Capstone Project : Customer Churn Prediction

Haresh P Tayade

PGP-DSBA

Online Sept'2021

Date-31/07/2022

Customer Churn Prediction

Review Parameters	***Section***
Defining problem statement	***Introduction of the business problem***
Need of the study/project	
Understanding business/social opportunity	
Understanding how data was collected in terms of time, frequency and methodology	***Data Report***
Visual inspection of data (rows, columns, descriptive details)	
Understanding of attributes (variable info, renaming if required)	
Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)	***Exploratory data analysis***
Bivariate analysis (relationship between different variables , correlations)	
Removal of unwanted variables (if applicable)	
Missing Value treatment (if applicable)	
Outlier treatment (if required)	
Variable transformation (if applicable)	
Addition of new variables (if required)	
Is the data unbalanced? If so, what can be done? Please explain in the context of the business	***Business insights from EDA***
Any business insights using clustering (if applicable)	
Any other business insights	

Introduction to Business Problem

Defining Problem statement

A 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 a single account can have multiple customers. Hence, by losing one account the company might be losing more than one customer.

Need of Study/Project

The study of the customer churn is essential to companies usually. We have a greater focus on customer acquisition and keep retention as a secondary priority and E-commerce industry is the most competitive industry with high acquisition costs and changing tariff plans. It can cost Two times more to attract a new customer than it does to retain an existing one. Additionally, the existing customers have the potential to get the company new customers via word of referral helping the business in indirect selling. The study would be to focus on discovering the impact of the independent variables on the dependent variables and also understand the Root Cause Analysis of the factors leading the customer churning on a very high level.

Understanding Business/Social Opportunities

The Root Cause Analysis of the bleeding factors will indicate the severity of the existing issue and the creation of a churn prediction model will define the direction of the business in the upcoming days, which in turn will help the company to restructure its business model for customer retention and increasing goodwill. The company will be able to assess the touch points

Data Report

Understanding how data was collected in terms of time frequency and methodology

The data has been collected via –

- *KYC details of the customers.*
- *Customer care call assistance details.*
- *Aggregated revenue collection details on different periods.*
- *Demographic details captured from login devices.*

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

Head of Dataset

	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status	rev_per_month	Complain_ly
0	20000	1	4	3.0	6.0	Debit Card	Female	3.0	3	Super	2.0	Single	9	1.0
1	20001	1	0	1.0	8.0	UPI	Male	3.0	4	Regular Plus	3.0	Single	7	1.0
2	20002	1	0	1.0	30.0	Debit Card	Male	2.0	4	Regular Plus	3.0	Single	6	1.0
3	20003	1	0	3.0	15.0	Debit Card	Male	2.0	4	Super	5.0	Single	8	0.0
4	20004	1	0	1.0	12.0	Credit Card	Male	2.0	3	Regular Plus	5.0	Single	3	0.0
5	20005	1	0	1.0	22.0	Debit Card	Female	3.0	NaN	Regular Plus	5.0	Single	2	1.0
6	20006	1	2	3.0	11.0	Cash on Delivery	Male	2.0	3	Super	2.0	Divorced	4	0.0
7	20007	1	0	1.0	6.0	Credit Card	Male	3.0	3	Regular Plus	2.0	Divorced	3	1.0
8	20008	1	13	3.0	9.0	E wallet	Male	2.0	4	Regular Plus	3.0	Divorced	2	1.0
9	20009	1	0	1.0	31.0	Debit Card	Male	2.0	5	Regular Plus	3.0	Single	2	0.0

Tail of Dataset

	AccountID	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status	rev_per_month	Complain_
11255	31255	0	10	1.0	34.0	Credit Card	Male	3.0	2	Super	1.0	Married	9	
11256	31256	0	13	1.0	19.0	Credit Card	Male	3.0	5	HNI	5.0	Married	7	
11257	31257	0	1	1.0	14.0	Debit Card	Male	3.0	2	Super	4.0	Married	7	
11258	31258	0	23	3.0	11.0	Credit Card	Male	4.0	5	Super	4.0	Married	7	
11259	31259	0	8	1.0	22.0	Credit Card	Male	3.0	2	Super	3.0	Married	5	

Info of Dataset

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

Description of Dataset

	Churn	Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status	rev_per_month	Complain_
count	11260.000000	11158.0	11148.000000	11158.000000	11151	11152	11162.000000	11148.0	11163	11144.000000	11048	11158.0	10903.000
unique	NaN	38.0	NaN	NaN	5	4	NaN	7.0	7	NaN	3	59.0	N
top	NaN	1.0	NaN	NaN	Debit Card	Male	NaN	4.0	Super	NaN	Married	3.0	N
freq	NaN	1351.0	NaN	NaN	4587	6328	NaN	4569.0	4062	NaN	5860	1746.0	N
mean	0.168384	NaN	1.653929	17.867091	NaN	NaN	2.902526	NaN	NaN	3.066493	NaN	NaN	0.285
std	0.374223	NaN	0.915015	8.853269	NaN	NaN	0.725584	NaN	NaN	1.379772	NaN	NaN	0.451
min	0.000000	NaN	1.000000	4.000000	NaN	NaN	0.000000	NaN	NaN	1.000000	NaN	NaN	0.000
25%	0.000000	NaN	1.000000	11.000000	NaN	NaN	2.000000	NaN	NaN	2.000000	NaN	NaN	0.000
50%	0.000000	NaN	1.000000	16.000000	NaN	NaN	3.000000	NaN	NaN	3.000000	NaN	NaN	0.000
75%	0.000000	NaN	3.000000	23.000000	NaN	NaN	3.000000	NaN	NaN	4.000000	NaN	NaN	1.000
max	1.000000	NaN	3.000000	132.000000	NaN	NaN	5.000000	NaN	NaN	5.000000	NaN	NaN	1.000

Check Null Values

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

- *Percentile values of most of the variables suggest presence of NaN values, indicating Data has lot of missing values which may need further treatment basis their data types.*
- *Mean and Median values of all the variables have fair differences, which indicates data must be skewed.*
- *Values of Max values, Min Values compared to 75% and 25%, indicates high number of outliers.*
- *Some of the continuous variables is found to be of 'object datatype, suggesting presence of unwanted variables,*
- *Gender column is supposed to have only 2 variables, but value count Suggest of 4 unique values, suggesting possible renaming of some of the values*
- *Let's check for null values, duplicate values, outliers and skewness.*
- *There are Lot of Null Values in the given Dataset.*

Understanding the Attributes(variable info, renaming if required)

Checking for unique variable

```
Tenure
[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]


Payment
['Debit Card' 'UPI' 'Credit Card' 'Cash on Delivery' 'E wallet' nan]


Gender
['Female' 'Male' 'F' nan 'M']


Account_user_count
[3 4 nan 5 2 '@' 1 6]


account_segment
['Super' 'Regular Plus' 'Regular' 'HNI' 'Regular +' nan 'Super Plus'
 'Super +']


Marital_Status
['Single' 'Divorced' 'Married' nan]


rev_per_month
[9 7 6 8 3 2 4 10 1 5 '+' 130 nan 19 139 102 120 138 127 123 124 116 21
 126 134 113 114 108 140 133 129 107 118 11 105 20 119 121 137 110 22 101
 136 125 14 13 12 115 23 122 117 131 104 15 25 135 111 109 100 103]
```

