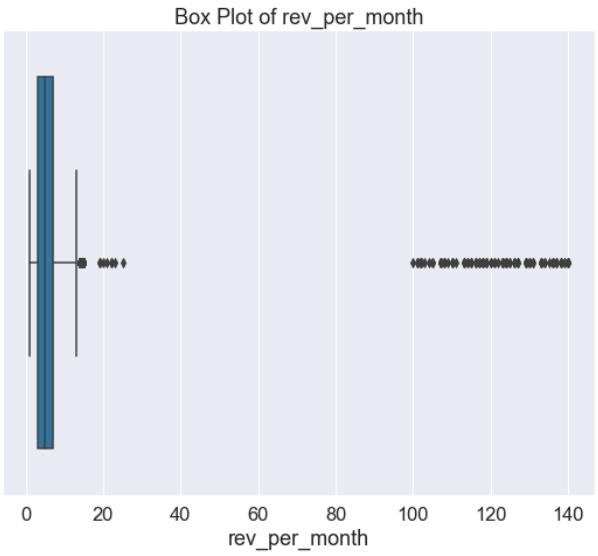


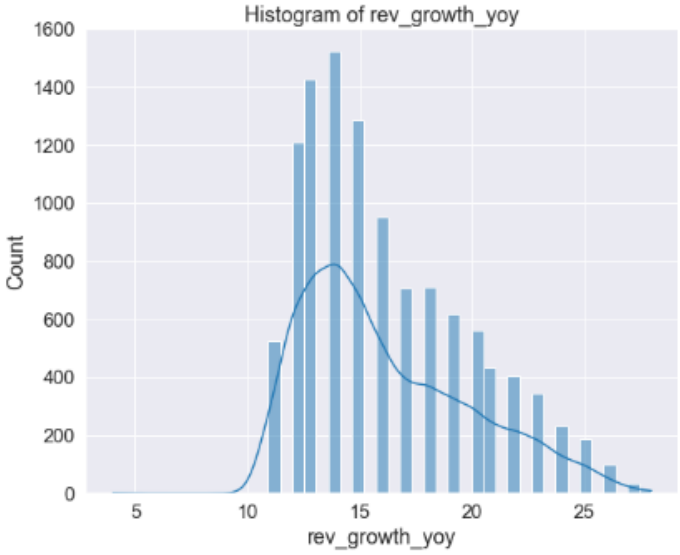


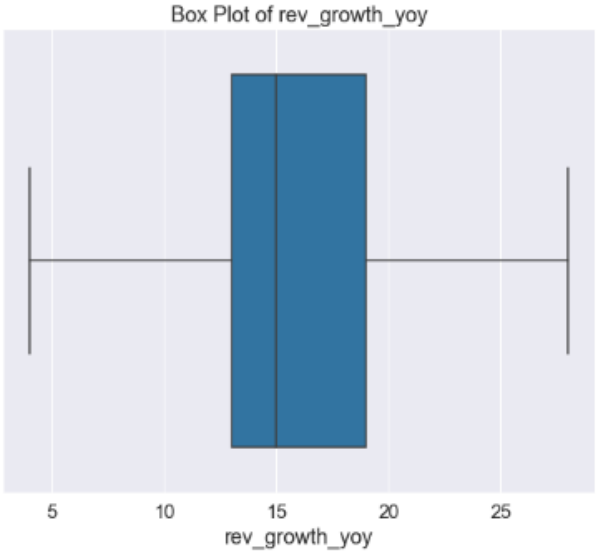


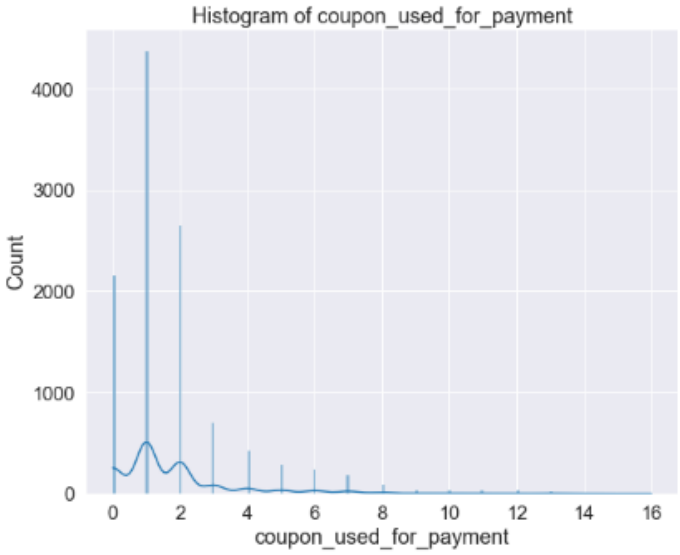


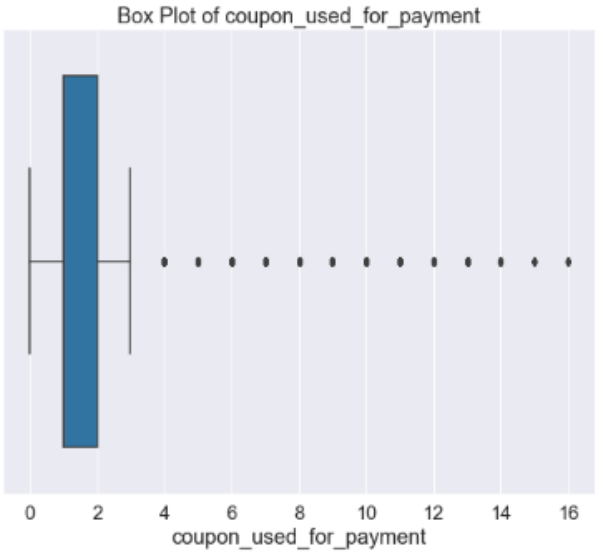


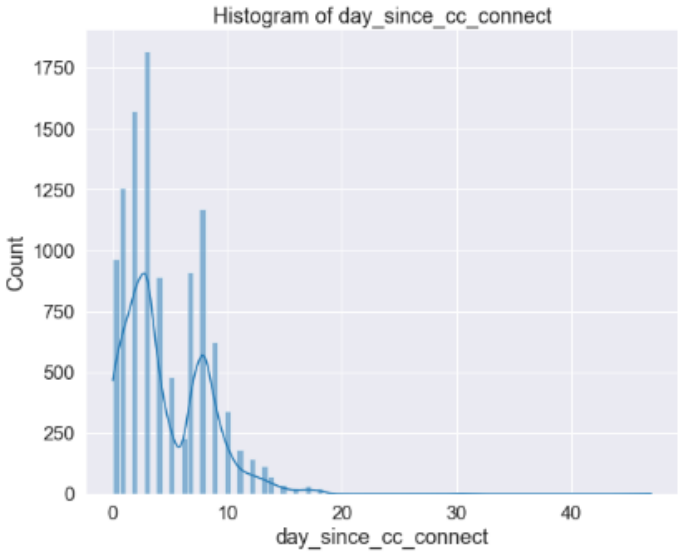


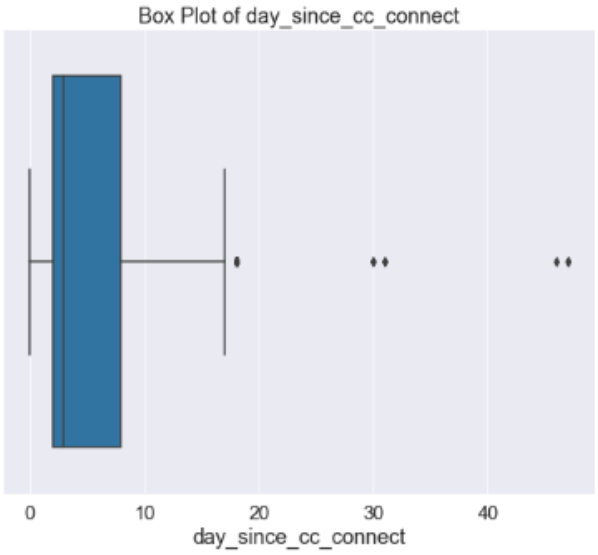

Histogram of rev_growth_yoy
Count
rev_growth_yoy


Box Plot of rev_growth_yoy
rev_growth_yoy


Histogram of coupon_used_for_payment
Count
coupon_used_for_payment


Box Plot of coupon_used_for_payment
coupon_used_for_payment


Histogram of day_since_cc_connect
Count
day_since_cc_connect


Box Plot of day_since_cc_connect
day_since_cc_connect

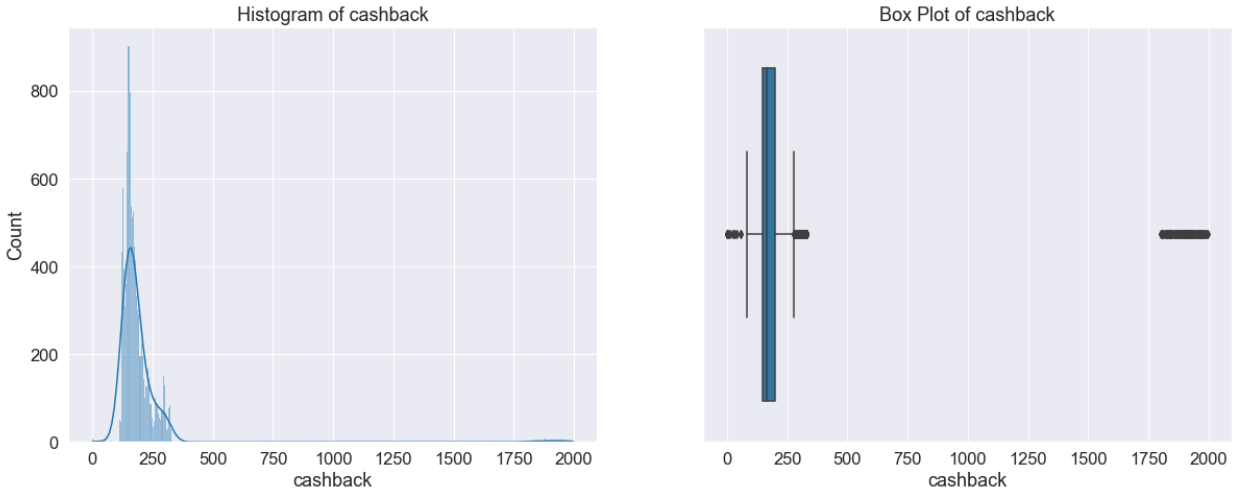


Figure 1. Histogram and Box Plot of Continuous Numerical Features.

Skewness & Kurtosis for Numerical Continuous Variables

- Skewness is a measure of lack of symmetry in distribution.
- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

Feature	Skewness	Kurtosis
Tenure	3.90	23.37
CC contacted LY	1.42	8.23
Rev per month	9.09	86.96
Rev growth YoY	0.75	-0.22
Coupons used for payment	2.58	9.10
Day since cc connect	1.27	5.33
Cashback	8.77	81.11

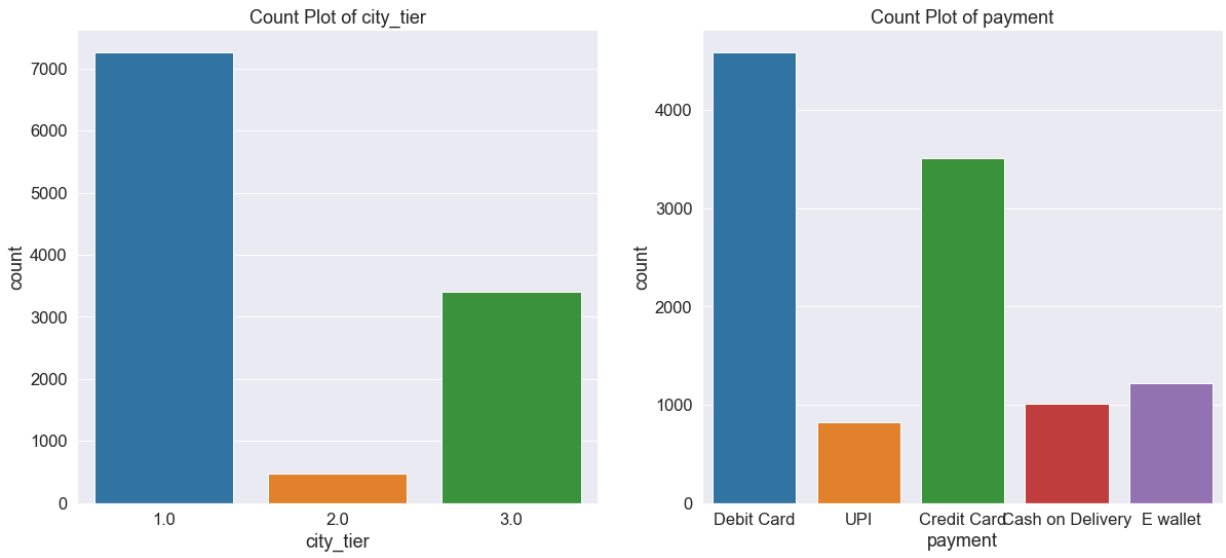
Insights

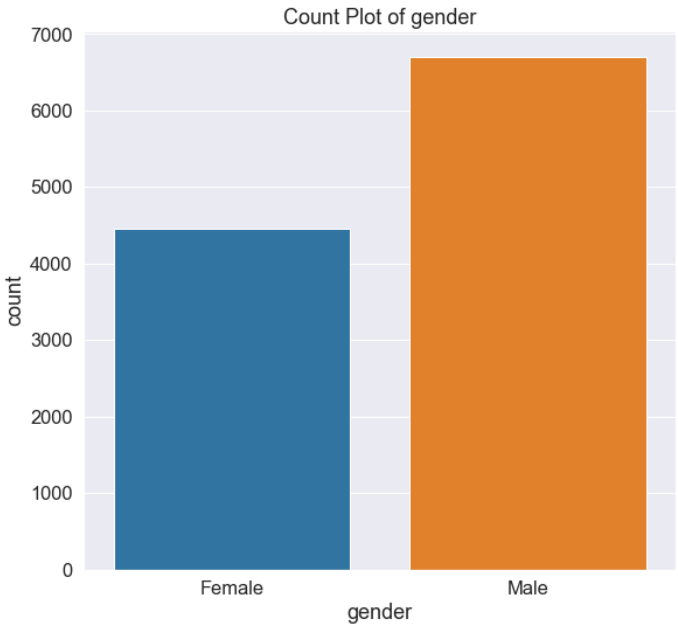
From the above plots and tables, we can conclude the below points,

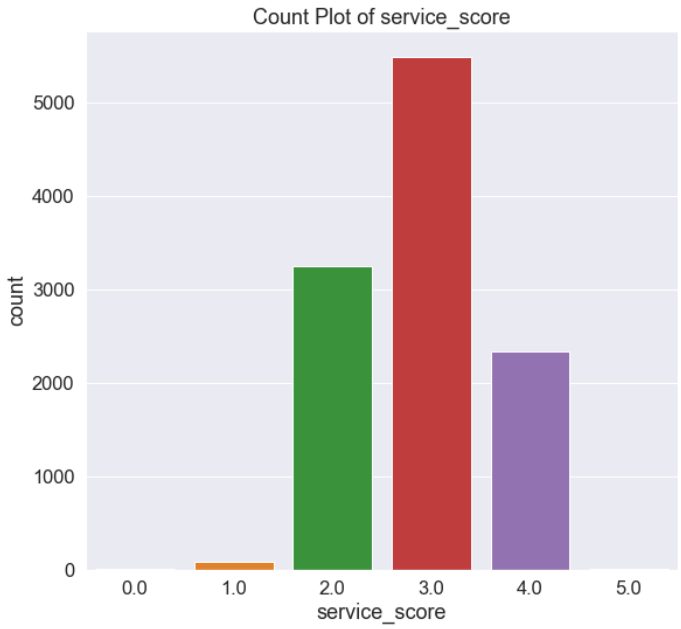
1. All features are having **right-skewed** distribution.
2. Rev_growth_yoy has negative kurtosis (-0.22). All remaining features have **positive kurtosis** values.

