Ecommerce: Customer Churn Prediction

<u>Developing Churn Prediction Model and provide Business Recommendations on Campaign</u>

<u>Capstone Project Report</u>

<u>Submitted to</u>

GREAT LAKES
INSTITUTE OF MANAGEMENT

Submitted by

G.R. Krishnaraj

Under the Guidance of

Ms. Keya Choudhury Ganguli

Batch –PGP-DSBA (May 2021)

Year of completion (June 2022)

Table of contents:

1.Introdution

1.1 Objective of the study:

An E-com company is facing a challenge in the sector due to presence of rival companies offering high attractive subscription plans to is customer base. Due to this to this company is facing a serious challenge regarding customer retentions

Hence, the company wants to develop model through which they can do churn prediction of the accounts and provide segmented offers to the potential churners.

1.2 In-scope

The company to stay afloat needs to assess its options, create strategies through which its able to retain its customer ad also to over time attract more customers and into the platform and so to maintain positive year on year growth

1.3 Out of scope

The major focus on maximum customer retention onto the platform which in turn will lead to more sales which in turn leads to higher profits for the company

Secondly once the retention is achieved, having developed a big happy customer base it's a matter of tome that these happy customers are via word of mouth going to bring new customer on to the Platform

Working on every aspect and removing the obstacles which are leading to customer churns is going help the company reap profits in every way

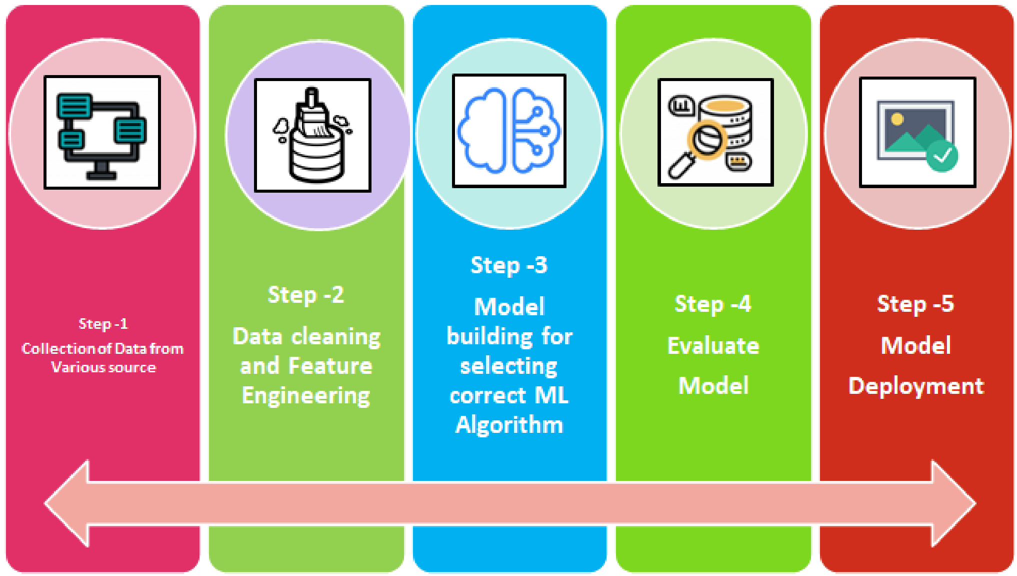
1.4 Tools and techniques Used

- **Used python version:3.10**
- **Used Microsoft excel for our data gathering and data cleaning**
- **Tableau for visualization**
- **Used the following visualization form python for better visualization and building models**
  `NumPy, pandas, matplotlib, pyplot, seaborn etc.`