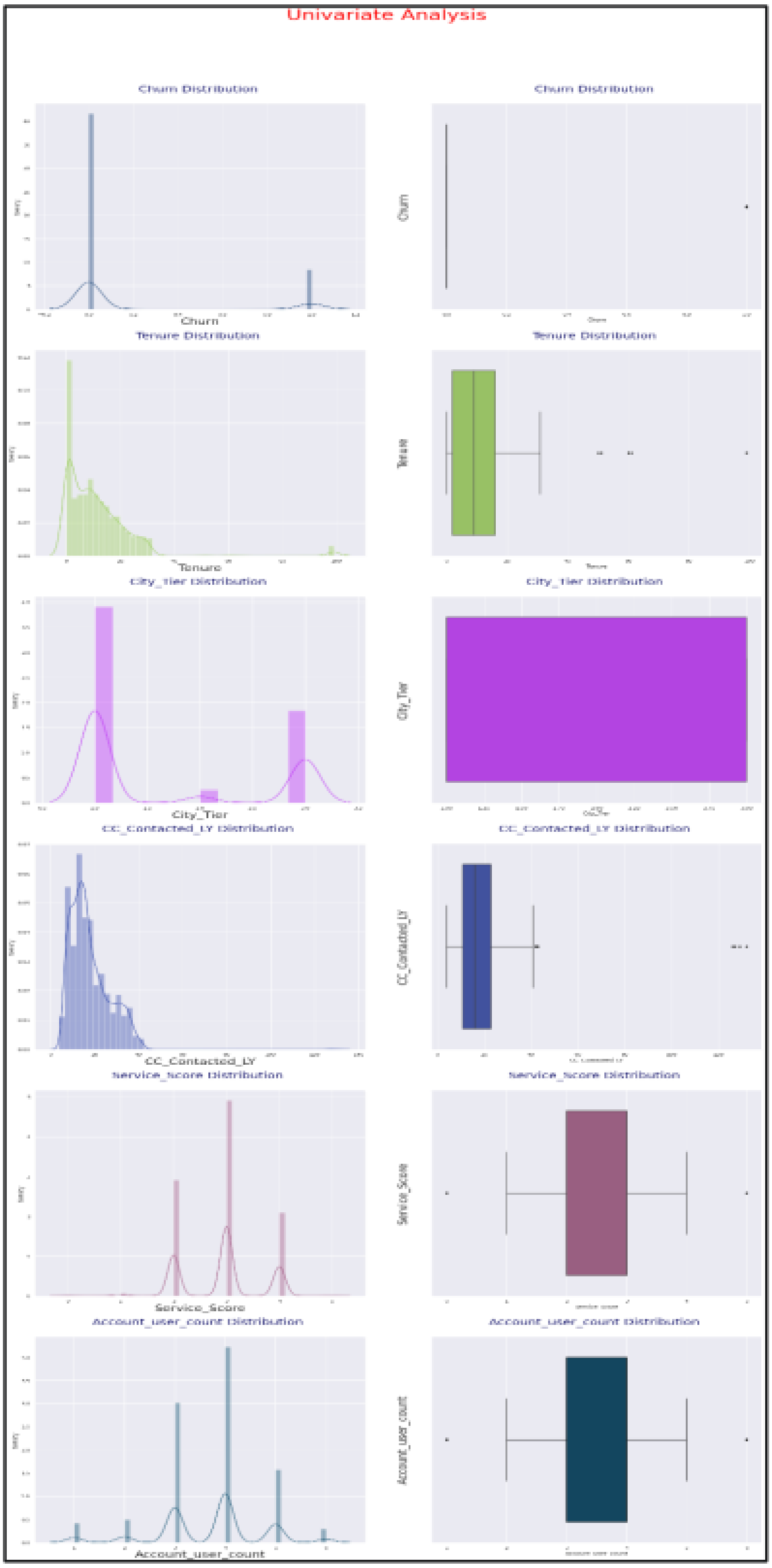
```
rev_growth_yoy
[11 15 14 23 22 16 12 13 17 18 24 19 20 21 25 26 '$' 4 27 28]


coupon_used_for_payment
[1 0 4 2 9 6 11 7 12 10 5 3 13 15 8 '#' '$' 14 '*' 16]


Day_Since_CC_connect
[5 0 3 7 2 1 8 6 4 15 nan 11 10 9 13 12 17 16 14 30 '$' 46 18 31 47]


cashback
[159.93 120.9 nan ... 227.36 226.91 191.42]


Login_device
['Mobile' 'Computer' '&&&&' nan]
```

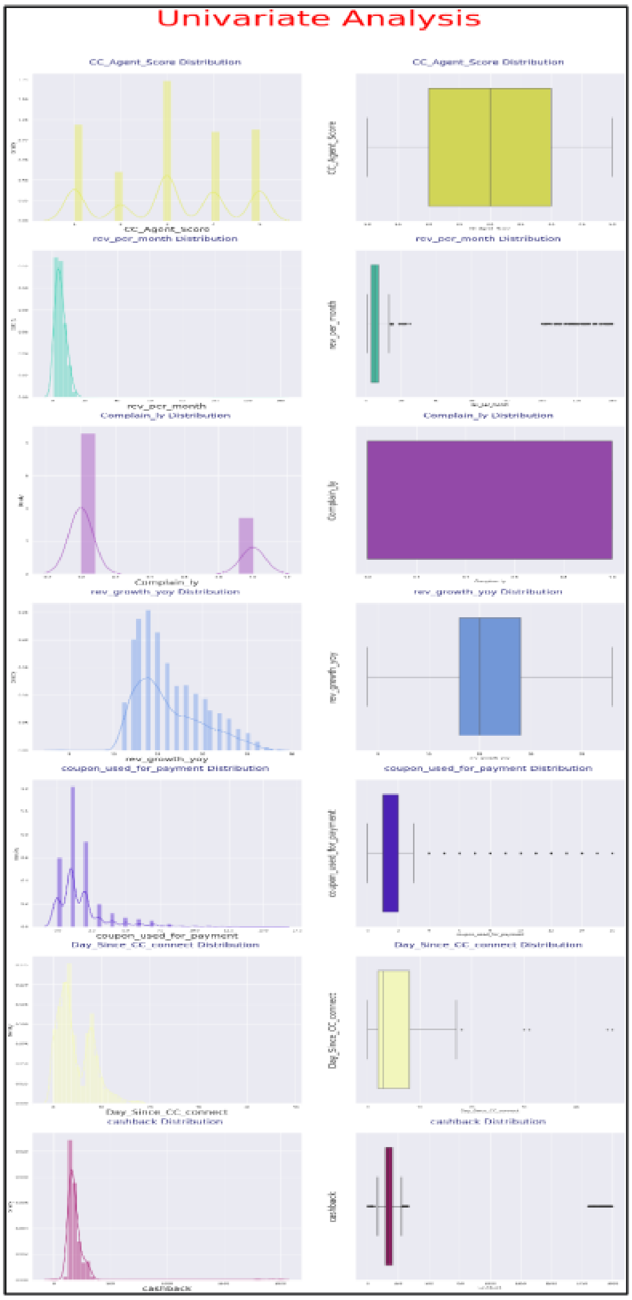
Converting the string variables to Numerical for Preprocessing

```
<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
```

Exploratory data analysis

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


Univariate Analysis
Churn Distribution
Churn Distribution
Churn
Churn
Tenure Distribution
Tenure Distribution
Tenure
Tenure
City_Tier Distribution
City_Tier Distribution
City_Tier
City_Tier
CC_Contacted_LY Distribution
CC_Contacted_LY Distribution
CC_Contacted_LY
CC_Contacted_LY
Service_Score Distribution
Service_Score Distribution
Service_Score
Service_Score
Account_user_count Distribution
Account_user_count Distribution
Account_user_count
Account_user_count


Univariate Analysis
CC_Agent_Score Distribution
CC_Agent_Score
rev_per_month Distribution
rev_per_month
Complain_ly Distribution
Complain_ly
rev_growth_yoy Distribution
rev_growth_yoy
coupon_used_for_payment Distribution
coupon_used_for_payment
Day_Since_CC_connect Distribution
Day_Since_CC_connect
cashback Distribution
cashback

Skewness

```
churn.skew()
```

```
Churn                      1.772606
Tenure                     3.895707
City_Tier                  0.737107
CC_Contacted_LY            1.422977
Service_Score              0.003891
Account_user_count        -0.393100
CC_Agent_Score            -0.142149
rev_per_month              9.093909
Complain_ly                0.950876
rev_growth_yoy             0.752474
coupon_used_for_payment    2.575199
Day_Since_CC_connect       1.273021
cashback                   8.770766
dtype: float64
```