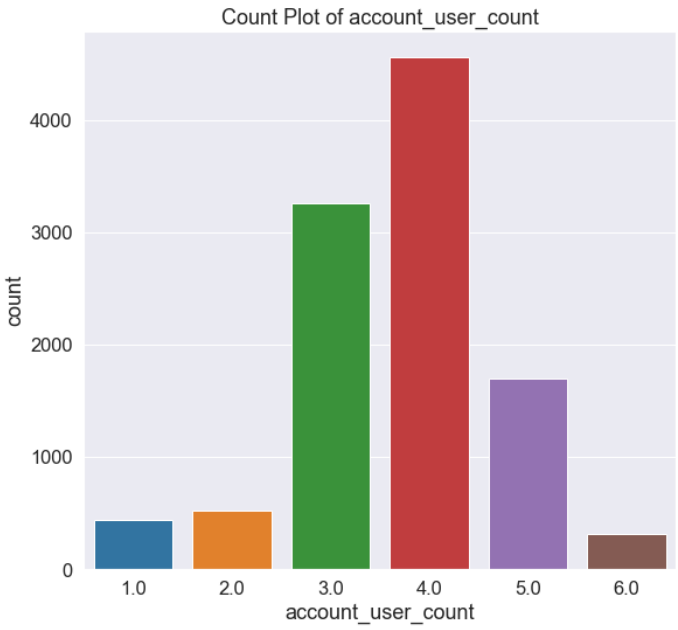
Table 5. Skewness and Kurtosis of Numerical Features.

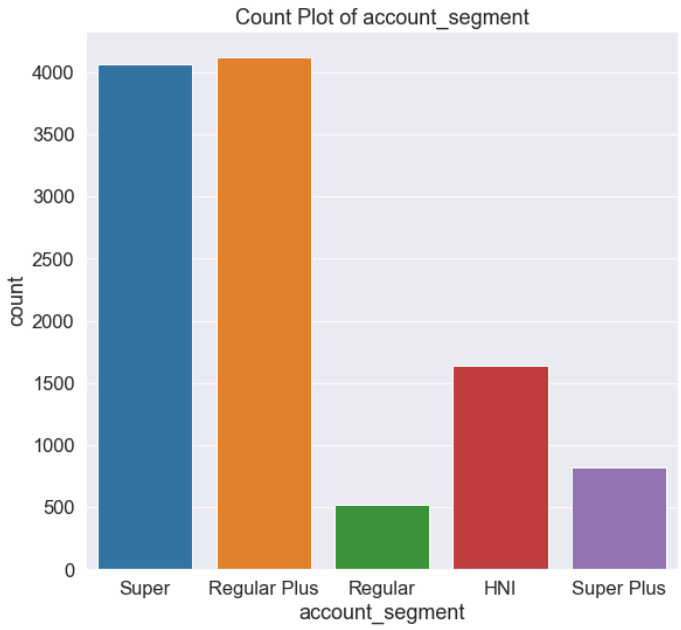
Count Plots for Categorical Features



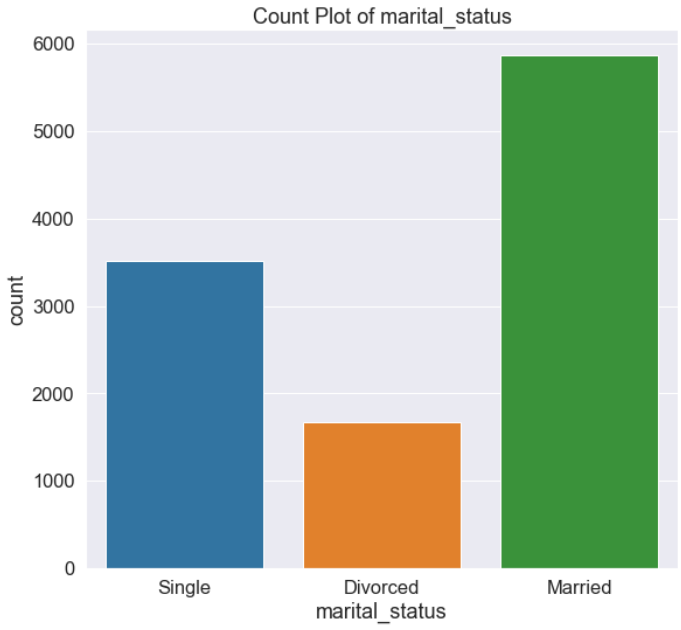

Count Plot of gender
count
gender
Female
Male
0
1000
2000
3000
4000
5000
6000
7000


Count Plot of service_score
count
service_score
0.0
1.0
2.0
3.0
4.0
5.0
0
1000
2000
3000
4000
5000


Count Plot of account_user_count
count
account_user_count
1.0
2.0
3.0
4.0
5.0
6.0
0
1000
2000
3000
4000


Count Plot of account_segment
count
account_segment
Super
Regular Plus
Regular
HNI
Super Plus
0
500
1000
1500
2000
2500
3000
3500
4000


Count Plot of cc_agent_score
count
cc_agent_score
1.0
2.0
3.0
4.0
5.0
0
500
1000
1500
2000
2500
3000
3500


Count Plot of marital_status
count
marital_status
Single
Divorced
Married
0
1000
2000
3000
4000
5000
6000

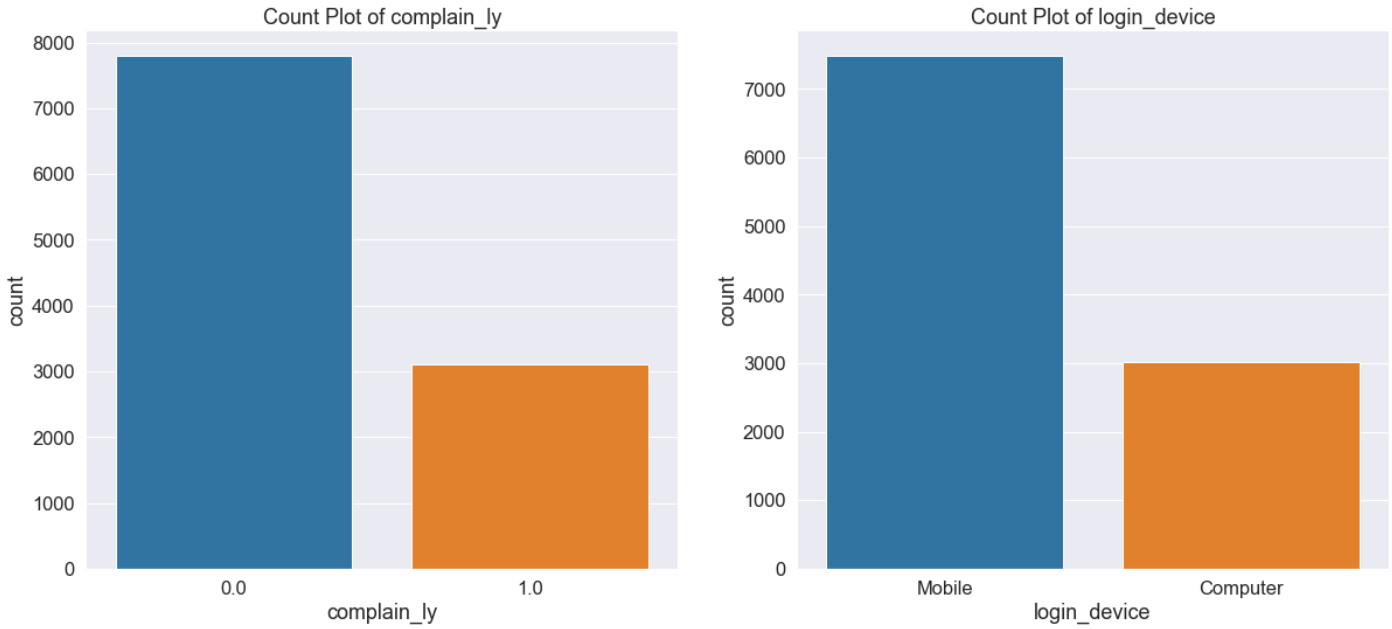


Figure 2. Count Plots for Categorical Features.

Insights and Business Implications

- **Highest** number of Customers from **Tier 1 city (65%).** The order of Cites according to decreasing number of customers is as below.

  Tier 1 (65%) > Tier 3 (31%) > Tier 2 (4%)

- **Highest number** of Customers are preferring to **pay through debit cards**. The order of type of payment according to decreasing number of customers is as below.

  Debit Card (41%) > Credit Card (31%) > E-Wallet (11%) > Cash on Delivery (9%) > UPI (7%)

- There is a greater number of male (60%) customers.

  Male (60%) > Female (40%)

- Most of the customers have given service scores of 2, 3 and 4.
- Highest number of accounts have four mapped customers.
- **Highest** number of accounts belong **to super and regular plus segments**. The order of account segments according to decreasing number of customers is as below.

  Regular Plus (37%) > Super (36%) > HNI (15%) > Super Plus (7%) > Regular (5%)

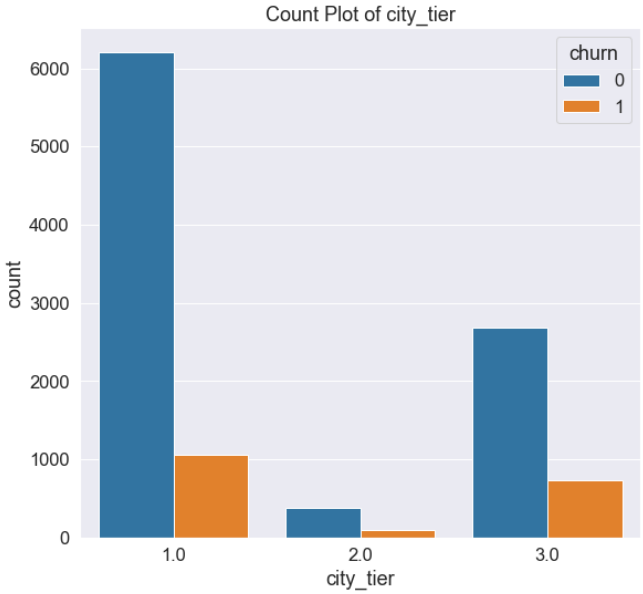
- **Highest** number of customers are **married** and **the lowest** number of customers **are divorced**. The order of marital status according to the decreasing number of customers is as below.

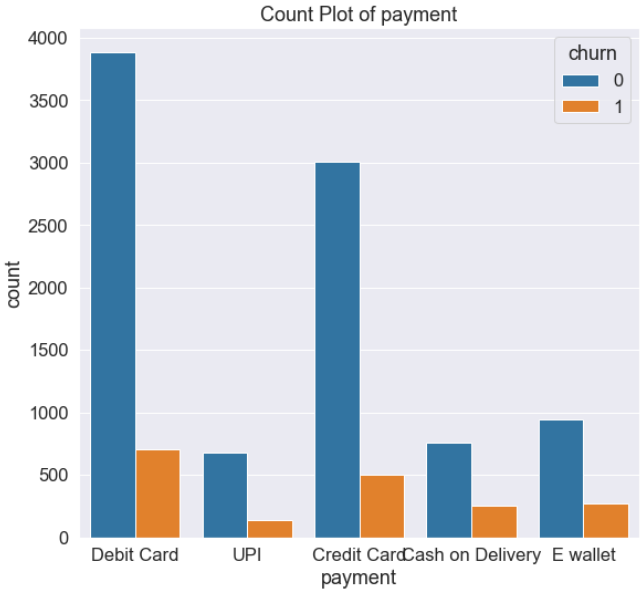
  Married (53%) > Single (32%) > Divorced (15%)

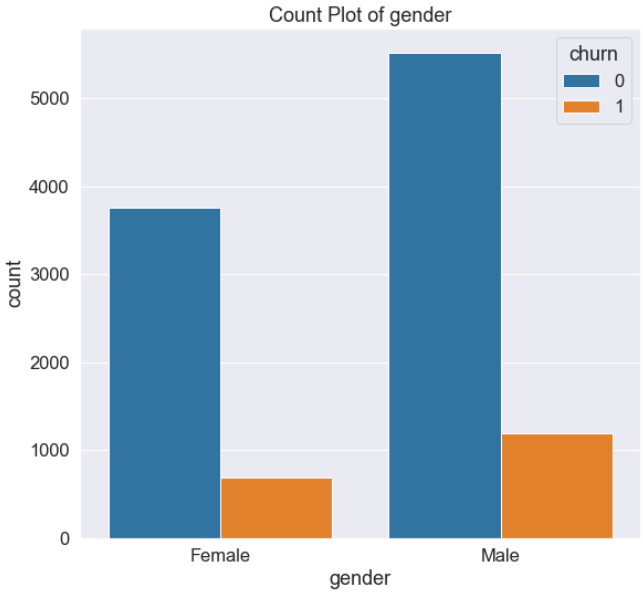
- The number of customers who have **not given complaints (71%)** is **more than** the customers who **gave any complaints (29%)** in the last year.
- The number of customers **using mobile (71%)** is **more than** the number of customers **using computers (29%).**

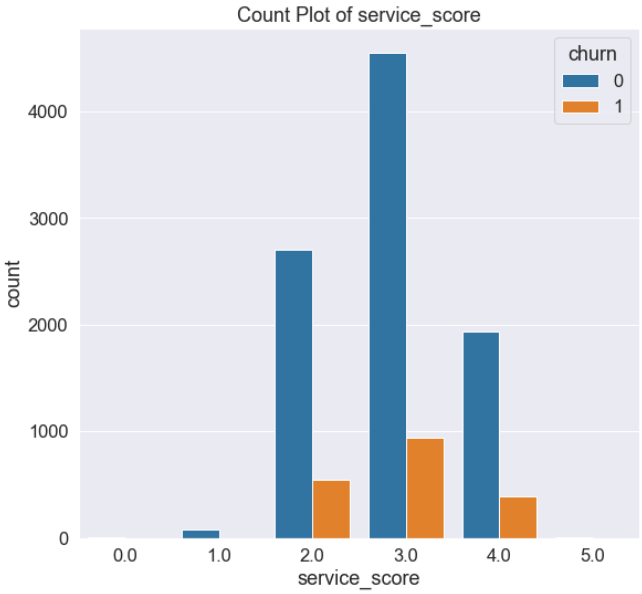
Bivariate Analysis

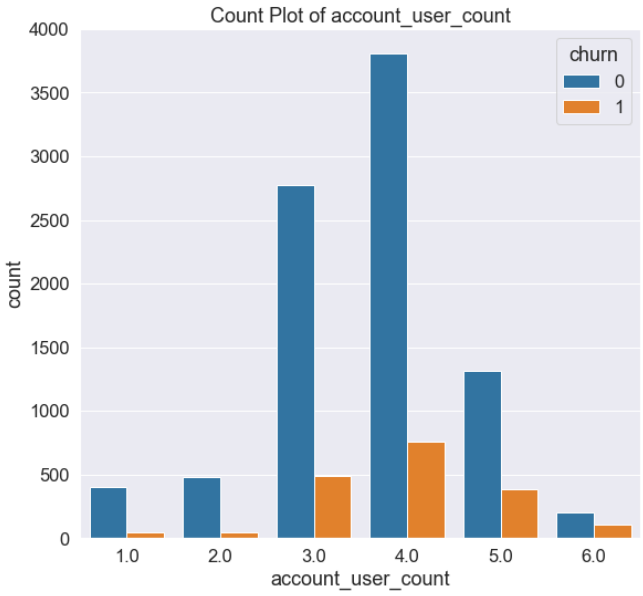
Count Plot of Categorical Features with Churn as Hue

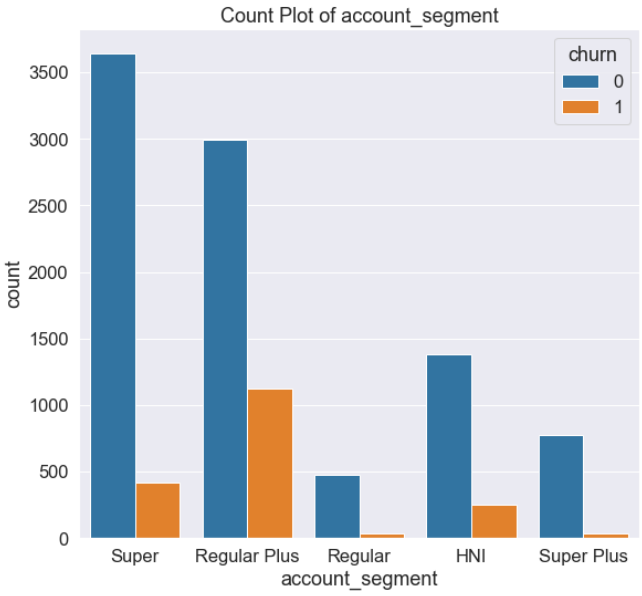












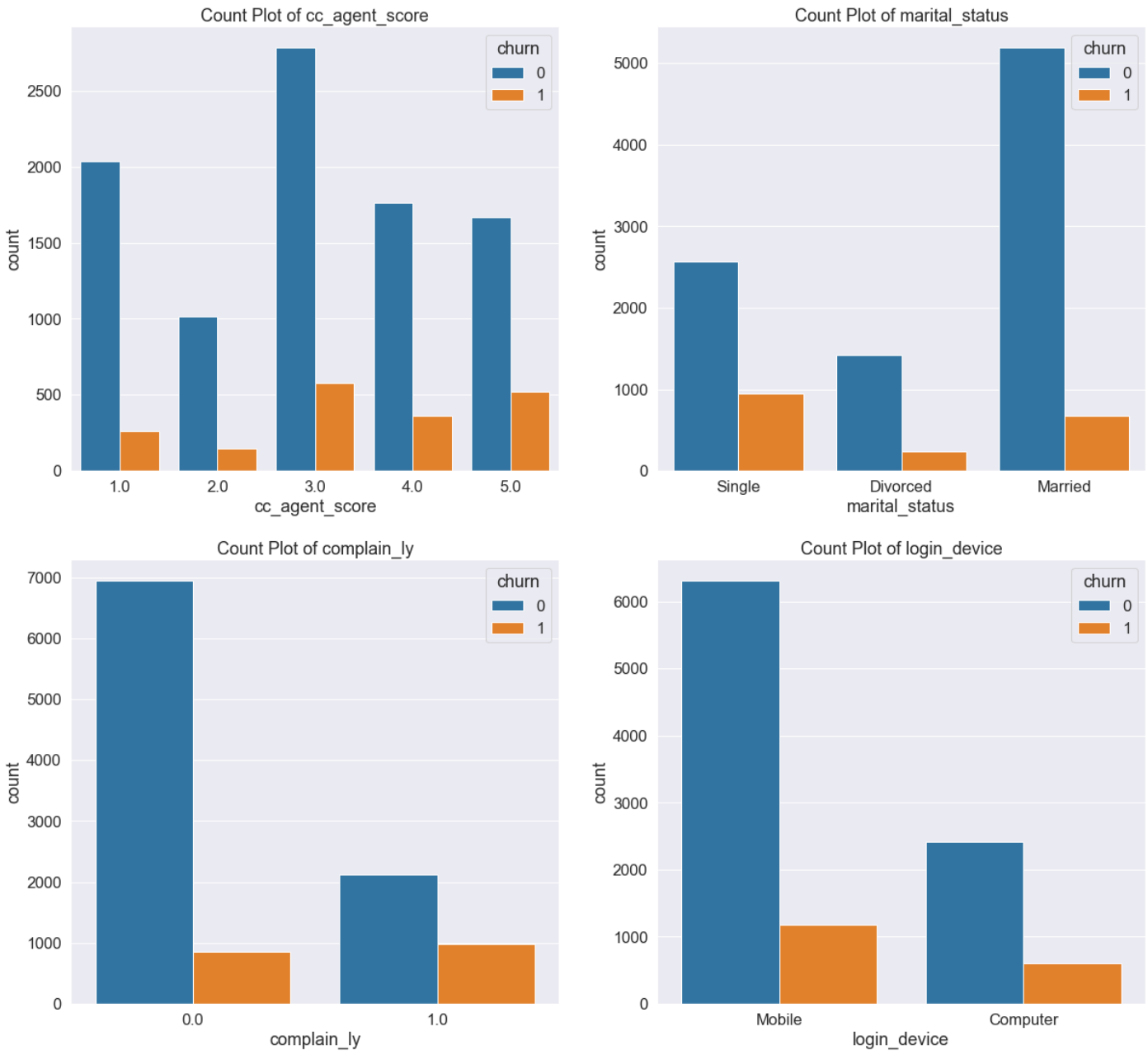


Figure 3. Count Plot of Categorical Features with Churn as Hue.

Insights and Business Implications

- Customers **from tier 3 cities are churning more** than those from other cities. The **decreasing order** of churn rate is as below.

  Tier 3 (21.4%) > Tier 2 (20%) > Tier 1(14.5%)

- Customers paying through **cash on delivery are churning more** than those who prefer other payment methods. The decreasing order of churn rate is as below.

  Cash on delivery (25%) > E-wallet (22.7%) > UPI (17.4%) > Debit card (15.3%) > Credit Card (14.2%)

- Male customers are slightly churning more than female customers.

  Males (17.7%) > Females (15.5%)

- Customers who have given service scores like 2, 3 and 4 are churning. Other customers are not churning.

- Accounts mapped with 5 to 6 customers are churning more than other accounts.
- Customers from **regular plus and HNI segments are churning more** than those from other segments. The decreasing order of churn rate is as below.

Regular Plus (27.3%) > HNI (15.6%) > Super (10.2%) > Regular (7.7%) > Super Plus (4.9%)

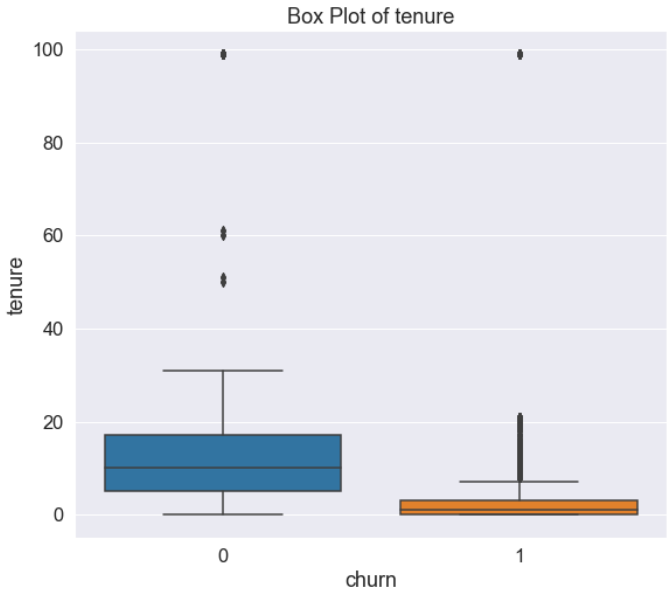
- Customers who have given a high score to customer care agents are churning more than other customers.
- Customers who have **not married are churning more** than other customers. The decreasing order of churn rate is as below.

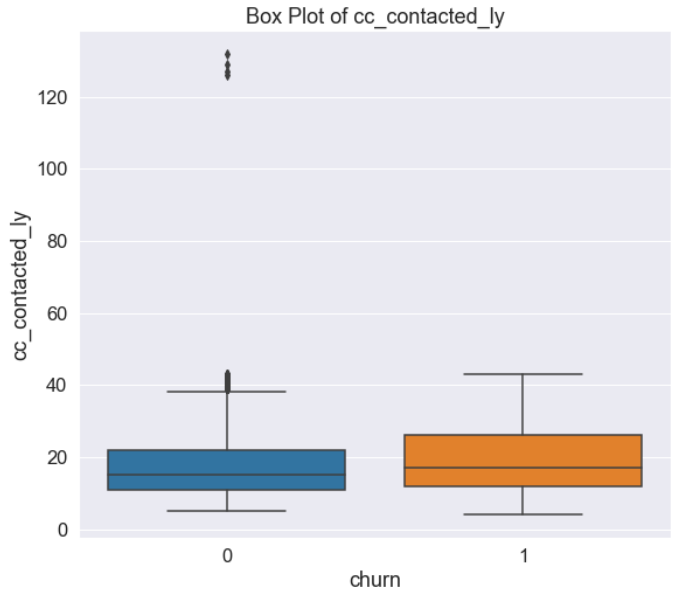
Single (26.9%) > Divorced (14.6%) > Married (11.5%)

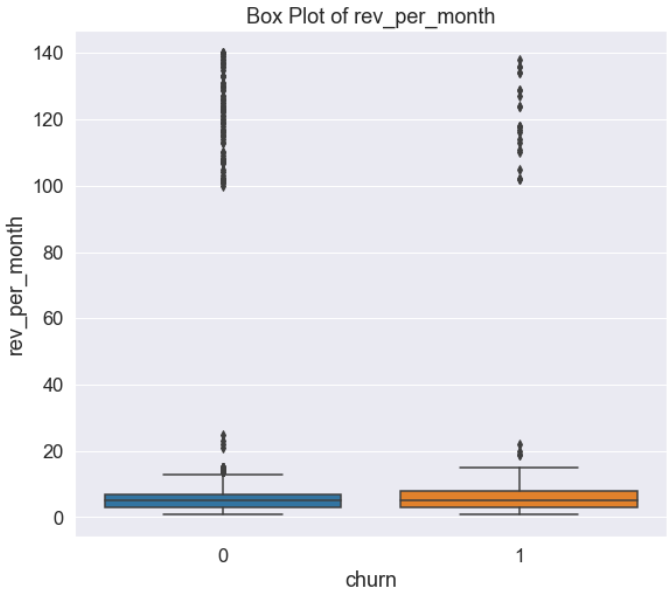
- Customers who have **given complaints (31.8%) are churning more** than those who have not given any complaints (10.9%).
- Churn rate is slightly more in customers who are using a computer (19.8%) compared to mobile (15.7%).

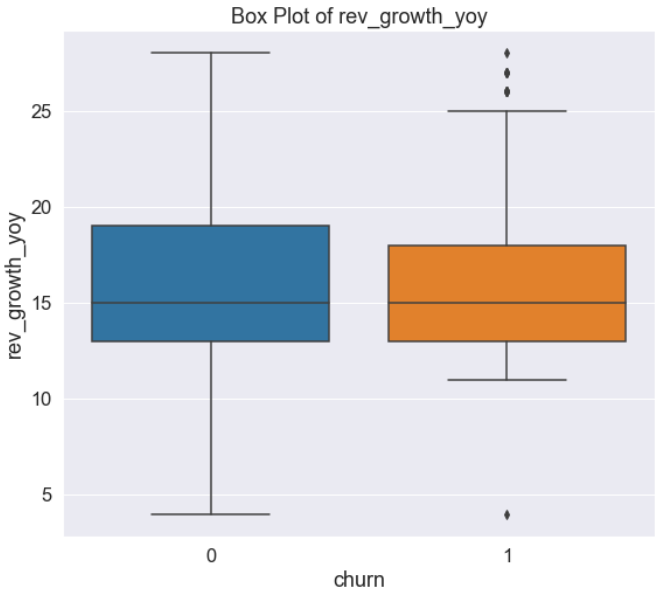
Bivariate Analysis for Numerical Features