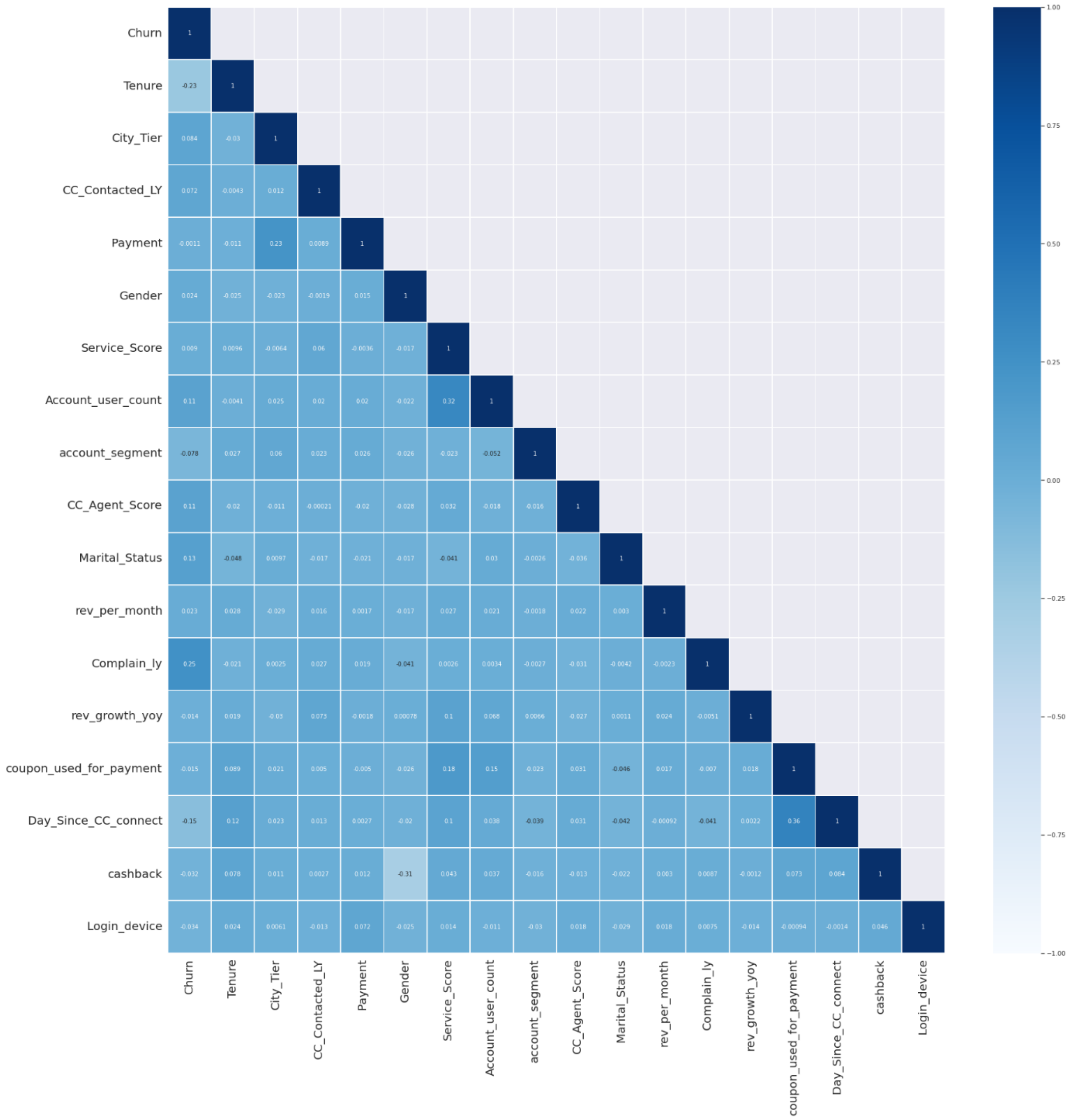
- *We observe from the plots that the variables 'Tenure', ,'CC_Contacted_LY', 'Service_Score','Account_user_count', 'rev_per_month','rev_growth_yoy', coupon_used_for_payment', 'Day_Since_CC_connect','cashback' have outliers*
- *Also the data in case of all variables are skewed, in either direction*
- *Data in its original form may not follow normal distribution.*

Multi-variate analysis

Correlation between the variables

	Churn	Tenure	City_Tier	CC_Contacted_LY	Service_Score	Account_user_count	CC_Agent_Score	rev_per_month	Complain_ly	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback
Churn	1.000000	-0.233150	0.084135	0.072071	0.008991	0.107292	0.105796	0.022538	0.251488	-0.013877	-0.014826	-0.147956	-0.032382
Tenure	-0.233150	1.000000	-0.030223	-0.004261	0.009608	-0.004052	-0.020031	0.028431	-0.021419	0.018824	0.089171	0.122612	0.078416
City_Tier	0.084135	-0.030223	1.000000	0.011835	-0.006443	0.025498	-0.011479	-0.028521	0.002478	-0.030025	0.021124	0.023158	0.010516
CC_Contacted_LY	0.072071	-0.004261	0.011835	1.000000	0.060052	0.020351	-0.000209	0.015675	0.027244	0.072913	0.004969	0.012938	0.002679
Service_Score	0.008991	0.009608	-0.006443	0.060052	1.000000	0.323327	0.032135	0.026691	0.002643	0.103374	0.181914	0.099770	0.042961
Account_user_count	0.107292	-0.004052	0.025498	0.020351	0.323327	1.000000	-0.017522	0.020600	0.003411	0.067616	0.146081	0.037573	0.037057
CC_Agent_Score	0.105796	-0.020031	-0.011479	-0.000209	0.032135	-0.017522	1.000000	0.022167	-0.031459	-0.027159	0.030661	0.030808	-0.012599
rev_per_month	0.022538	0.028431	-0.028521	0.015675	0.026691	0.020600	0.022167	1.000000	-0.002262	0.024114	0.016548	-0.000923	0.002974
Complain_ly	0.251488	-0.021419	0.002478	0.027244	0.002643	0.003411	-0.031459	-0.002262	1.000000	-0.005122	-0.007023	-0.041111	0.008734
rev_growth_yoy	-0.013877	0.018824	-0.030025	0.072913	0.103374	0.067616	-0.027159	0.024114	-0.005122	1.000000	0.018341	0.002206	-0.001157
coupon_used_for_payment	-0.014826	0.089171	0.021124	0.004969	0.181914	0.146081	0.030661	0.016548	-0.007023	0.018341	1.000000	0.361735	0.072861
Day_Since_CC_connect	-0.147956	0.122612	0.023158	0.012938	0.099770	0.037573	0.030808	-0.000923	-0.041111	0.002206	0.361735	1.000000	0.084465
cashback	-0.032382	0.078416	0.010516	0.002679	0.042961	0.037057	-0.012599	0.002974	0.008734	-0.001157	0.072861	0.084465	1.000000

Correlation PLot



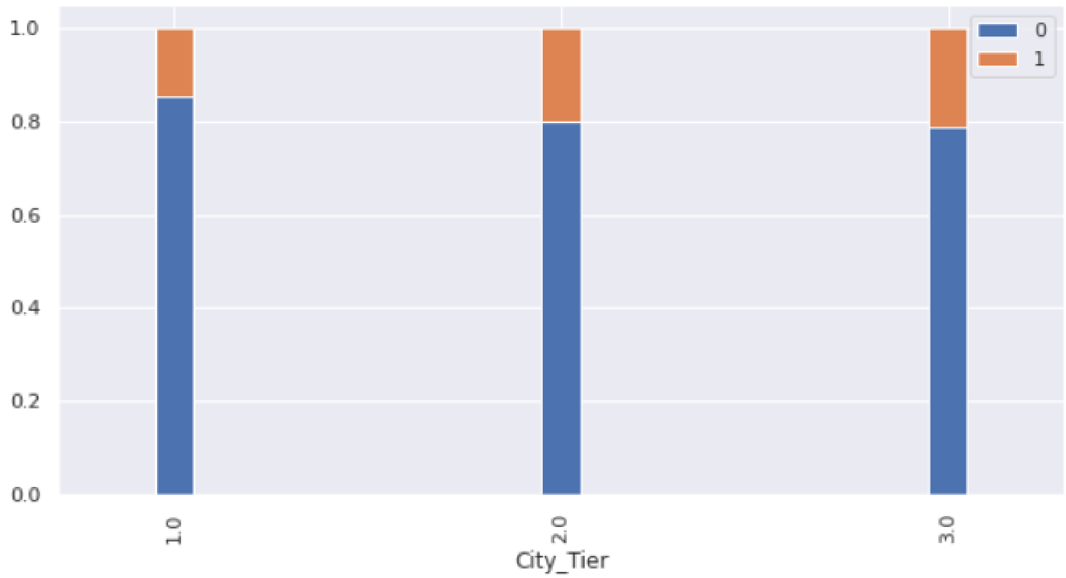
Independent among each others shows very less collinearity