Box Plot of Numerical Features with Churn as Hue









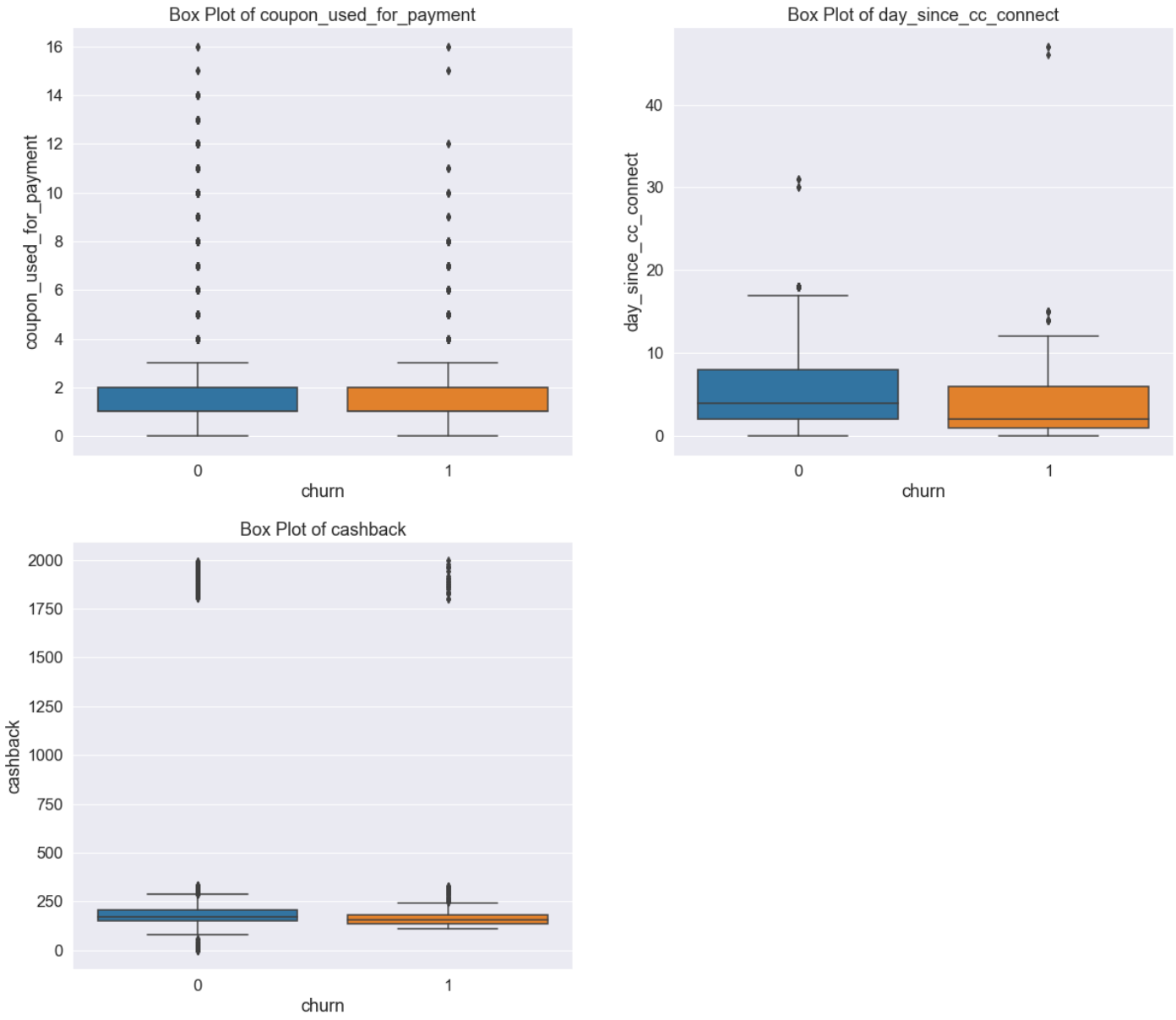


Figure 4. Box Plot of Numerical Features with Churn as Hue.

Median Values of Numerical Features

Target / Feature	Not Churning	Churning
Tenure	10	1
CC Contacted LY	15	17
Revenue per Month	5	5
Revenue growth	15	15
Coupon used for Payment	1	1
Day since cc connect	4	2
Cashback	168.3	152.7

Insights

1. Median tenure of churning customers (1) is very much less than those who are not churning (10).
2. The median number of days since customers contacted is slightly less for churning customers (2) than for non-churning customers (4).
3. The median monthly cashback generated is slightly less for churning customers (152.7) than non-churning customers (168.3).
4. Other numerical features are influencing the target much.

Pair Plot of Numerical Features

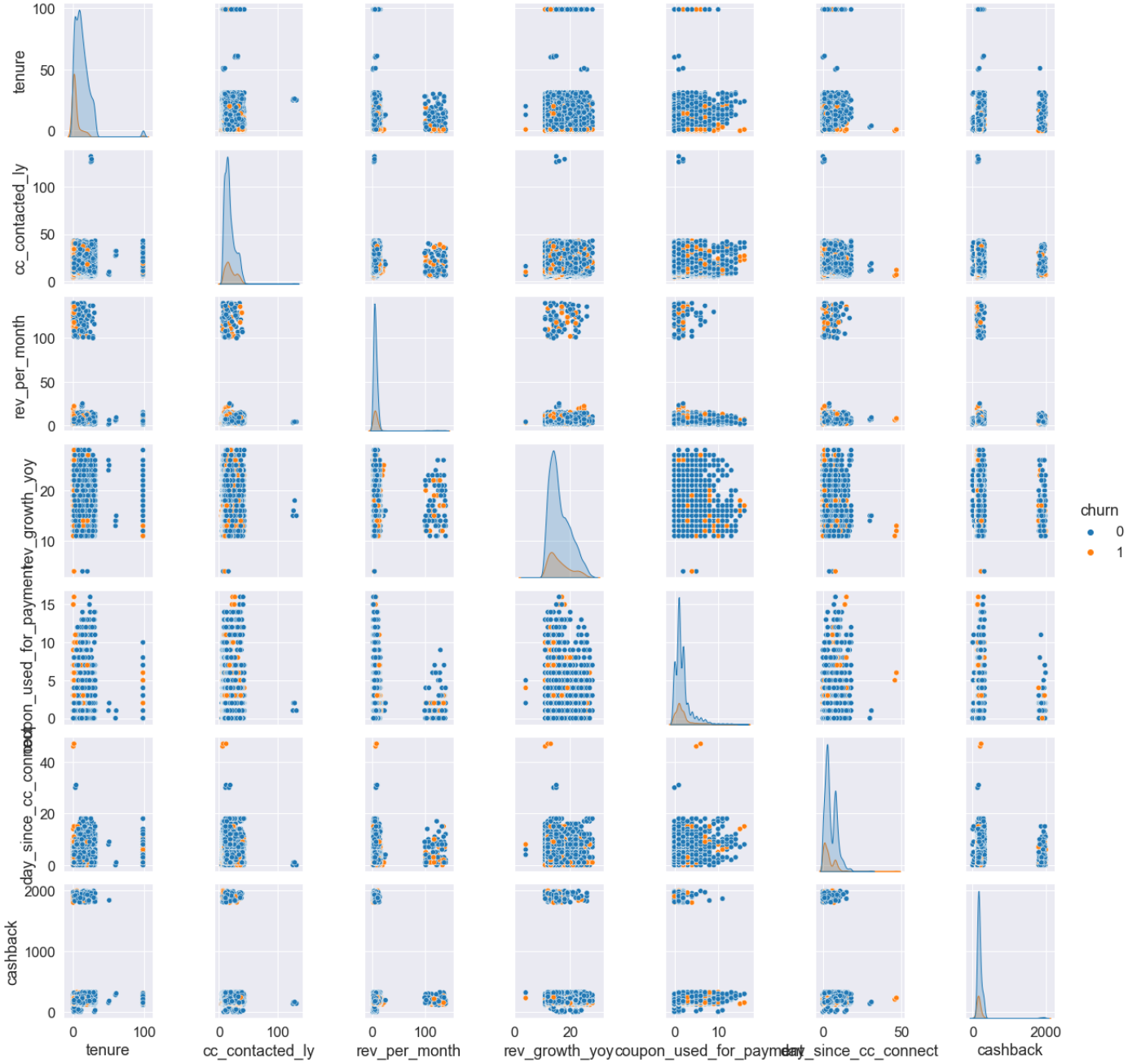


Figure 5. Pair Plot.

Correlation Coefficients

	tenure	cc_contacted_ly	rev_per_month	rev_growth_yoy	coupon_used_for_payment	day_since_cc_connect	cashback
tenure	1.00	-0.00	0.03	0.02	0.09	0.12	0.08
cc_contacted_ly	-0.00	1.00	0.02	0.07	0.00	0.01	0.00
rev_per_month	0.03	0.02	1.00	0.02	0.02	-0.00	0.00
rev_growth_yoy	0.02	0.07	0.02	1.00	0.02	0.00	-0.00
coupon_used_for_payment	0.09	0.00	0.02	0.02	1.00	0.36	0.07
day_since_cc_connect	0.12	0.01	-0.00	0.00	0.36	1.00	0.08
cashback	0.08	0.00	0.00	-0.00	0.07	0.08	1.00

Heat with Correlation Coefficients

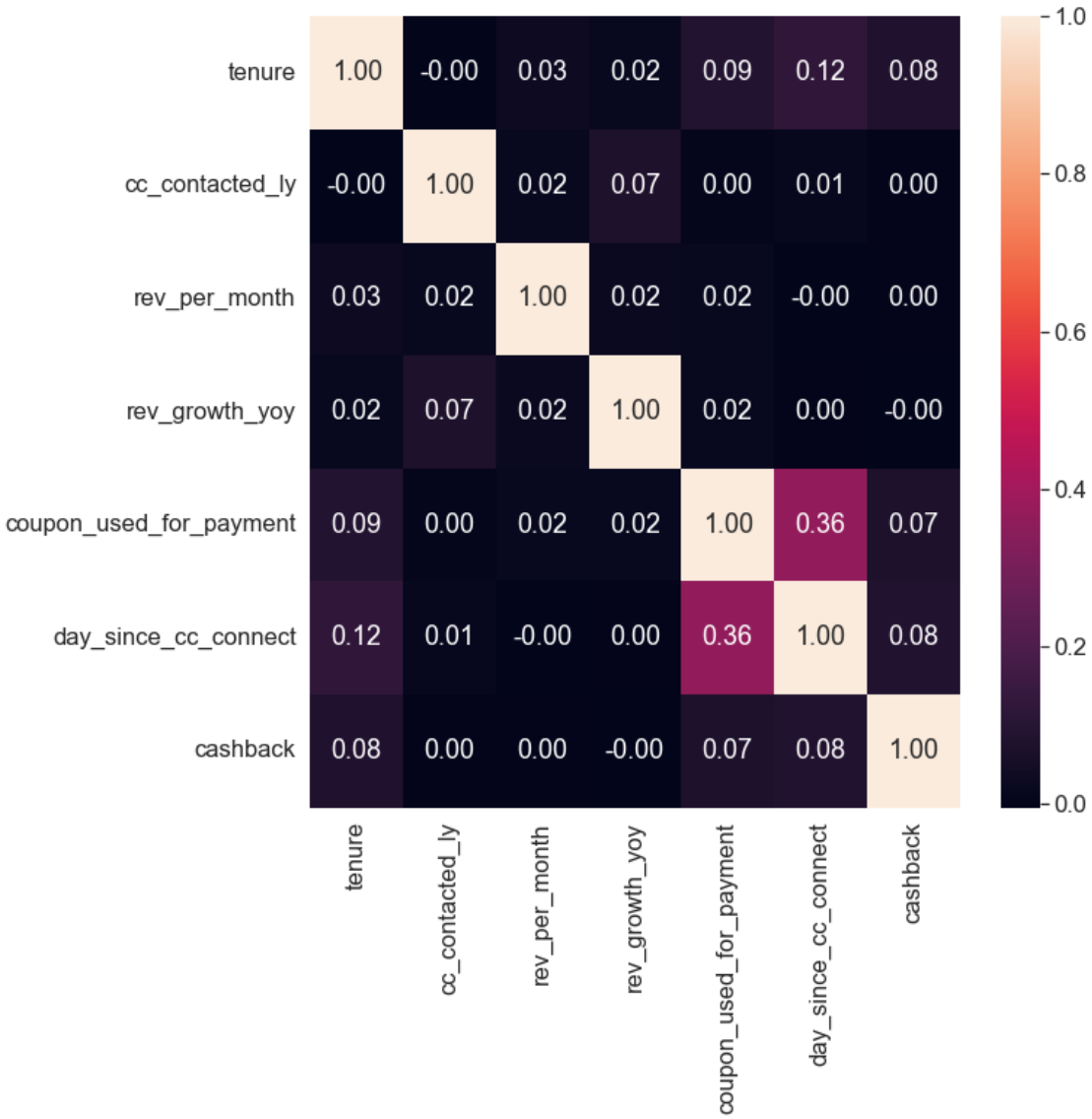


Figure 6. Heat Map with Correlation Coefficients.

Note: There are no significant correlations between the predictor variables.

Multivariate Analysis

Churn across City Tier and Complain in last year

Figure 7. Churn across City Tier and Complain in last year

Churn across Account Segment and Complain in last year

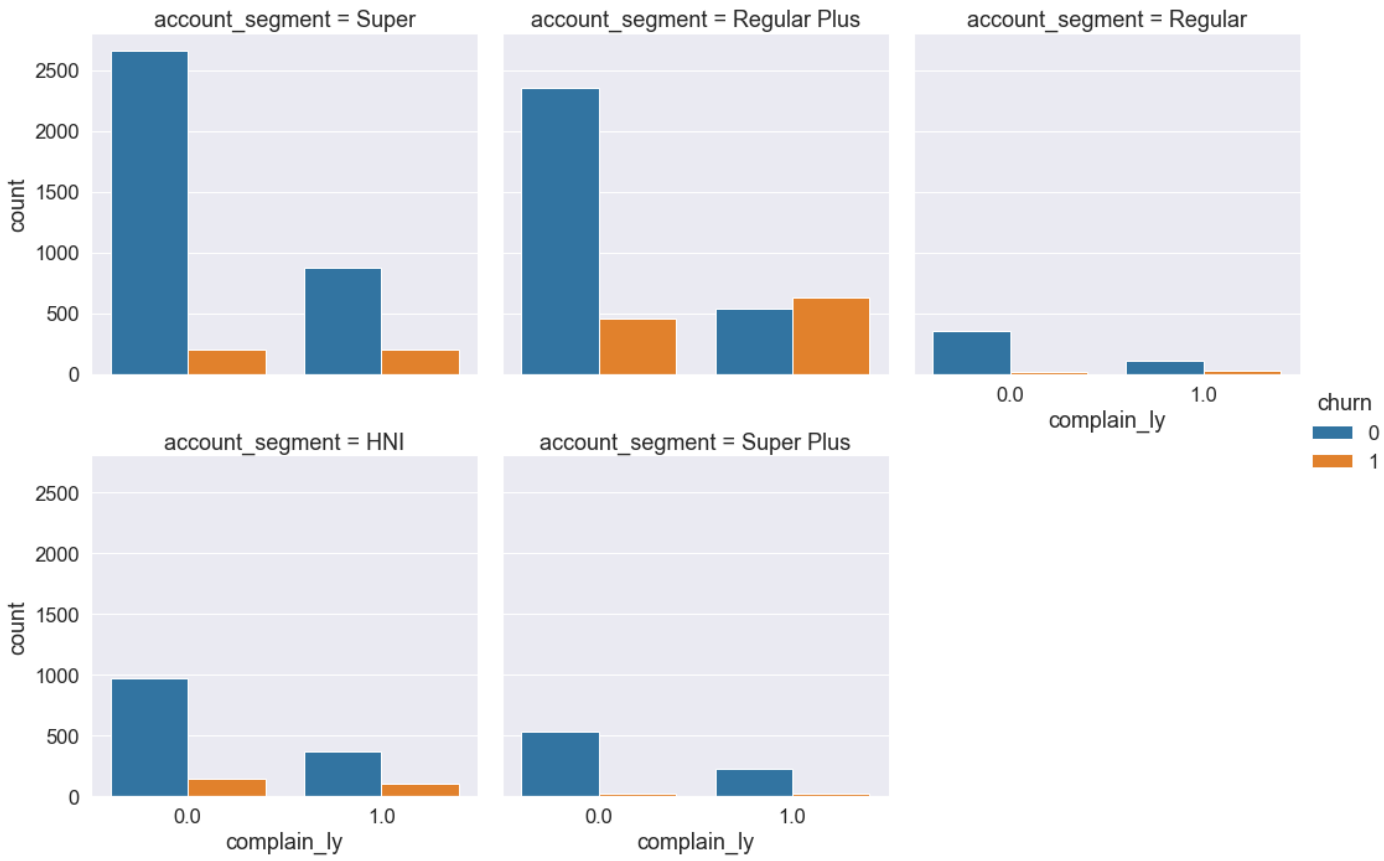


Figure 8. Churn across Account Segment and Complain in last year

Churn across Account Segment and City Tier in last year

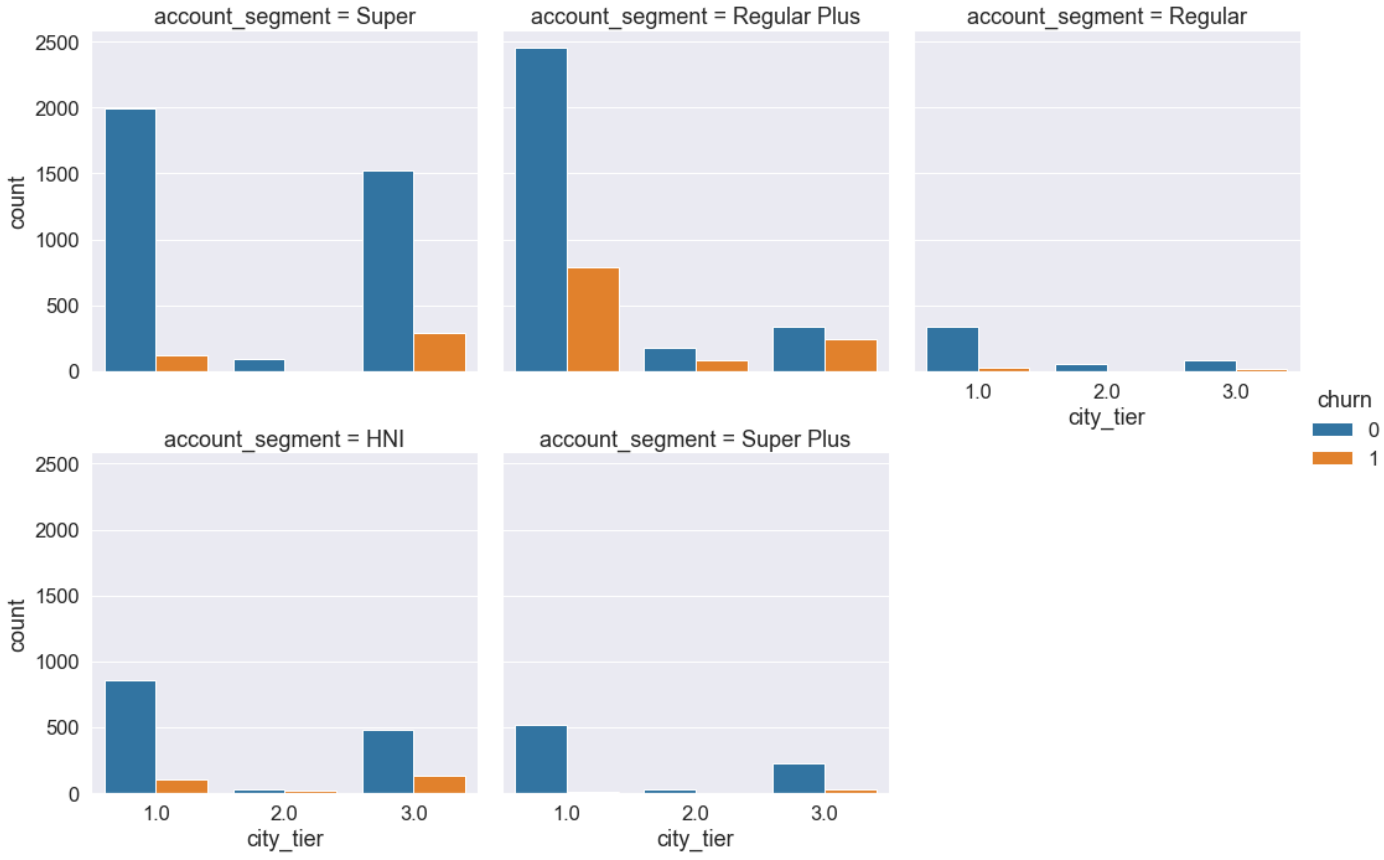


Figure 9. Churn across Account Segment and City Tier in last year

Churn across Tenure and Day since cc connect

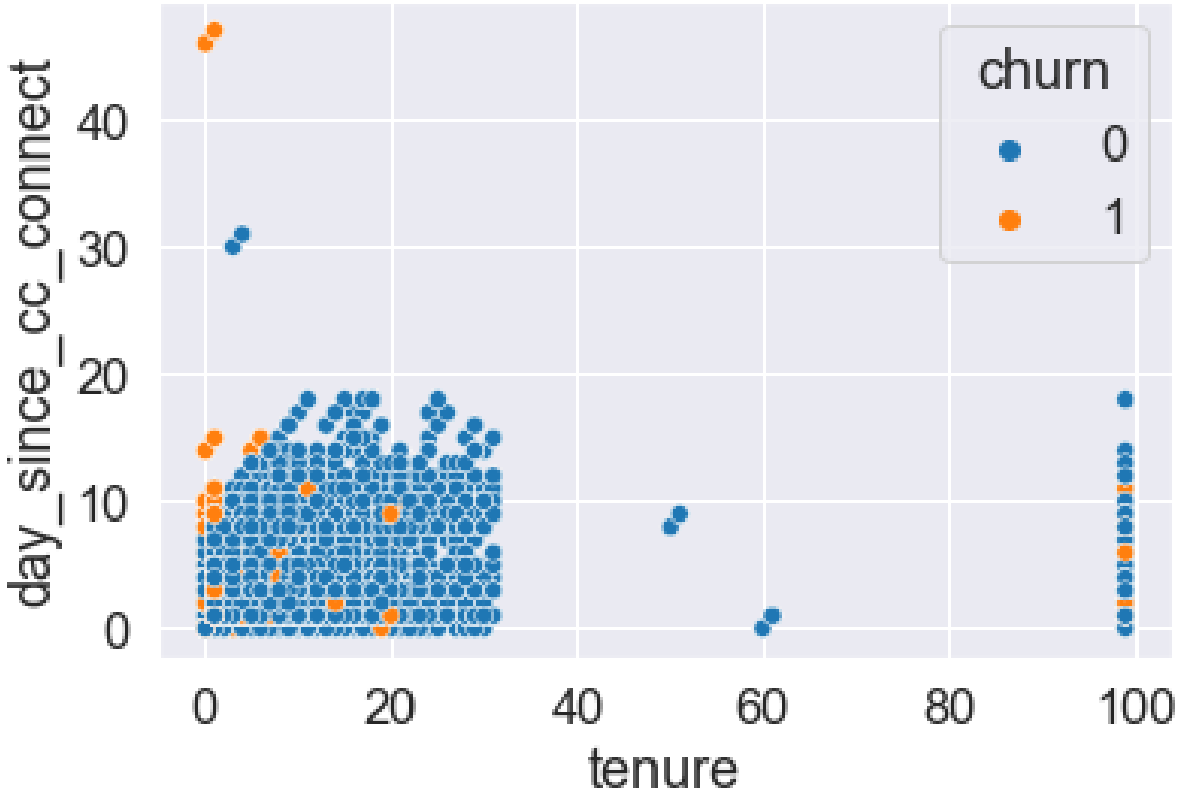


Figure 10. Churn across Day since cc connects and Tenure.

Insights and Business Implications

- Customers who have **given complaints are churning more** than those who have not given any complaints across all tiers of cities.
- Customers who have **given any complaints** are **churning maximum in tier 3 cities (39%)**. Customers who have **not given any complaints** are **churning maximum in tier 2 cities (15%)**.
- Customers who have **given complaints are churning more** than those who have not given any complaints across all segments of customers.
- Customers from the **regular plus segment are churning more** than those from other segments irrespective of whether they have given any complaints in the last year.

The churn rate of Regular Plus customers given any complaint (54%)

The churn rate of Regular Plus customers not given any complaint (16%)

- Customers **from tier 3 cities are churning more** than those from other cities across all segments of customers.
- Customers **from regular plus segments are churning more** than those from other segments across all tiers of cities.
- Customers with low tenure and contacted recently are churning more than those with high tenure and contacted customer care long before.

Q3. Data Cleaning and Pre-Processing

Missing Value Treatment

Features	Number of Null Values	Percentage of Null Values	Features	Number of Null Values	Percentage of Null Values
Rev per month	791.0	7.0	Rev per month	0	0
Login device	760.0	6.7	Login device	0	0
Cashback	473.0	4.2	Cashback	0	0
Account user count	444.0	3.9	Account user count	0	0
Day since cc connect	358.0	3.2	Day since cc connect	0	0
Complain ly	357.0	3.2	Complain ly	0	0
Tenure	218.0	1.9	Tenure	0	0
Marital status	212.0	1.9	Marital status	0	0
cc agent score	116.0	1.0	cc agent score	0	0
City tier	112.0	1.0	City tier	0	0
Payment	109.0	1.0	Payment	0	0
Gender	108.0	1.0	Gender	0	0
cc contacted ly	102.0	0.9	cc contacted ly	0	0
Service score	98.0	0.9	Service score	0	0
Account segment	97.0	0.9	Account segment	0	0
Rev growth yoy	3.0	0.0	Rev growth yoy	0	0
Coupon used for payment	3.0	0.0	Coupon used for payment	0	0

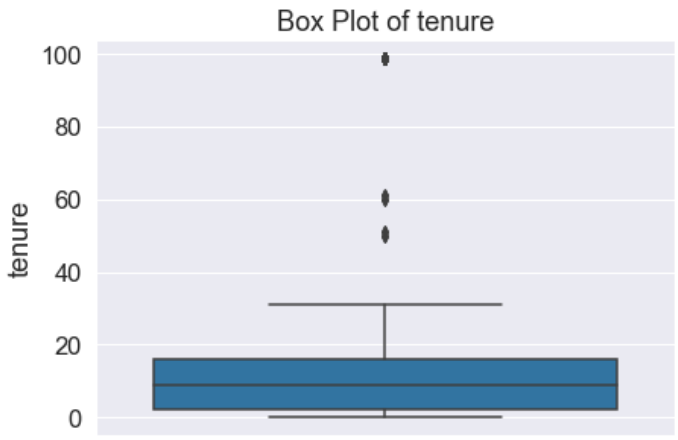
Table 6. Percentage of Null Values in Each Feature before and after Treating.

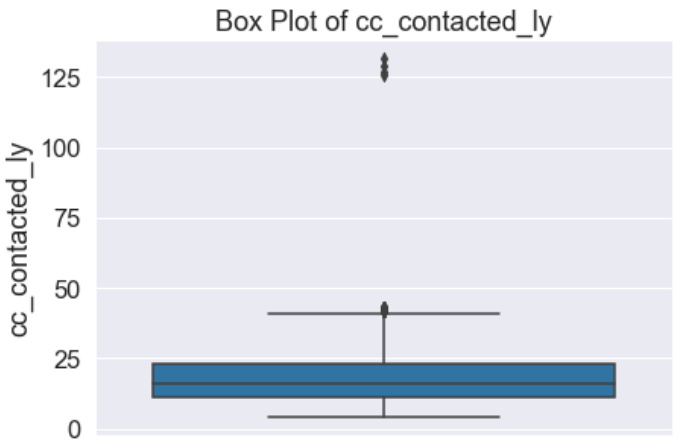
Approach

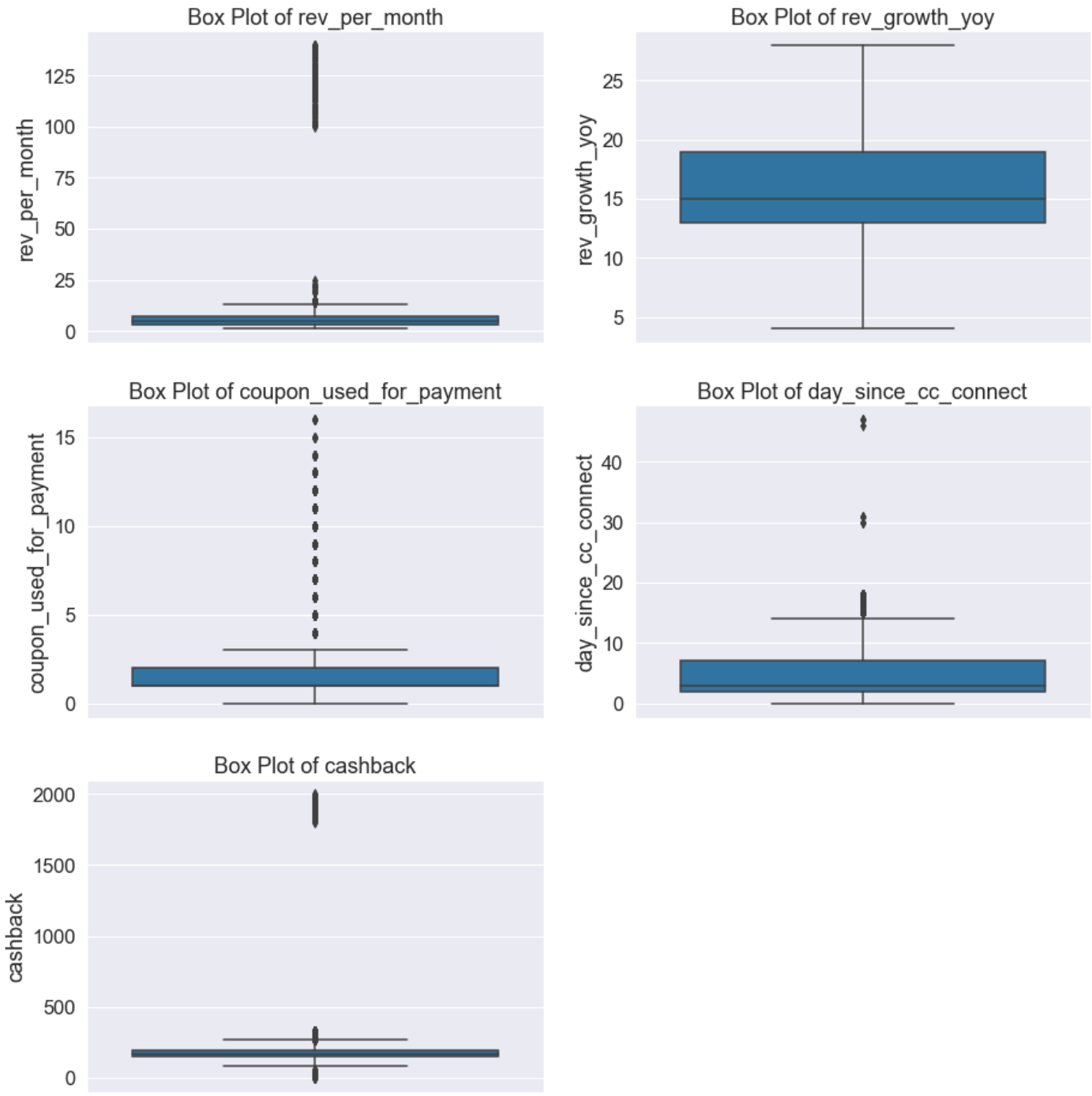
- If any record is **not having any value** for any feature. Then it is identified as **a null value**.
- As the **data has outliers**, the null values in **numerical features are imputed** with their respective **median values**.
- The null values in **categorical features are imputed** with their respective **mode values**.