Pairplot

Independent variables shows very less correlation between each other

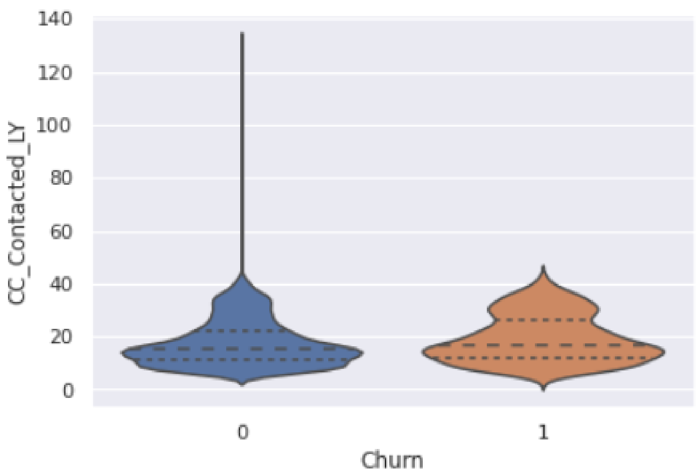
Bi-Variate analysis

Churn vs City_Tier



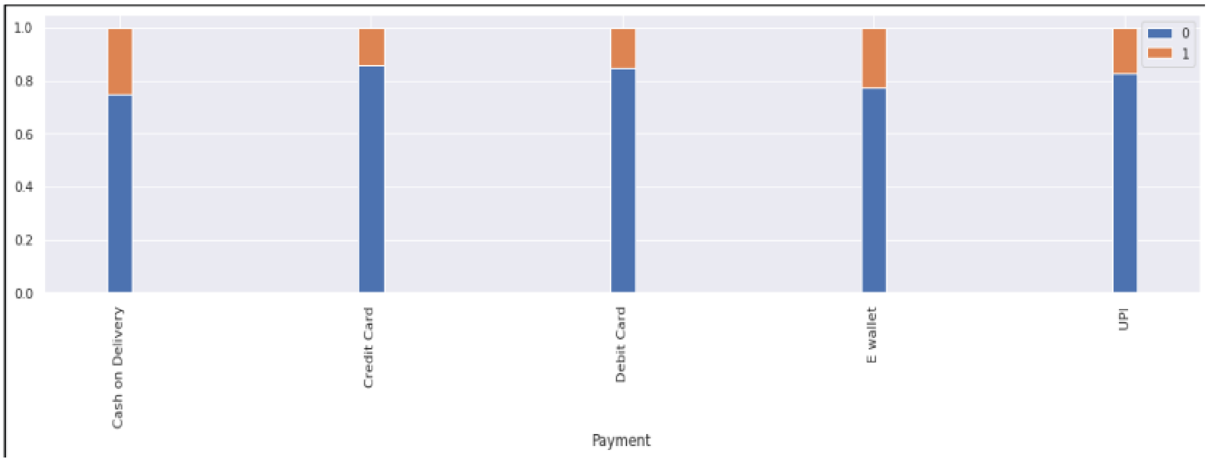
The consumers belonging to Tier 1 cities churns less compared to Tier 2 and Tier 3 cities. The Tier 3 cities show the maximum impact in churn ratio, possibly, due to the lack of infrastructure development and customer service delays and may be preferring the traditional local owned cable operators to DTH connections.

Churn vs CC_Contacted_LY



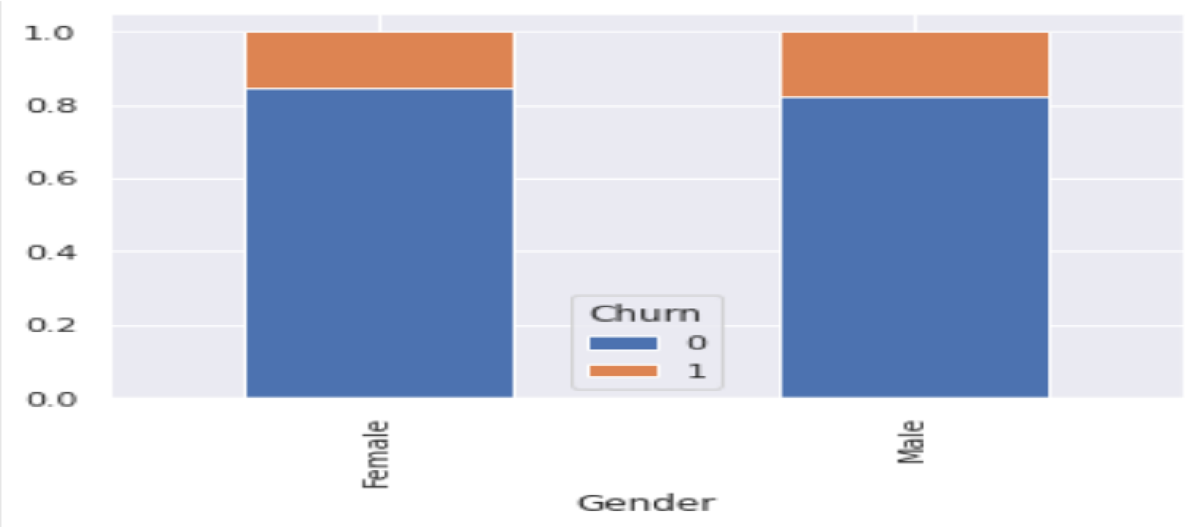
The consumers who have contacted Customer care in around 35 or more time seems to be churning more. However above feature does not suggest significant difference pattern of churning among customers with those who do not churn.

Churn VS Payment



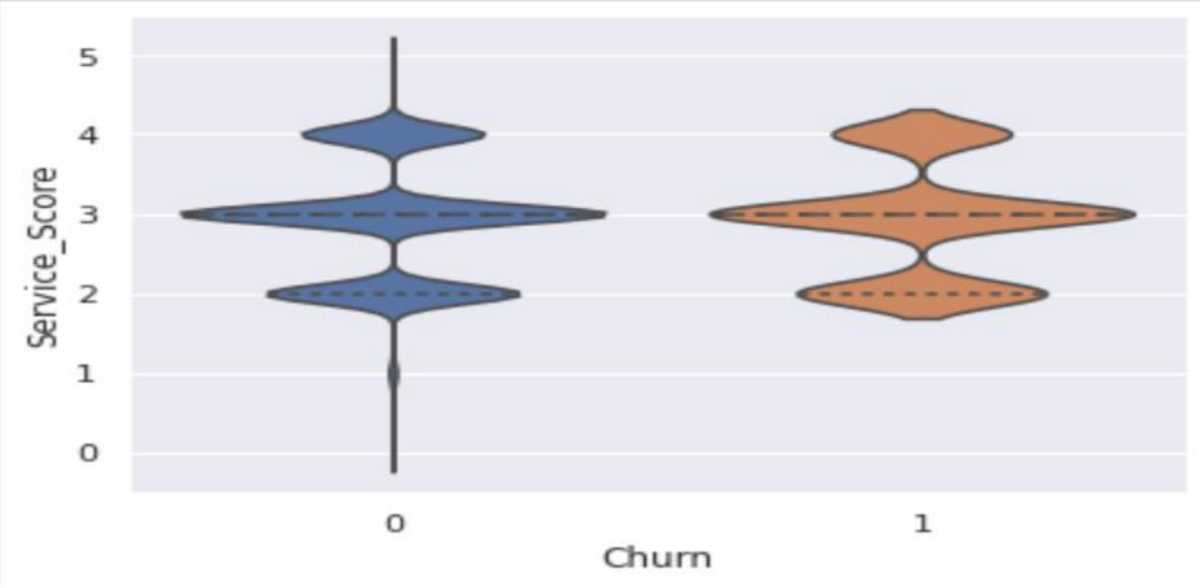
The consumers with payment modes 'Cash on Delivery' and 'E wallet' have churned inflated more than the ones using 'Credit/Debit Cards' and 'UPI'. This signifies that the consumers using 'Credit/Debit Cards' and 'UPI' have had the privilege of cashback offers and points added post each Transactions whereas the other two payment mode does not offer any kind of monetary benefit.

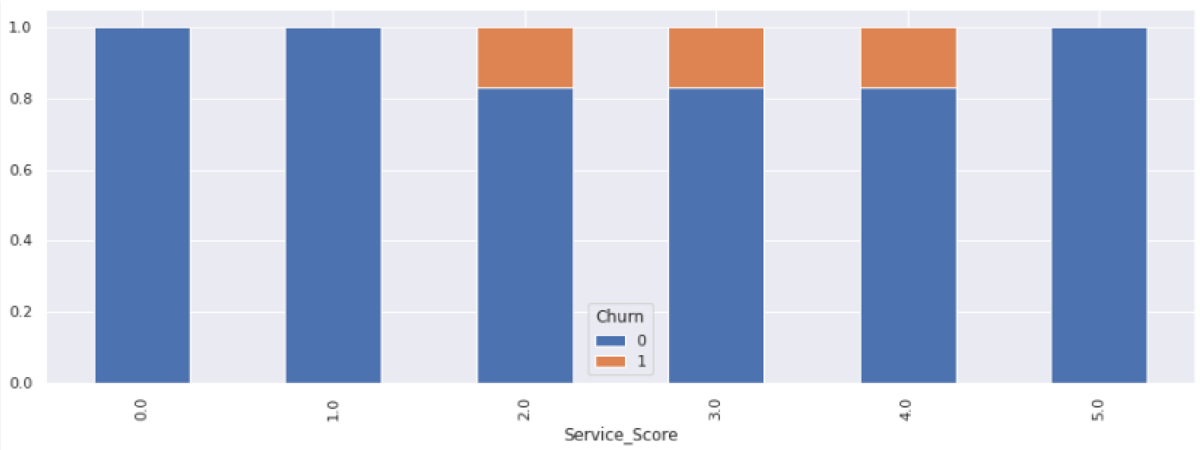
Churn VS Gender



There is not much significant difference in percentage of churning customers among both the genders.