Outlier Treatment

Let us check outliers visually by plotting box plots for each feature.







As we can see in the above box plots, outliers are present in almost all the features. Let us quantify it.

Percentage of Outliers in Each Feature

Feature	No. of Outliers	Percentage of Outliers	Feature	No. of Outliers	Percentage of Outliers
Coupon used for payment	1380.0	12.3	Coupon used for payment	0	0
Cashback	986.0	8.8	Cashback	0	0
Rev per month	185.0	1.6	Rev per month	0	0
Tenure	139.0	1.2	Tenure	0	0
Day since cc connect	130.0	1.2	Day since cc connect	0	0
cc contacted ly	42.0	0.4	cc contacted ly	0	0
Rev growth yoy	0.0	0.0	Rev growth yoy	0	0

Table 7. Percentage of Outliers in Each Feature before and after Treating.

Approach

- The values which lie **outside the range [Q1-1.5*IQR, Q3+1.5*IQR]** are identified as outliers.
- The outliers in the above dataset are treated by the **capping** (replacing the values above the upper range with Q3+1.5*IQR) and **flooring** (replacing the values below the lower range with Q1-1.5*IQR) technique.

Need for variable transformation (if any)

All the features in the dataset **should have numerical values** before building any machine learning model. Hence, let us transform the categorical variables into numerical ones by following label encoding and one hot encoding techniques.

Encoding of Categorical Data

- There is one ordinal categorical variable i.e., account segment.
- There are four nominal categorical variables i.e., payment, gender, marital status and login device.
- Account segment ordinary variable is label encoded as per below order preference.

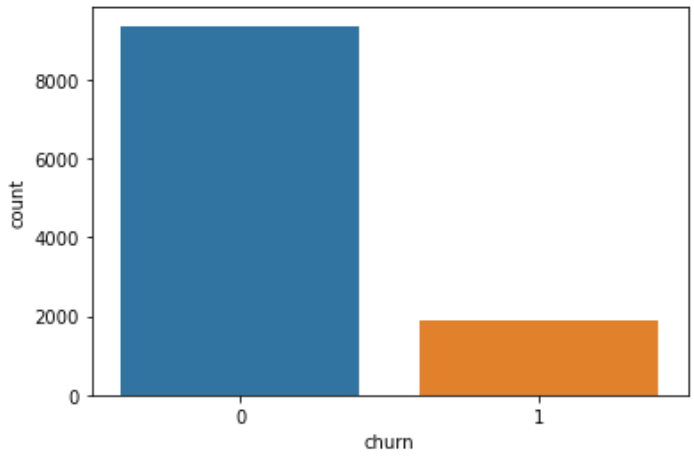
Sublevel in Account Segment	Label encoding
Regular	0
Regular Plus	1
Super	2
Super Plus	3
HNI	4

- Nominal categorical variables are encoded in a different way for **tree-based models (label encoding)** and **distance & weight-based models (one hot encoding – creating dummies).**

Splitting the Dataset into Predictors and Target Sets

- In the given dataset churn is the target variable, let us split the dataset into two parts.
- Except churn feature, all remaining features are taken in the predictor's dataset.
- Churn feature is considered as the target variable.

Distribution of Classes in Target Feature



- The percentage of customers who are not churning is 83.2%
- The percentage of customers who are churning is 16.8%
- The data **is unbalanced** with the classes in the target feature but not so bad. As of now, we can proceed with the model building process.

- Later, if the performance of the model is not up to the mark, we can apply oversampling techniques such as SMOTE for the training dataset to improve the model performance.

Splitting the Dataset into Train and Test Datasets

- Let us split the predictor's dataset and target dataset into train and test sets in the ratio of 70:30.

Table 8. Sample of the Train and Test Datasets for Tree-based Models.

	7580	5198	1929	3427	6249
tenure	11.0	22.0	15.0	15.0	0.0
city_tier	1.0	3.0	1.0	1.0	1.0
cc_contacted_ly	22.0	14.0	14.0	14.0	22.0
payment	Debit Card	E wallet	Debit Card	Credit Card	Credit Card
gender	Male	Male	Male	Female	Female
service_score	0.0	4.0	3.0	3.0	3.0
account_user_count	4.0	4.0	3.0	4.0	5.0
account_segment	2	2	0	2	2
cc_agent_score	5.0	5.0	1.0	3.0	4.0
marital_status	Married	Married	Married	Single	Divorced
rev_per_month	7.0	6.0	2.0	8.0	6.0
complain_ly	0.0	0.0	0.0	0.0	0.0
rev_growth_yoy	15.0	12.0	13.0	15.0	15.0
coupon_used_for_payment	1.0	1.0	3.5	2.0	1.0
day_since_cc_connect	2.0	3.0	9.0	2.0	0.0
cashback	163.88	179.86	271.44	175.03	149.77
login_device	Mobile	Mobile	Computer	Mobile	Mobile

	784	6943	3709	6439	5310
tenure	0.0	0.0	10.0	0.0	9.0
city_tier	1.0	3.0	3.0	1.0	1.0
cc_contacted_ly	31.0	22.0	23.0	10.0	25.0
payment	Debit Card	E wallet	E wallet	Credit Card	Debit Card
gender	Female	Male	Female	Female	Male
service_score	2.0	2.0	4.0	2.0	3.0
account_user_count	1.0	3.0	5.0	3.0	2.0
account_segment	1	1	2	1	2
cc_agent_score	2.0	5.0	5.0	4.0	5.0
marital_status	Single	Married	Divorced	Single	Married
rev_per_month	2.0	5.0	6.0	4.0	3.0
complain_ly	1.0	0.0	0.0	0.0	0.0
rev_growth_yoy	12.0	20.0	14.0	19.0	12.0
coupon_used_for_payment	1.0	1.0	2.0	1.0	2.0
day_since_cc_connect	1.0	0.0	3.0	1.0	11.0
cashback	129.94	119.9	181.47	120.96	177.08
login_device	Mobile	Computer	Mobile	Mobile	Mobile

Necessity of Scaling

- Generally, Scaling improves the performance of **all distance-based models like Linear Discriminant Analysis and KNN** and **weight-based models like Artificial Neural Networks and Logistic Regression**. Even Scaling influences, the coefficients obtained for different features in the logistic regression model. By scaling, units can be avoided in coefficients and standardized coefficients are obtained. Also scaling improves the speed of convergence of the models.
- If we don't scale the data, it gives higher weightage to features which have higher magnitude. Hence, it is always advisable to **bring all the features to the same scale** before proceeding to model building.
- In this dataset, the magnitudes of the statistical parameters like Mean, Standard Deviation, Minimum and Maximum are significantly different for all features (Refer below table). **Hence, scaling is required to bring all the features into a common scale before proceeding to model building.**
- Z-Score method is used to scale the data, i.e., finding the z-score value for each observation in the dataset by using the following formula.

$$Z\ Score = \frac{(x - \mu)}{Sigma}$$

Where, x = Value of the observation

μ = Mean

Sigma = Standard Deviation

- **Scaling is required for Logistic Regression, Linear Discriminant Analysis, Artificial Neural Networks, and KNN models.** The scaled dataset is used for these models.
- For other models like **Naive Bayes, Bagging, Random Forest, Ada Boosting, Gradient Boosting and Extreme Gradient Boosting** models, **scaling is not required**. Hence, the non-scaled dataset is used for these models.

Table 9. Mean and Standard Deviation of All Numeric Features.

Feature/Measure	Tenure	cc contacted ly	Rev per month	Rev growth yoy	Coupon used for payment	Day since cc connect	Cashback
count	11042.0	11158.0	10469.0	11257.0	11257.0	10902.0	10787.0
mean	11.0	17.9	6.4	16.2	1.8	4.6	196.2
std	12.9	8.9	11.9	3.8	2.0	3.7	178.7
min	0.0	4.0	1.0	4.0	0.0	0.0	0.0
25%	2.0	11.0	3.0	13.0	1.0	2.0	147.2
50%	9.0	16.0	5.0	15.0	1.0	3.0	165.3
75%	16.0	23.0	7.0	19.0	2.0	8.0	200.0
max	99.0	132.0	140.0	28.0	16.0	47.0	1997.0

Sample of Scaled Datasets

Predictor variables have been **scaled by using the z-score method**. Initially, the training dataset has been scaled by using its mean and standard deviation. Then test dataset has been **scaled by using train dataset parameters** (mean and standard deviation) **to avoid data leakage**.

Table 10. Samples of Train and Test Datasets after Scaling (for weight-based models).

	7580	5198	1929	3427	6249
tenure	0.085906	1.319059	0.534326	0.534326	-1.147247
city_tier	1.000000	3.000000	1.000000	1.000000	1.000000
cc_contacted_ly	0.497497	-0.441786	-0.441786	-0.441786	0.497497
service_score	0.000000	4.000000	3.000000	3.000000	3.000000
account_user_count	4.000000	4.000000	3.000000	4.000000	5.000000
account_segment	2.000000	2.000000	0.000000	2.000000	2.000000
cc_agent_score	5.000000	5.000000	1.000000	3.000000	4.000000
rev_per_month	0.596296	0.249534	-1.137511	0.943057	0.249534
complain_ly	0.000000	0.000000	0.000000	0.000000	0.000000
rev_growth_yoy	-0.317924	-1.124191	-0.855436	-0.317924	-0.317924
coupon_used_for_payment	-0.434493	-0.434493	1.843459	0.476688	-0.434493
day_since_cc_connect	-0.723805	-0.436752	1.285565	-0.723805	-1.297911
cashback	-0.303254	0.061574	2.152373	-0.048696	-0.625389
payment_Credit_Card	0.000000	0.000000	0.000000	1.000000	1.000000
payment_Debit_Card	1.000000	0.000000	1.000000	0.000000	0.000000
payment_E_wallet	0.000000	1.000000	0.000000	0.000000	0.000000
payment_UPI	0.000000	0.000000	0.000000	0.000000	0.000000
gender_Male	1.000000	1.000000	1.000000	0.000000	0.000000
marital_status_Married	1.000000	1.000000	1.000000	0.000000	0.000000
marital_status_Single	0.000000	0.000000	0.000000	1.000000	0.000000
login_device_Mobile	1.000000	1.000000	0.000000	1.000000	1.000000

	784	6943	3709	6439	5310
tenure	-1.147247	-1.147247	-0.026199	-1.147247	-0.138303
city_tier	1.000000	3.000000	3.000000	1.000000	1.000000
cc_contacted_ly	1.554191	0.497497	0.614908	-0.911428	0.849728
service_score	2.000000	2.000000	4.000000	2.000000	3.000000
account_user_count	1.000000	3.000000	5.000000	3.000000	2.000000
account_segment	1.000000	1.000000	2.000000	1.000000	2.000000
cc_agent_score	2.000000	5.000000	5.000000	4.000000	5.000000
rev_per_month	-1.137511	-0.097227	0.249534	-0.443988	-0.790750
complain_ly	1.000000	0.000000	0.000000	0.000000	0.000000
rev_growth_yoy	-1.124191	1.025854	-0.586680	0.757099	-1.124191
coupon_used_for_payment	-0.434493	-0.434493	0.476688	-0.434493	0.476688
day_since_cc_connect	-1.010858	-1.297911	-0.436752	-1.010858	1.859671
cashback	-1.078114	-1.307330	0.098331	-1.283130	-0.001894
payment_Credit_Card	0.000000	0.000000	0.000000	1.000000	0.000000
payment_Debit_Card	1.000000	0.000000	0.000000	0.000000	1.000000
payment_E_wallet	0.000000	1.000000	1.000000	0.000000	0.000000
payment_UPI	0.000000	0.000000	0.000000	0.000000	0.000000
gender_Male	0.000000	1.000000	0.000000	0.000000	1.000000
marital_status_Married	0.000000	1.000000	0.000000	0.000000	1.000000
marital_status_Single	1.000000	0.000000	0.000000	1.000000	0.000000
login_device_Mobile	1.000000	0.000000	1.000000	1.000000	1.000000

Distribution of Target Variable Classes in Train and Test Datasets

Class	Total Data	Training Data	Testing Data
Not Churn (0)	83.16%	83.18%	83.13%
Churn (1)	16.84%	16.82%	16.87%

Note:

- Churning customers are almost equally distributed in both training and testing datasets.

Variables removed or added and why (if any)

Chi-Square Test

- We have removed Account ID which is not required for the model building process.
- Let us **identify other insignificant variables** by performing **a chi-square test** for each categorical variable with the target variable.

Table 11. P-Values in chi-square test for each categorical variable.

Feature	P Values of Chi-Square Test
City tier	0
Payment	0
Gender	0.002502
Service score	0.002469
Account user count	0
Account segment	0
cc agent score	0
Marital status	0
Complain ly	0
Login device	0

As **p-value is less than 0.05 for all categorical features** in the chi-square test. Hence, all the categorical features are influencing the target variable **and they are important for model building**.

Multicollinearity

There is no significant correlation between the predictor variables. Anyway, let us check the multicollinearity by checking the Variance Inflation Factor values for each numerical feature.

Feature	VIF
Cashback	1.4
Day since cc connect	1.3
Tenure	1.3
Coupons used for payment	1.2
Rev per month	1.1

- From the VIF table, it can be noticed that all features are having VIF of less than 5. Hence, there is **no significant multicollinearity between the variables**.
- Hence, it is **not required to drop any variable based on VIF** values.

Let us build the Logistic Regression model in the stats library to check the significance of the predictor variable in predicting the target

The Regression model is built in the stats library by using a scaled dataset. The summary of the model is shown in the below table.

Dep. Variable:	churn	No. Observations:	7882
Model:	Logit	Df Residuals:	7860
Method:	MLE	Df Model:	21
Date:	Sat, 04 Jun 2022	Pseudo R-squ.:	0.3279
Time:	13:30:09	Log-Likelihood:	-2400.3
converged:	True	LL-Null:	-3571.1
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.9963	0.309	-12.931	0.000	-4.602	-3.391
tenure	-1.5404	0.067	-22.890	0.000	-1.672	-1.409
city_tier	0.2113	0.046	4.560	0.000	0.120	0.302
cc_contacted_ly	0.2285	0.037	6.143	0.000	0.156	0.301
service_score	-0.1205	0.057	-2.099	0.036	-0.233	-0.008
account_user_count	0.3198	0.041	7.813	0.000	0.240	0.400
account_segment	-0.1890	0.048	-3.927	0.000	-0.283	-0.095
cc_agent_score	0.2819	0.028	10.179	0.000	0.228	0.336
rev_per_month	0.3803	0.037	10.154	0.000	0.307	0.454
complain_ly	1.6788	0.078	21.505	0.000	1.526	1.832
rev_growth_yoy	-0.0700	0.038	-1.835	0.066	-0.145	0.005
coupon_used_for_payment	0.1829	0.044	4.139	0.000	0.096	0.269
day_since_cc_connect	-0.3246	0.049	-6.681	0.000	-0.420	-0.229
cashback	0.0326	0.062	0.529	0.597	-0.088	0.153
payment_Credit_Card	-0.8945	0.131	-6.851	0.000	-1.150	-0.639
payment_Debit_Card	-0.6955	0.124	-5.590	0.000	-0.939	-0.452
payment_E_wallet	-0.2313	0.164	-1.412	0.158	-0.553	0.090
payment_UPI	-0.8360	0.172	-4.865	0.000	-1.173	-0.499
gender_Male	0.2808	0.077	3.638	0.000	0.130	0.432
marital_status_Married	-0.3099	0.113	-2.752	0.006	-0.531	-0.089
marital_status_Single	0.6811	0.114	5.988	0.000	0.458	0.904
login_device_Mobile	-0.4136	0.081	-5.086	0.000	-0.573	-0.254

Note:

- Cashback, Revenue growth YOY and Payment through E-Wallet features have a p-value of more than 0.05. Hence, they are not significant.
- Anyway, these features are not dropped at this stage. Because these features may be important to predict the target in other models.
- Hence, let us review these features again after building other models also.

Q4. Model building

Selection of Models

- As we need to **predict the churn of the customers**, it is a **classification type supervised learning** problem. Let us build all classifier type machine learning models to predict the churn of the customers.
- Initially, **let us start with basic models** like Decision Trees, Artificial Neural Networks, and Logistic Regression with default hyperparameters **then we will build ensemble models** like Bagging, Random Forest, Ada Boosting, Gradient Boosting, and Extreme Gradient Boosting models with default hyperparameters.
- Various models are built with their respective default hyperparameters shown in the below table.

Table 12. Default Hyperparameters of All Models.

Model	Default Hyperparameter
Decision Tree	Criterion = gini, Maximum depth = None
Random Forest	Number Estimators = 100, Criterion = gini, Maximum Depth = None, Maximum Features = auto, Random State = 1
Artificial Neural Networks	Number of Hidden Layers = 1, Number of Neurons = 100, Activation = relu, Solver = adam.
Logistic Regression	Penalty = 'l2', Tolerance = 0.0001, Random State = 1, Solver = lbfgs.
Linear Discriminant Analysis (LDA)	Solver = SVD, Tolerance = 0.0001
k-nearest neighbours' (KNN)	Number of neighbours = 5, Weights = uniform, Metric = Makowski
Bagging	Base Estimator = None, Number of Estimators = 10
AdaBoosting	Base Estimator = None, Number of Estimators = 50,
Gradient Boosting	Number Estimators = 100, Maximum depth = 3, Random state = 1, Tolerance = 0.0001
Extreme Gradient Boosting	Number Estimators = 100, Maximum depth = 6, Random state = 1, Tolerance = 0.0001

Evaluation of Models

- After building various models with default hyperparameters, their **performance is evaluated** for both train and test datasets based on **True Positives, True Negatives, False Positives and False Negatives**.
- Performance metrics like **Accuracy, Recall, Precision and F1 score** are calculated for all the models to evaluate them.

Performance Metrics of Models with Default Hyperparameters

Model	True Positives		True Negatives		False Positives		False Negatives	
	Train	Test	Train	Test	Train	Test	Train	Test
Decision Tree	1326	476	6556	2732	0	76	0	94
Random Forest	1326	478	6556	2797	0	11	0	92
Artificial Neural Networks	1203	453	6505	2748	51	60	123	117
Logistic Regression	600	258	6347	2719	209	89	726	312
Linear Discriminant Analysis (LDA)	561	250	6349	2715	207	93	765	320

Model	True Positives		True Negatives		False Positives		False Negatives	
	Train	Test	Train	Test	Train	Test	Train	Test
Naïve Bayes	731	301	6090	2598	466	210	595	269
KNN	1207	473	6515	2772	41	36	119	97
Bagging	1309	457	6553	2779	3	29	17	113
AdaBoosting	754	322	6314	2709	242	99	572	248
Gradient Boosting	844	345	6385	2739	171	69	482	225
Extreme Gradient Boosting	1323	482	6556	2779	0	29	3	88

Table 13. Performance Metrics of Models with Default Hyperparameters.

Metric/Model	Accuracy		ROC-AUC		Precision		Recall		F1 Score	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Decision Tree	1	0.95	1	0.9	1	0.86	1	0.84	1	0.85
Random Forest	1	0.97	1	0.99	1	0.98	1	0.84	1	0.9
ANN	0.98	0.95	0.99	0.97	0.96	0.88	0.91	0.79	0.93	0.84
Logistic Regression	0.88	0.88	0.87	0.86	0.74	0.74	0.45	0.45	0.56	0.56
LDA	0.88	0.88	0.86	0.85	0.73	0.73	0.42	0.44	0.54	0.55
Naive Bayes	0.87	0.86	0.82	0.81	0.61	0.59	0.55	0.53	0.58	0.56
KNN	0.98	0.96	1	0.98	0.97	0.93	0.91	0.83	0.94	0.88
Bagging	1	0.96	1	0.98	1	0.94	0.99	0.8	0.99	0.87
AdaBoosting	0.9	0.9	0.91	0.91	0.76	0.76	0.57	0.56	0.65	0.65
Gradient Boosting	0.92	0.91	0.95	0.93	0.83	0.83	0.64	0.61	0.72	0.7
Extreme Gradient Boosting	1	0.97	1	0.99	1	0.94	1	0.85	1	0.89

- From the above metrics table, we can notice that a few models **like Decision Tree, Random Forest, ANN, Bagging and Extreme Gradient Boosting models are overfitted slightly.** Let us **tune the hyperparameters** of these models to eliminate overfitting.
- Apart from that **class level metrics like recall, precision and F1 score are low**.
- Hyperparameters of all models are tuned **to eliminate overfitting and to improve metrics**.

Efforts to improve model performance

- The models discussed in the above section are **tuned with optimum hyperparameters by using GridsearchCV** to eliminate overfitting and to improve the performance of the model.
- The tuned hyperparameters for each model are listed in the below table.
- **The models are rebuilt by using below-tuned hyperparameters**.

Table 14. Tuned Hyperparameters of All Models.

Model	Tuned Hyperparameter
Decision Tree	Maximum depth = 14
Random Forest	Number of Estimators: 201, Maximum Features: 9 Maximum Depth: 16
Artificial Neural Networks	Number of Neurons: 350
Logistic Regression	Solver: newton-CG
Linear Discriminant Analysis (LDA)	Solver: svd
Bagging	Number of Estimators: 25
AdaBoosting	Number of Estimators: 151
Gradient Boosting	Number of Estimators: 351 Maximum Depth: 10
Extreme Gradient Boosting	Maximum Depth: 9 Number of Estimators: 201

Q5. Model validation

Performance Metrics of Models with Tuned Hyperparameters

Model	True Positives		True Negatives		False Positives		False Negatives	
	Train	Test	Train	Test	Train	Test	Train	Test
Decision Tree	1292	467	6543	2726	13	82	34	103
Random Forest	1325	487	6556	2791	0	17	1	83
Artificial Neural Networks	1321	516	6555	2772	1	36	5	54
Logistic Regression	600	258	6346	2719	210	89	726	312
Linear Discriminant Analysis (LDA)	561	250	6349	2715	207	93	765	320
Naïve Bayes	731	301	6090	2598	466	210	595	269
KNN	1207	473	6515	2772	41	36	119	97
Bagging	1326	489	6555	2777	1	31	0	81
AdaBoosting	776	332	6317	2697	239	111	550	238
Gradient Boosting	1326	512	6556	2795	0	13	0	58
Extreme Gradient Boosting	1326	499	6556	2785	0	23	0	71

Table 15. Performance Metrics of Models with Tuned Hyperparameters.

Metric/Model	Accuracy		ROC-AUC		Precision		Recall		F1 Score	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Decision Tree	0.99	0.95	1	0.91	0.99	0.85	0.97	0.82	0.98	0.83
Random Forest	1	0.97	1	0.99	1	0.97	1	0.85	1	0.91
ANN	1	0.97	1	0.99	1	0.93	1	0.91	1	0.92
Logistic Regression	0.88	0.88	0.87	0.86	0.74	0.74	0.45	0.45	0.56	0.56
LDA	0.88	0.88	0.86	0.85	0.73	0.73	0.42	0.44	0.54	0.55
Naive Bayes	0.87	0.86	0.82	0.81	0.61	0.59	0.55	0.53	0.58	0.56
KNN	0.98	0.96	1	0.98	0.97	0.93	0.91	0.83	0.94	0.88
Bagging	1	0.97	1	0.98	1	0.94	1	0.86	1	0.9
AdaBoosting	0.9	0.9	0.92	0.91	0.76	0.75	0.59	0.58	0.66	0.66
Gradient Boosting	1	0.98	1	0.99	1	0.98	1	0.9	1	0.94
Extreme Gradient Boosting	1	0.97	1	0.99	1	0.96	1	0.88	1	0.91

- From the above table, we can notice that **the accuracy of the test dataset is slightly less than the accuracy of the training dataset for almost all the models**. Hence, we can conclude that **overfitting is reduced for all models except for the Decision Tree model**.
- Let us compare the performance metrics of all the models obtained for the test dataset to find out the most optimized model (Best Model).

Comparison of Models and Selecting the Best Model

- As there is a **class imbalance in the target variable**, it is **not recommended to rely on only accuracy**. We should **refer to class-level metrics like recall, precision and F1 score**.
- In this project, the **best model** is selected based on the **recall metric**. The **best model** should have **a high recall**. It means the model should **predict almost all the customers who are about churn as churning** customers. Hence, the company can **try to retain those customers** by providing customer-specific offers.

Table 16. Performance Metrics of all Models with Tuned Hyperparameters for the Test Dataset.

Model/Metric	Accuracy	ROC-AUC	Precision	Recall	F1 Score
Decision Tree	0.95	0.91	0.85	0.82	0.83
Random Forest	0.97	0.99	0.97	0.85	0.91
ANN	0.97	0.99	0.93	0.91	0.92
Logistic Regression	0.88	0.86	0.74	0.45	0.56
LDA	0.88	0.85	0.73	0.44	0.55
Naive Bayes	0.86	0.81	0.59	0.53	0.56
KNN	0.96	0.98	0.93	0.83	0.88
Bagging	0.97	0.98	0.94	0.86	0.9
AdaBoosting	0.9	0.91	0.75	0.58	0.66
Gradient Boosting	0.98	0.99	0.98	0.9	0.94
Extreme Gradient Boosting	0.97	0.99	0.96	0.88	0.91

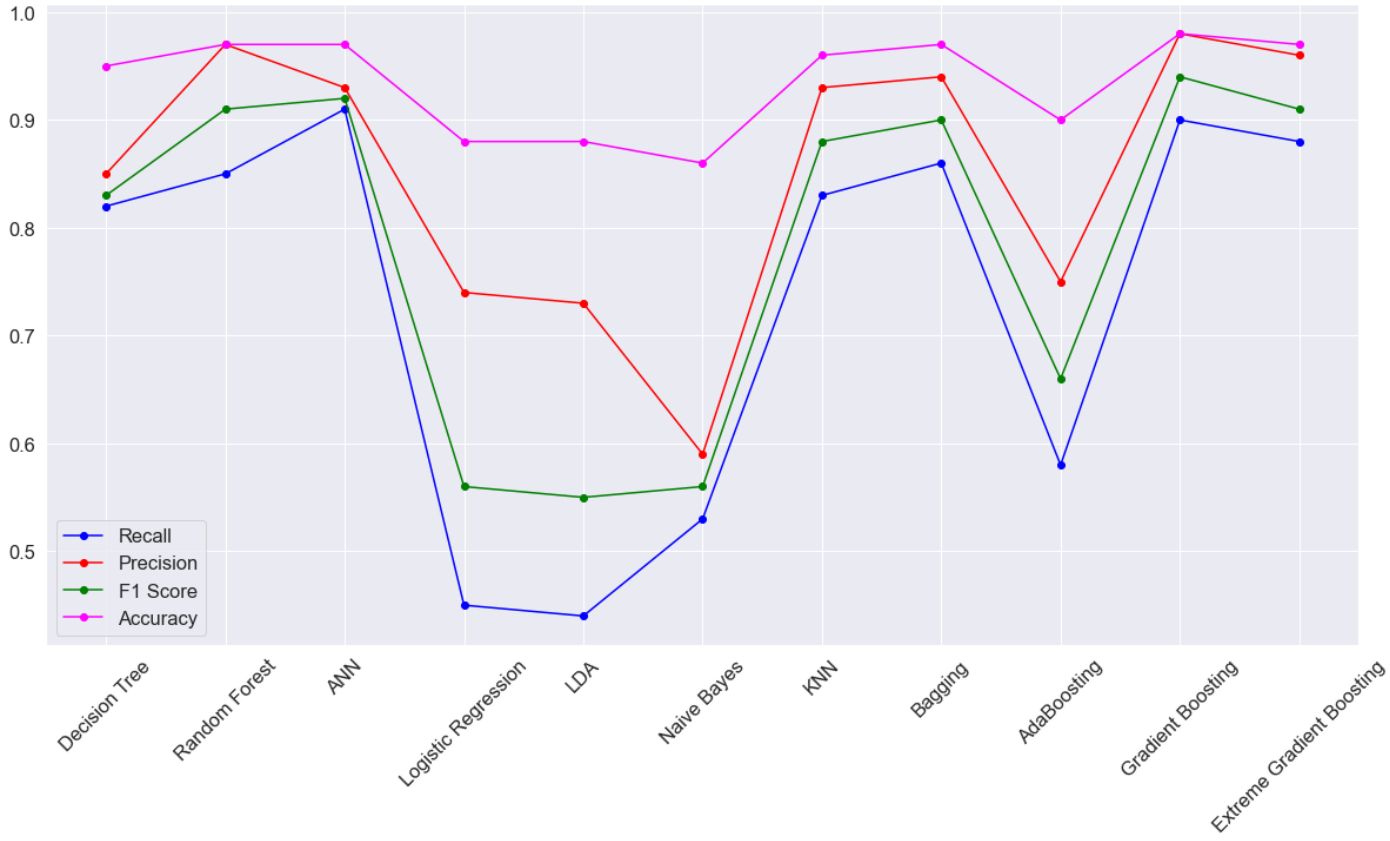


Figure 11. Performance Metrics of all Models with Tuned Hyperparameters for the Test Dataset.