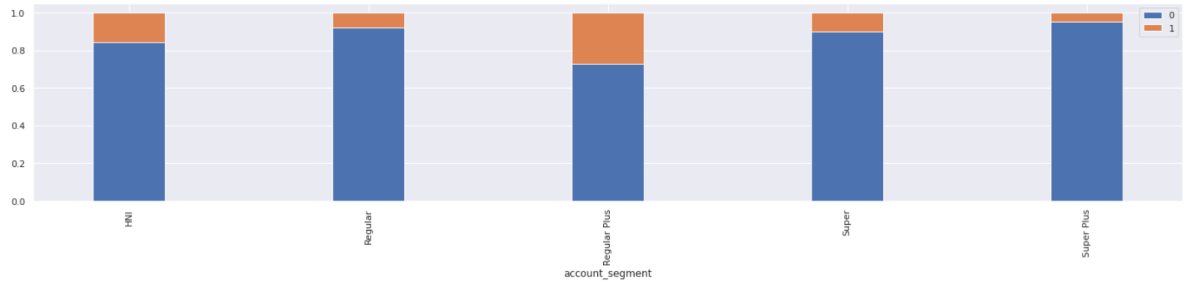
Churn VS Service Score





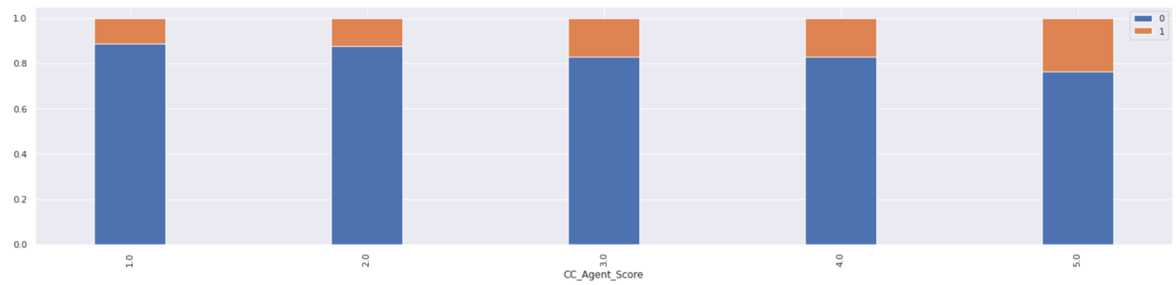
The consumers who have rated the service score from 2.0 – 4.0 are the ones who are more likely to churn compared to the ones given the lowest rating. The consumers with medium satisfaction rate might have been the ones having high contact ratio and the service score rates have been aggregated

Churn vs account segment



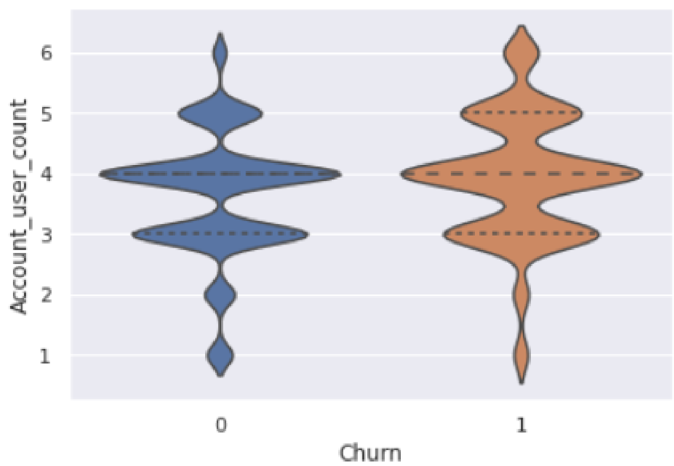
The customers using Regular Plus accounts seem to be churning more among others, possibly due to inadequate feature when compared to premiums