- From the above plot, we can notice that **the recall is less than the precision for most of the models**. But in churn prediction problems**, recall (having low false negatives) for the minority class is more important than precision (having low false positives),** because retaining customers by providing customer-specific offers involves less cost than attracting new customers.
- Hence, it is **very important to predict all the customers who are on the verge of churning accurately**. In this process, even if a few customers who are not likely to churn are predicted as churned is acceptable.
- As we have more gap between the precision and recall for most of the models, let us **try to improve the recall at the expense of precision without affecting the F1 Score much by changing the threshold value to 0.4 instead of 0.5** (default threshold).

Table 17. Performance Metrics of all Models with a threshold of 0.4 for the Test Dataset.

Model/Metric	Accuracy	ROC-AUC	Precision	Recall	F1 Score
Decision Tree	0.95	0.91	0.85	0.82	0.83
Random Forest	0.97	0.99	0.92	0.9	0.91
ANN	0.97	0.99	0.92	0.92	0.92
Logistic Regression	0.88	0.86	0.66	0.55	0.6
LDA	0.88	0.85	0.63	0.52	0.57
Naive Bayes	0.86	0.81	0.51	0.62	0.56
KNN	0.96	0.98	0.93	0.83	0.88
Bagging	0.97	0.98	0.9	0.89	0.89
AdaBoosting	0.9	0.91	0.17	1	0.29

Model/Metric	Accuracy	ROC-AUC	Precision	Recall	F1 Score
Gradient Boosting	0.98	0.99	0.97	0.9	0.94
Extreme Gradient Boosting	0.98	0.99	0.97	0.92	0.94

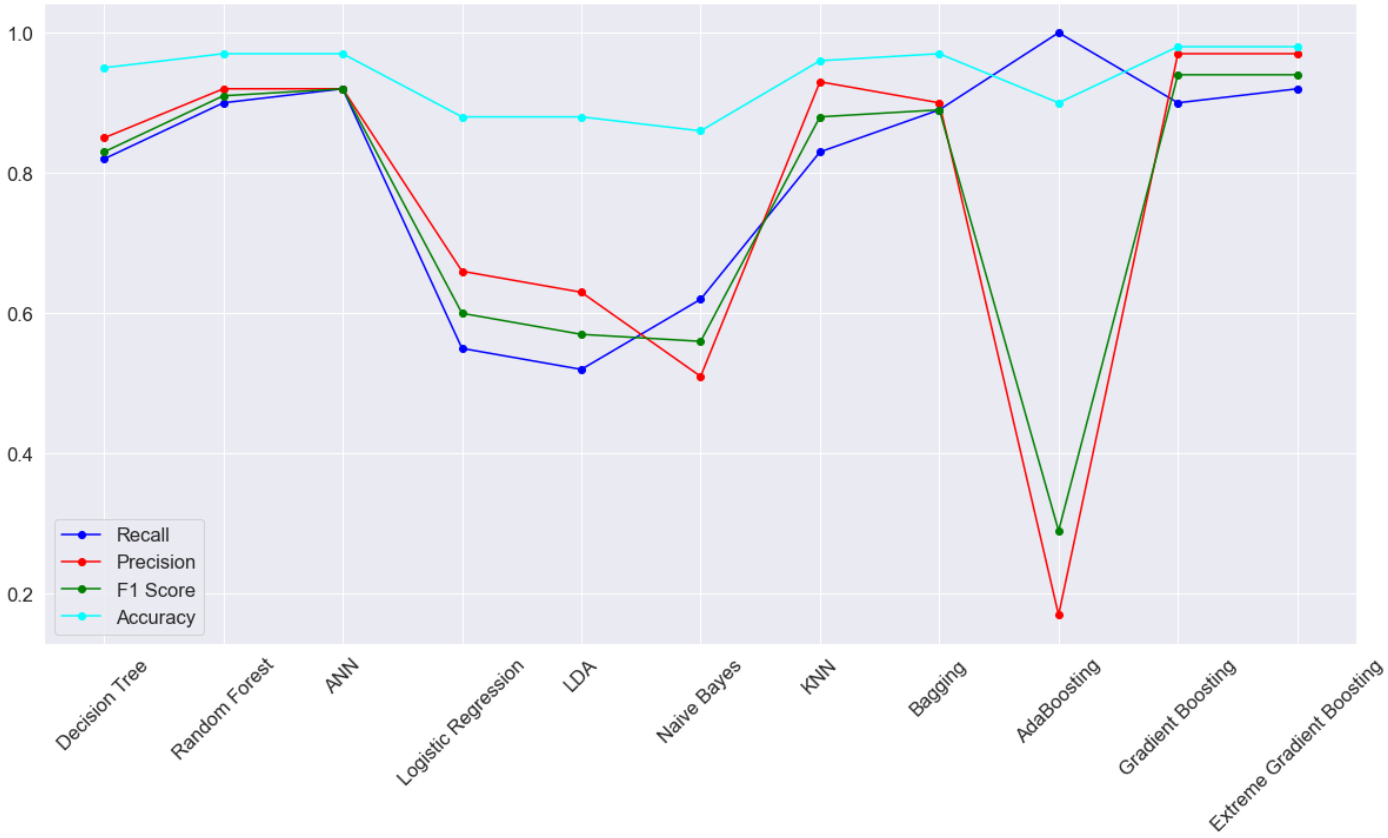


Figure 12. Performance Metrics of all Models with Tuned Hyperparameters and threshold of 0.4 for the Test Dataset.

From the above table and plot, we can derive the below inferences.

- Extreme Gradient Boosting (F1 Score = 0.94), Gradient Boosting (F1 Score = 0.94), Artificial Neural Networks (F1 Score = 0.92), and Random Forest models (F1 Score = 0.91) are the **most optimized models based their F1 scores**.
- As class-level performance metrics like Precision, Recall, and F1 **Score are very much near to the above four models**, we can select any one model from the above four models as the best **model based on interpretation requirements by business and computational power**.
- As Extreme Gradient Boosting Model (Precision = 0.97, Recall = 0.92, F1 Score = 0.94) is performing **slightly better than the remaining three models**, let us consider **the Extreme Gradient Boosting Model as the best model for predicting churn**.

Q6. Final interpretation / recommendation

Interpretation of the Extreme Gradient Boosting (XGB) Model

- As the recall is 0.92, this model may miss 8 customers who are likely to churn out of 100 customers. **Those 8 customers are predicted as not churn by the model**. The DTH company may not provide any

customer-specific offers to these 8 customers and these customers may churn over time. **The business has to spend more money to attract new customers in place of the churned customers.**

- As the precision is 0.97, this model **may add 3 customers who are not likely to churn out of 100 customers to the churning customers' list. Those 3 customers are predicted to churn by the model but actually, they are not likely to churn.** The DTH company may provide customer-specific offers to these 3 customers also along with the remaining 97 customers. This is a loss to the business. **Unnecessarily business is spending on these three customers to retain them even though they are not likely to churn.** Anyway, it is only for three customers, it would be less amount only.

Feature Importance

Feature	IMP
Tenure	0.2
Complain Last Year	0.18
City Tier	0.08
Account Segment	0.07
CC Agent Score	0.07
Day since CC Connect	0.06
Payment	0.06
Login Device	0.06
Marital Status	0.05
Gender	0.05

- According to XGB Model, **Tenure, any complaints received in the last 12 months, account segment, city tier, and customer care agent score are the five most important features** for predicting the churn.

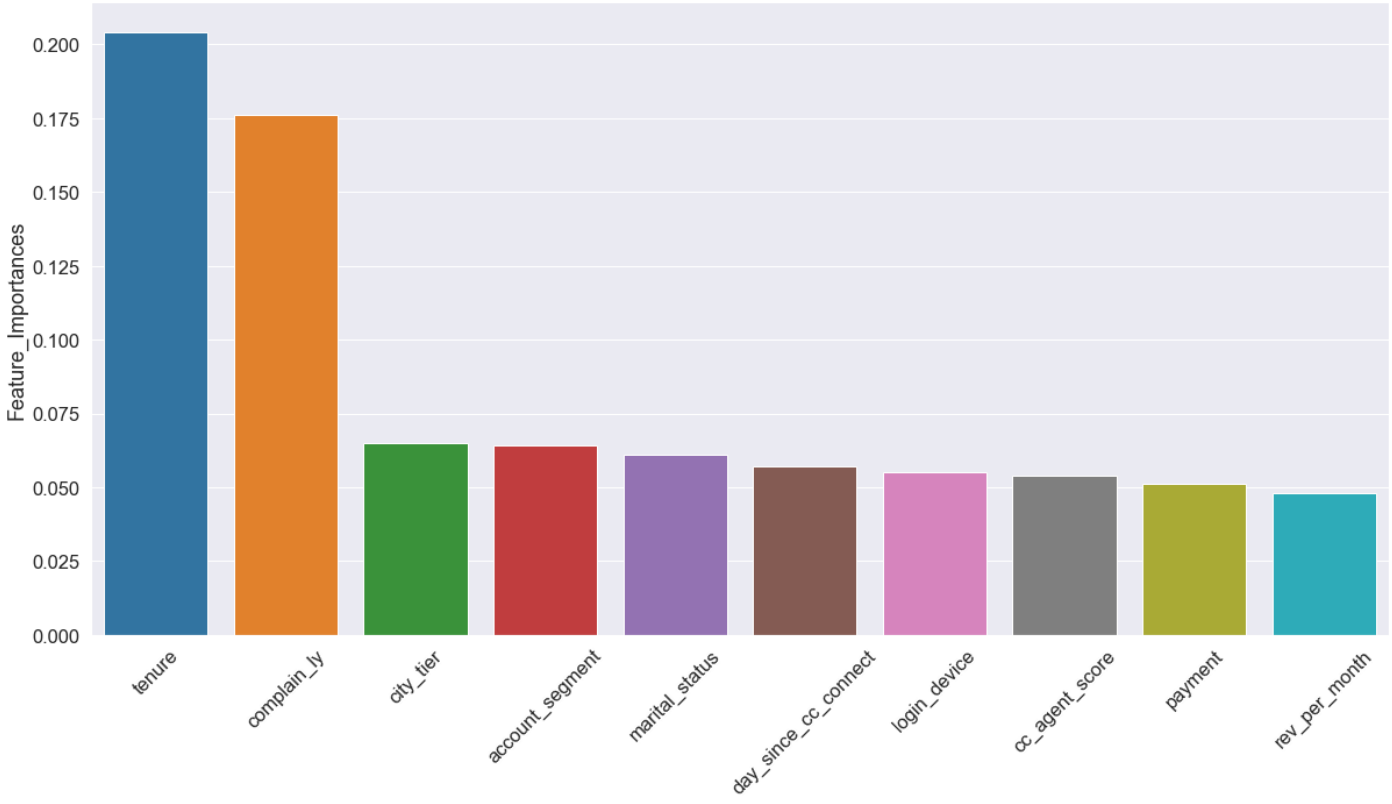


Figure 13. Features Importance of Top 10 Features in Extreme Gradient Boosting Model.

Recommendations

The following are the recommendations given to the DTH company.

- **Extreme Gradient Boosting model** can be used for churn prediction.

Measure	Tenure	
	Not Churning	Churning
Mean	12.4	4.3
Standard Deviation	12.6	12.1
Minimum	0	0
Q1	5	0
Median	10	1
Q3	17	3
Maximum	99	99

- 75th percentile of tenure of churned customers is less than the 25th percentile of tenure of not-churned customers. Hence, there is clear evidence that most of the **churned customers may be newly added customers**.
- **Newly joined customers should be given some offers** to reduce the churn rate.

Any Complaints/Churn	No	Yes
Not Churning	89%	68%
Churning	11%	32%

- Almost 32% of customers who raised any complaints in the last 12 months are churning. **It means most of the customers are not satisfied with the way complaints are resolved**. The Business should take some measurable actions on the **complaints resolution system.**

Churn/City Tier	Not Churning	Churning
Tier 1	85.5%	14.5%
Tier 2	80.0%	20.0%
Tier 3	78.6%	21.4%

- Customers from **tier 3 cities (21.4%) are churning more** than those from other cities.
- The company may review the customers who are residing in **tier 3 cities to attract them with suitable offers to retain** them for a long time.

Churn/Account Segment	Not Churning	Churning
HNI	84.4%	15.6%
Regular	92.3%	7.7%
Regular Plus	72.7%	27.3%
Super	89.8%	10.2%
Super Plus	95.1%	4.9%

- Customers from the **regular plus (27.3%) segment** are churning more than those from other segments.
- The company may think about this segment of customers **to provide customer-specific offers** to retain them for a long time.

Payment/Churn	Cash on Delivery	Credit Card	Debit Card	E-Wallet	UPI
Not Churning	75	85.8	84.7	77.3	82.6
Churning	25	14.2	15.3	22.7	17.4

- Customers who are paying through **cash on delivery (25%) and E-wallet (22.7%) are churning** more than other customers.
- Customers may not be getting any discounts and offers if they pay through cash on delivery and E-wallet.
- The Company can review these payment options and **add a few discounts to these payment options**.

Measure	Day Since CC Connect	
	Not Churning	Churning
Mean	5	3
Standard Deviation	4	4
Minimum	0	0
Q1	2	1
Median	4	2
Q3	8	6
Maximum	31	47

- **Churning customers** contacted customer care **recently** than not churning customers.
- Customers might be **churning if they are not satisfied** with the support provided by customer care.
- The Business should focus on improving the **customer care support system.**