Churn vs CC_Agent_Score



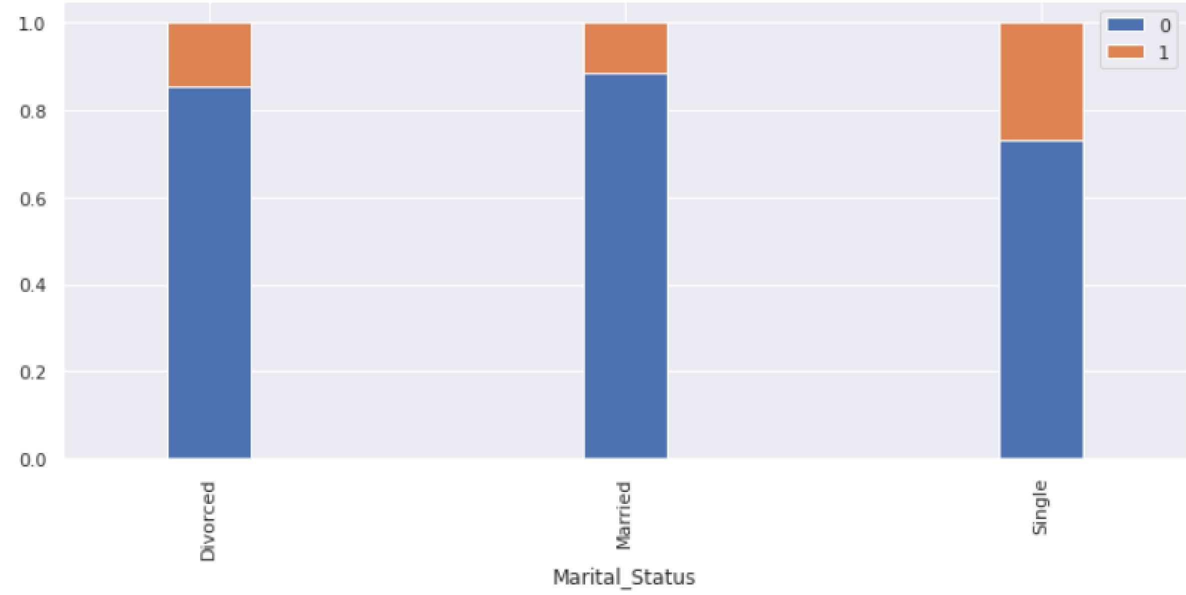
Churn vs Account_user_count

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ffa3fc8dc50>
```



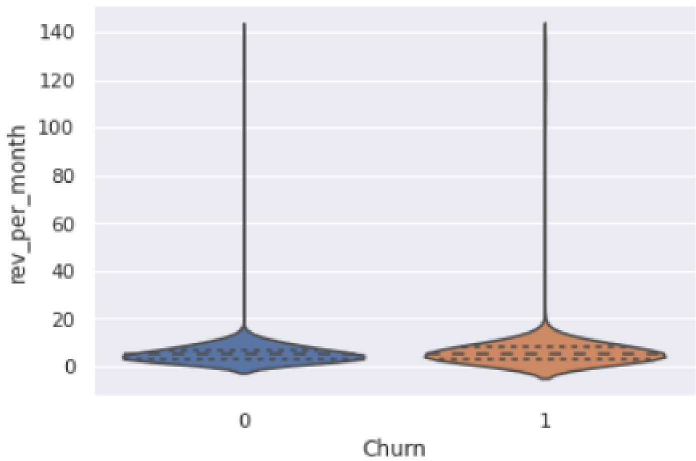
Churn vs Marital_Status

<matplotlib.legend.Legend at 0x7ffa3fcd2fd0>



The consumers belonging to the 'Single' and 'Divorced' category of marital status have churned more compared to the married consumers. The ratio of churning is sky-high in single consumers, perhaps, due to the fact that this segment is inclining more towards OTT platforms rather than relying on DTH.

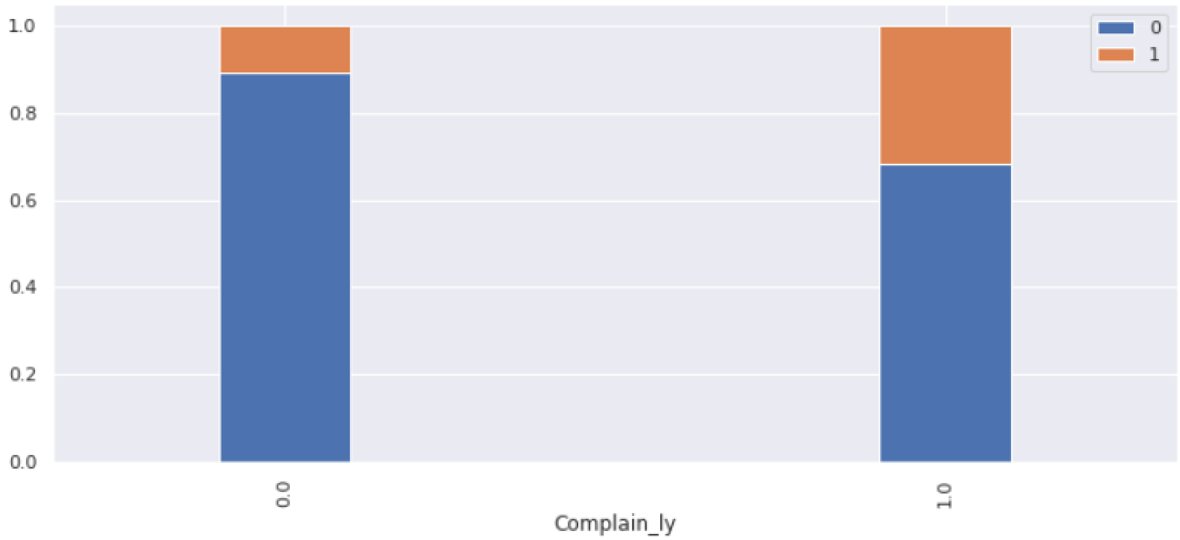
Churn vs rev_per_month



The revenue generated per month does not indicate significant pattern among churners and non-churners

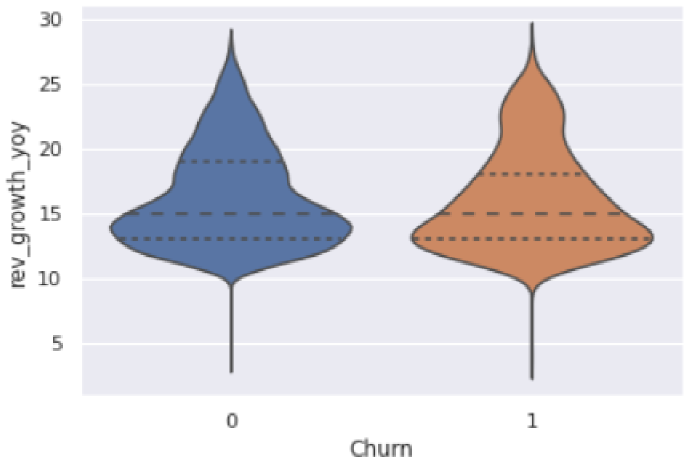
Churn vs Complain_ly

<matplotlib.legend.Legend at 0x7ffa3fb8e0d0>



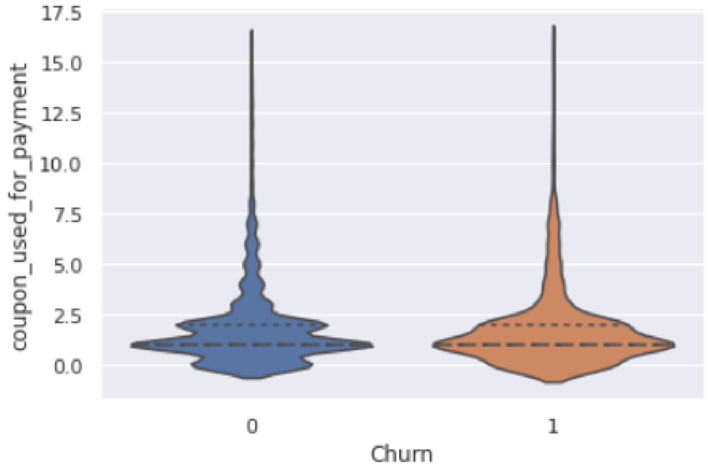
The customers with complaints registered have churned significantly more rather than the ones who have never registered any.

Churn vs rev_growth_yoy



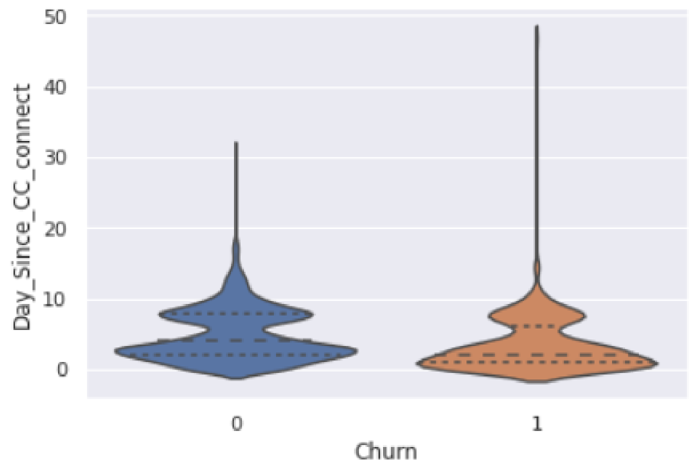
Percentage of Customers with yearly revenue growth above 23 seems to be churning more.

Churn vs coupon_used_for_payment



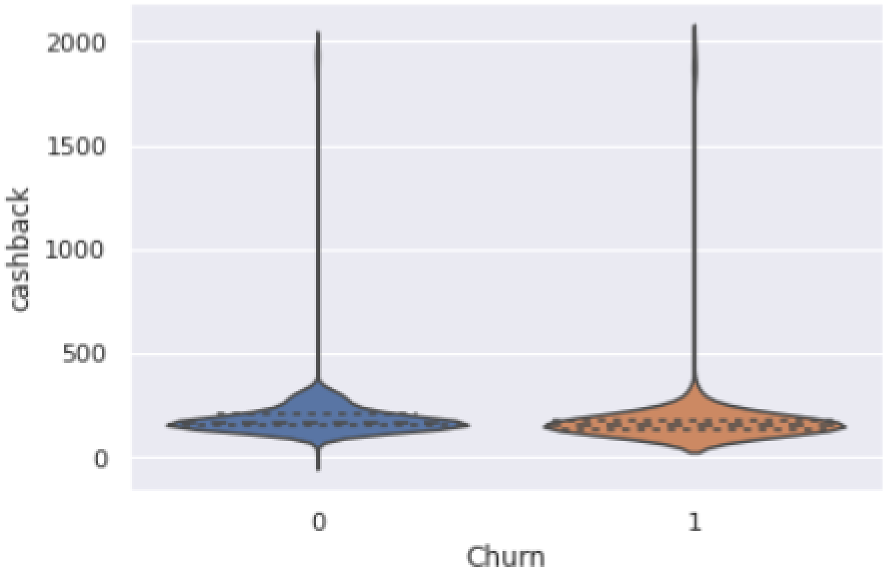
Percentage of Customers using less no of coupons seems to be churning more.

Churn vs Day_Since_CC_connect



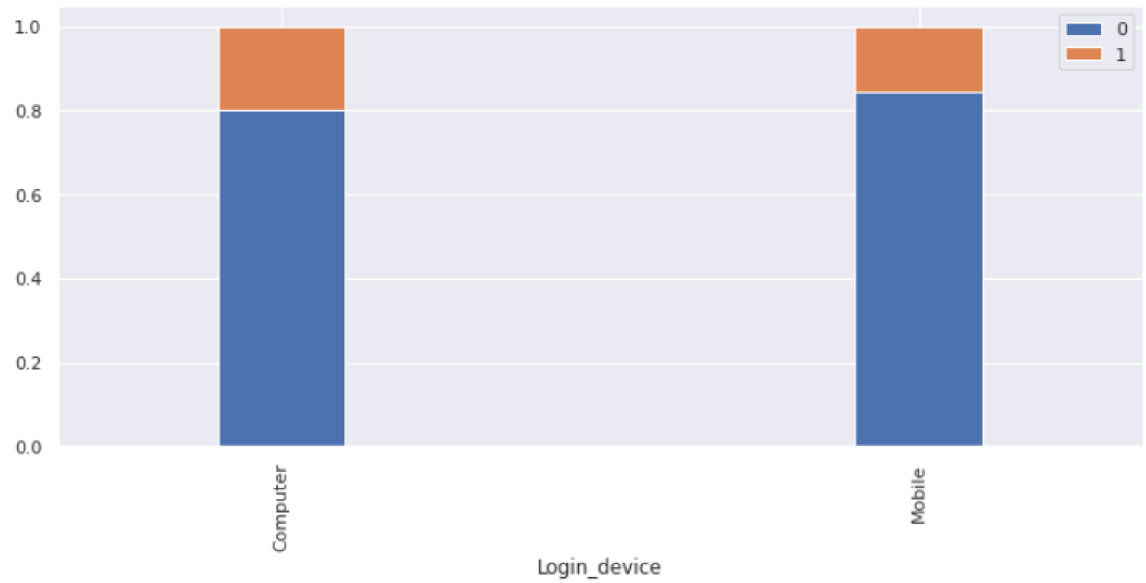
Customers who have not contacted with Customer care for longer period seems to be churning more compared to those with frequent connects.

Churn vs cashback



Churn vs Login_device

<matplotlib.legend.Legend at 0x7ffa3f91a810>



The consumers having logged in via mobile have churned less as per the graph plotted above. Possibly, due to the company providing services for payment, offers, product, customer care assistance solutions in form of an app

Now we do Train – Test Split

Splitting the data into training and Testing data to avoid overfitting and keep the data unseen.

X-head

Tenure	City_Tier	CC_Contacted_LY	Payment	Gender	Service_Score	Account_user_count	account_segment	CC_Agent_Score	Marital_Status	rev_per_month	Complain_ly	rev_growth_yoy	coupon_used_for_payment
4.0	3.0	6.0	2	0	3.0	3.0	3	2.0	2	9.0	1.0	11.0	1.0
0.0	1.0	8.0	4	1	3.0	4.0	2	3.0	2	7.0	1.0	15.0	0.0
0.0	1.0	30.0	2	1	2.0	4.0	2	3.0	2	6.0	1.0	14.0	0.0
0.0	3.0	15.0	2	1	2.0	4.0	3	5.0	2	8.0	0.0	23.0	0.0
0.0	1.0	12.0	1	1	2.0	3.0	2	5.0	2	3.0	0.0	11.0	1.0

Y-head

```
Y.head()
```

```
0    1
1    1
2    1
3    1
4    1
Name: Churn, dtype: int64
```

Dimension of Dataset of training and testing

```
x_train (7882, 17)
x_test (3378, 17)
y_train (7882,)
y_test (3378,)
```

Missing Values treatment

- KNN Imputer was first supported by Scikit-Learn in December 2019 when it released its version 0.22.
- This imputer utilizes the k-Nearest Neighbors method to replace the missing values in the datasets with the mean value from the parameter 'n_neighbors' nearest neighbors found in the training set.
- By default, it uses a Euclidean distance metric to impute the missing values.

Train set

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

Test set

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

Outlier Treatment

From the unvariate analysis performed below features are Categorical Data mostly of Ordinal Type, hence they do not require Outlier Treatment:

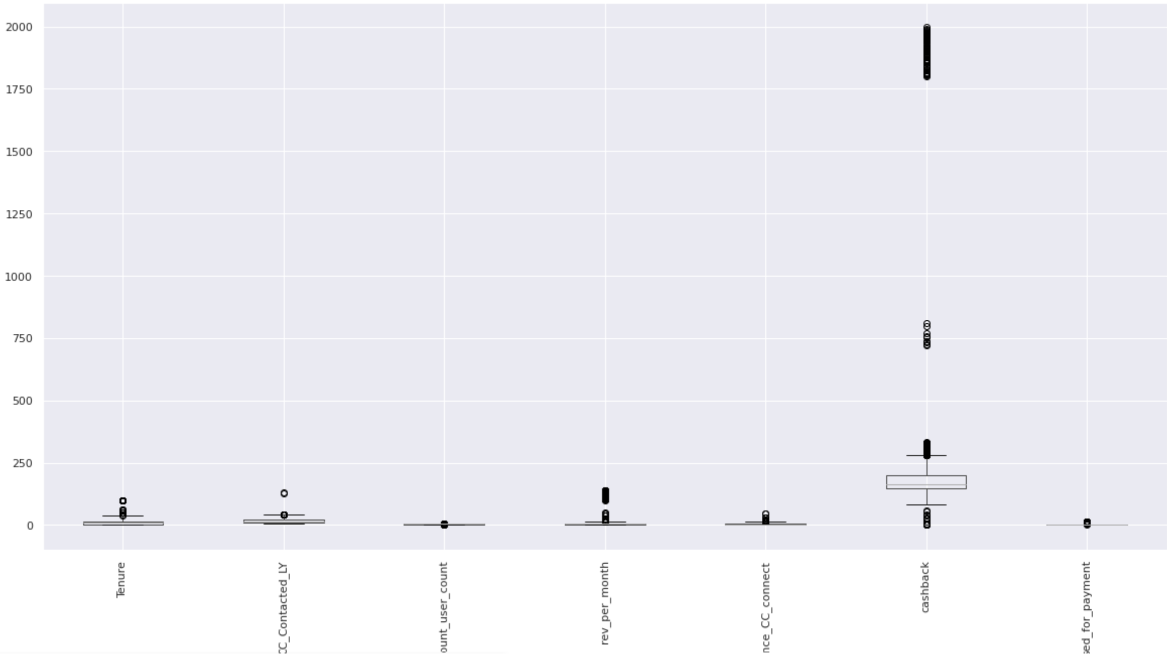
'Service_Score',

But Below Features require outlier treatment

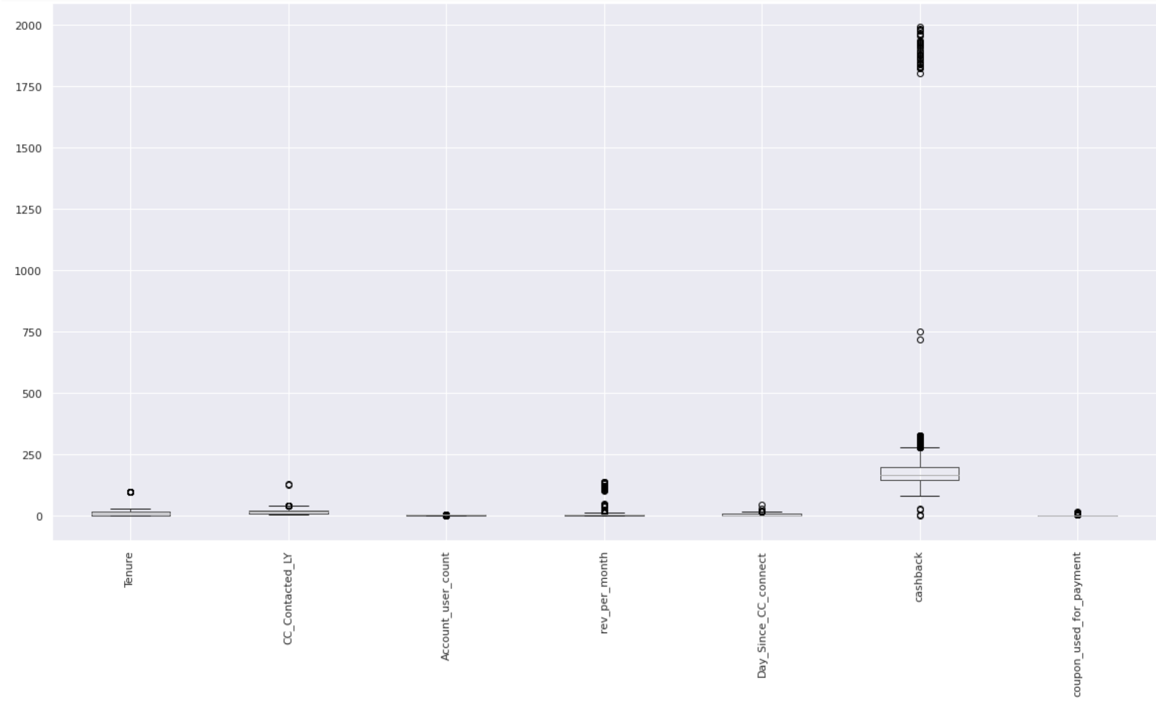
'Tenure',
'CC_Contacted_LY',
'Account_user_count',
'rev_per_month',
'rev_growth_yoy',
'Day_Since_CC_connect','cashback'
coupon_used_for_payment'

Checking outlier for training set

(array([1, 2, 3, 4, 5, 6, 7]), <a list of 7 Text major ticklabel objects>)



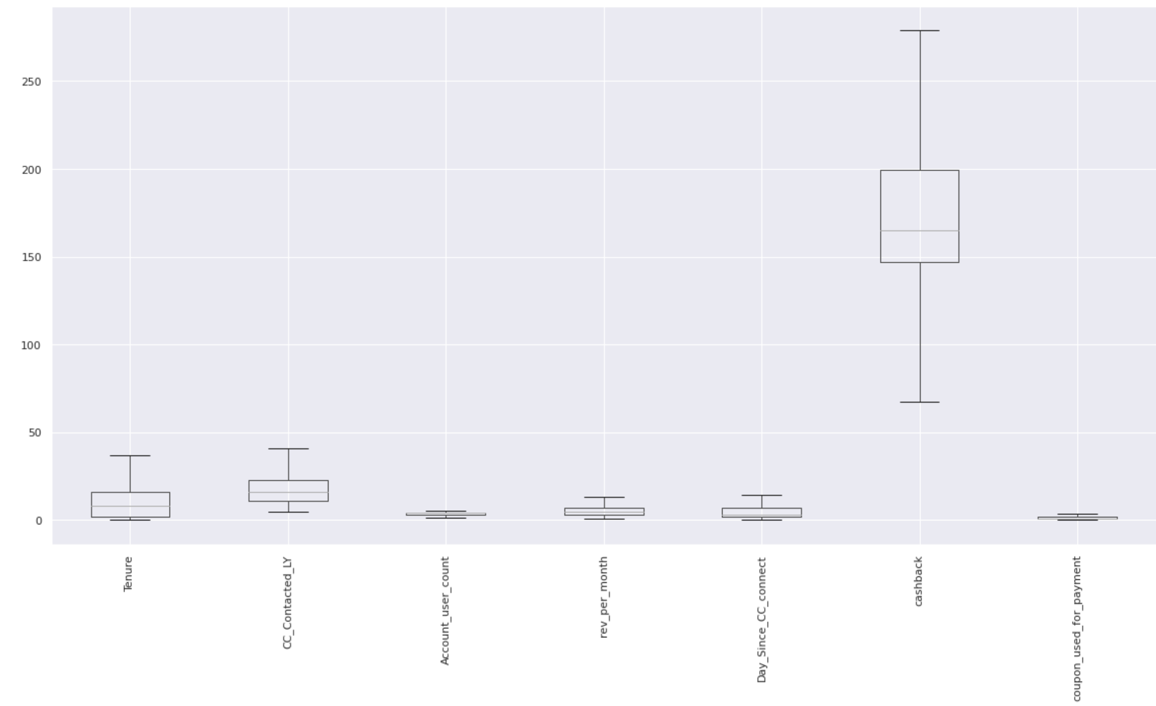
Checking outlier for testing set



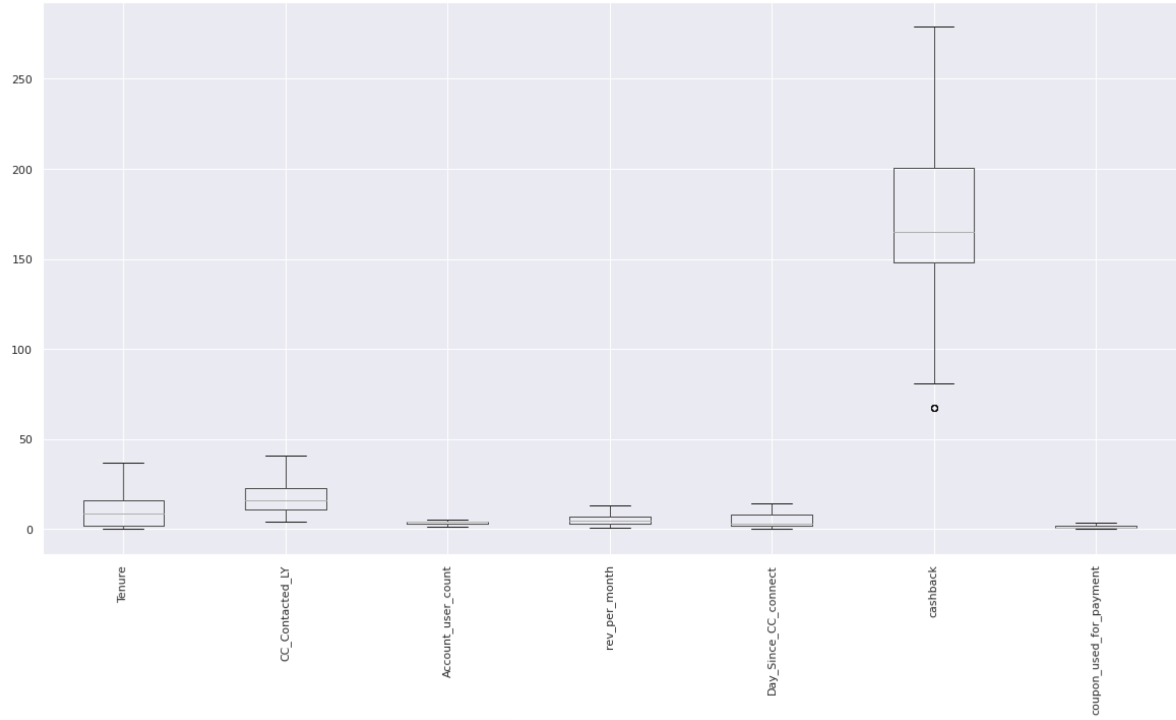
Remove outliers

Train set

(array([1, 2, 3, 4, 5, 6, 7]), <a list of 7 Text major ticklabel objects>)



Test set



Variable transformation (if applicable)

Standard Scaling the data:

- *Since features are based on different unit scale (especially revenue features and cashback) , it will be ideal to perform scaling of the data for the preparation of Data Models in future.*
- *It can be noted here that the higher magnitude of some variables will affected the distance between the two points which will impact the performance of distance based model as it will give higher weightage to variables which have higher magnitude.*
- *We do not want our algorithm to be affected by the magnitude of these variables and be biased towards variables with higher magnitude. To overcome this problem, we can bring down all the variables to the same scale.*
- *Various scaling techniques like Normalization, Min-max, Standard Scaler Technique can be used to scale the data.*

Business insights from EDA

Is the data unbalanced? If so, what can be done? Please explain in the context of the business

The dataset contains information of churning customers (17%) and non-churning customers (83%). Hence, it can be presumed that the data is imbalanced.

Such scenarios do occur in real world as generally the customer who will churn will always be less than those who have not.

Initially, we can try building our model based on the imbalanced proportion and if the model captures all the churning customers, we can continue with our implementation.

In case the model performs poorly, we can opt for the statistical technique - SMOTE – which creates artificial data points, thereby, to increase the values in the dataset in a balanced way.

```
0    0.831616
1    0.168384
Name: Churn, dtype: float64
```

Business insights using clustering

Clustering technique comes in handy in business models such as ours wherein we need to divide our customers into different segments which will help us to identify potential churners. This also helps us to determine fault in the service provided by the companies if any. The K-Means Clustering technique have been implemented to this dataset in order to create segments of consumers to understand the churning *behavior by categorizing them into different profiles. Out of the three profiles that has been created, interestingly in our case, Profile 2 only consists of customers who will churn.*

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Churn	1801.0	NaN	NaN	NaN	1.0	0.0	1.0	1.0	1.0	1.0	1.0
Tenure	1755.0	NaN	NaN	NaN	3.352137	9.754933	0.0	0.0	1.0	1.0	99.0
City_Tier	1786	3	Tier_1	998	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CC_Contacted_LY	1784.0	NaN	NaN	NaN	19.448991	8.937226	4.0	12.0	17.0	26.0	43.0
Payment	1780	5	Debit Card	671	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	1781	2	Male	1135	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Service_Score	1785.0	NaN	NaN	NaN	2.904762	0.703056	2.0	2.0	3.0	3.0	4.0
Account_user_count	1737.0	NaN	NaN	NaN	3.912493	1.014497	1.0	3.0	4.0	5.0	6.0
account_segment	1783	5	Regular Plus	1119	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CC_Agent_Score	1780.0	NaN	NaN	NaN	3.403371	1.336221	1.0	3.0	3.0	5.0	5.0
Marital_Status	1771	3	Single	923	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rev_per_month	1683.0	NaN	NaN	NaN	6.975639	13.706799	1.0	3.0	5.0	7.0	138.0
Complain_ly	1746.0	NaN	NaN	NaN	0.549255	0.497711	0.0	0.0	1.0	1.0	1.0
rev_growth_yoy	1801.0	NaN	NaN	NaN	16.139922	3.90048	4.0	13.0	15.0	19.0	28.0
coupon_used_for_payment	1801.0	NaN	NaN	NaN	1.524153	1.554777	0.0	1.0	1.0	2.0	15.0
Day_Since_CC_connect	1741.0	NaN	NaN	NaN	3.104538	3.149072	0.0	1.0	2.0	4.0	46.0
cashback	1718.0	NaN	NaN	NaN	178.043871	188.742912	110.09	133.48	151.84	173.325	1997.0
Login_device	1676	2	Mobile	1095	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Clus_Profile	1801.0	NaN	NaN	NaN	2.0	0.0	2.0	2.0	2.0	2.0	2.0

As indicated above:

- *People from Tier 3 City seems to be churning the most*
- *Customers who had to contact customer service more frequent around 15-25 times seems to be churning, this implies that they were not satisfied with the service provided.*
- *Accounts with around 4 users seems to be churning the most, which implies company can think of limiting the number of users per accounts*
- *Consumers who have received less coupons seems to be potential churners*

Conclusion:

• Improvement of infrastructures in Tier 3 cities to provide better customer service.

• Run offers for the customers who are using Regular Plus accounts in terms of more cashback coupons, or add channel counts.