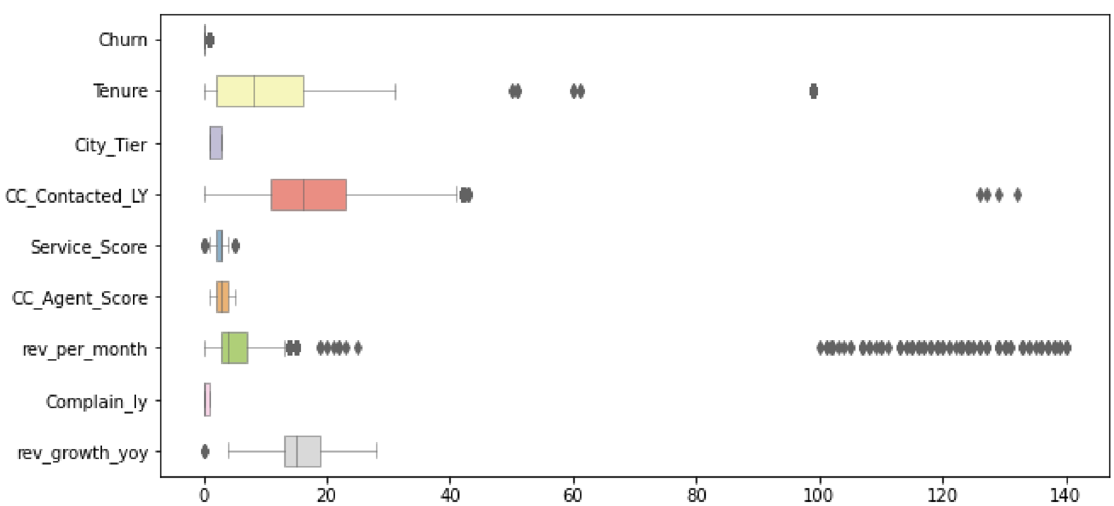
1.5 Analytical Approach

1. **Data extraction form data sources**
2. **Data cleaning and data preparation**
3. **Study of each variable by exploring the data**
4. **Study the variables for its relevance for the study**
5. **Identifying Y variable**
6. **Performing univariate analysis for all variables**
7. **Performing Bi-variant and Multi variate analysis for the variables required**
8. **Division of data into train and test**
9. **Model development**

10. **Model validation and model validation on test data**
11. **Invention strategies and recommendation**



1.6: The outliner check before EDA for numerical datatypes:



1.7: Table1: List of variables:

Variable	Description
AccountID	account unique identifier
Churn	account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_L12m	How many times all the customers of the account has contacted customer care in last 12months
Payment	Preferred Payment mode of the customers in the account
Gender	Gender of the primary customer of the account
Service_Score	Satisfaction score given by customers of the account on service provided by company
Account_user_count	Number of customers tagged with this account
account_segment	Account segmentation on the basis of spend
CC_Agent_Score	Satisfaction score given by customers of the account on customer care service provided by company
Marital_Status	Marital status of the primary customer of the account
rev_per_month	Monthly average revenue generated by account in last 12 months
Complain_l12m	Any complaints has been raised by account in last 12 months
rev_growth_yoy	revenue growth percentage of the account (last 12 months vs last 24 to 13 month)
coupon_used_l12m	How many times customers have used coupons to do the payment in last 12 months
Day_Since_CC_connect	Number of days since no customers in the account has contacted the customer care
cashback_l12m	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

2 Data cleaning and transformation:

2.1: In Excel:

The data has categorical and numerical type columns,

It has 11260 Rows and 18 columns

On further inspection we observed the presence of NAN values, blank spaces special characters within columns, which must be treated

2.2 In-Python

Iteration of feature engineering:

1.The outliner treatment was checked, and the missing value treatment is done

```
cashback                  0.041829
Day_Since_CC_connect      0.031705
Complain_ly               0.031705
Login_device              0.019627
Marital_Status            0.018828
CC_Agent_Score            0.010302
Account_user_count        0.009947
City_Tier                 0.009947
Payment                   0.009680
Gender                    0.009591
Tenure                    0.009059
rev_per_month             0.009059
CC_Contacted_LY           0.009059
Service_Score             0.008703
account_segment           0.008615
rev_growth_yoy            0.000000
coupon_used_for_payment   0.000000
Churn                     0.000000
dtype: float64
```

missing vale mv>2% <10 % impute by mode

2.We first treated each column with outliner treatment using quantile method

For the transformation we follow the below rule:

For categorical data:

if mv<2% --- mode

if mv>2%--- logical /mode

if mv>10 %---some new category

if mv>40%--- don't impute

For numerical data:

if mv<2% --- median

if mv>2%--- logical /median

if mv>10 %---regression

if mv>40%--- don't impute

3.The special characters and n nan values are replaced for the below columns

Tenure, CC_Contacted_LY, rev_per_month, rev_growth_yoy, cashback, Account_user_count

For the categorical values the nan is replaced by mode

For the numeri values the nan are replaced by median

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Churn	11260.0	NaN	NaN	NaN	0.168384	0.374223	0.0	0.0	0.0	0.0	1.0
Tenure	11158	38	1	1351	NaN	NaN	NaN	NaN	NaN	NaN	NaN
City_Tier	11148.0	NaN	NaN	NaN	1.653929	0.915015	1.0	1.0	1.0	3.0	3.0
CC_Contacted_LY	11158.0	NaN	NaN	NaN	17.867091	8.853269	4.0	11.0	16.0	23.0	132.0
Payment	11151	5	Debit Card	4587	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	11152	4	Male	6328	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Service_Score	11162.0	NaN	NaN	NaN	2.902526	0.725584	0.0	2.0	3.0	3.0	5.0
Account_user_count	11148	7	4	4569	NaN	NaN	NaN	NaN	NaN	NaN	NaN
account_segment	11163	7	Super	4062	NaN	NaN	NaN	NaN	NaN	NaN	NaN
CC_Agent_Score	11144.0	NaN	NaN	NaN	3.066493	1.379772	1.0	2.0	3.0	4.0	5.0
Marital_Status	11048	3	Married	5860	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rev_per_month	11158	59	3	1746	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Complain_ly	10903.0	NaN	NaN	NaN	0.285334	0.451594	0.0	0.0	0.0	1.0	1.0
rev_growth_yoy	11260	20	14	1524	NaN	NaN	NaN	NaN	NaN	NaN	NaN
coupon_used_for_payment	11260	20	1	4373	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Day_Since_CC_connect	10903	24	3	1816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cashback	10789	321	152	208	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Login_device	11039	3	Mobile	7482	NaN	NaN	NaN	NaN	NaN	NaN	NaN

2.3. Checking Multicollinearity & VIF

```
X_train = X_train.drop(["Login_device","Service_Score","Account_user_count","cashback","rev_growth_yoy"], axis=1)
X_test = X_test.drop(["Login_device","Service_Score","Account_user_count","cashback","rev_growth_yoy"], axis=1)
```

The VIF is checked and columns with score above 5% are removed

5.once the outliner and extreme values removed, we did the scaling process

6. The label encoding for the categorical variables are done

```
Gender,Account_segment,Account_user_count,Login_device,City_Tier,Service_Scor
e,CC_Agent_Score,Marital_Status,Payment,Complain_ly,Login_device
Cashback
```

3.Exploarotry Data analysis:

3.1 Data Summary:

The Exploratory Data analysis OF each variable is done to get the mean median min max and quartiles

```
df['rev_per_month'].describe(include="all").T
```

```
count    11260.000000
mean         5.915631
std         11.598273
min          0.000000
25%          3.000000
50%          4.000000
75%          7.000000
max        140.000000
Name: rev_per_month, dtype: float64
```

```
df['Tenure'].describe(include="all").T
```

```
count    11260.000000
mean        10.811634
std         12.844640
min          0.000000
25%          2.000000
50%          8.000000
75%         16.000000
max         99.000000
Name: Tenure, dtype: float64
```

```
df['CC_Contacted_LY'].describe(include="all").T
```

```
count    11260.000000
mean        17.705240
std          8.974194
min          0.000000
25%         11.000000
50%         16.000000
75%         23.000000
max        132.000000
Name: CC_Contacted_LY, dtype: float64
```

```
df['Day_Since_CC_connect'].describe(include="all").T
```

```
count    11260.000000
mean         4.580995
std          3.649867
min          0.000000
25%          2.000000
50%          3.000000
75%          7.000000
max         47.000000
Name: Day_Since_CC_connect, dtype: float64
```

```
df["coupon_used_for_payment"] = df["coupon_used_for_payment"].
df['coupon_used_for_payment'].describe(include="all").T
```

```
count    11260.000000
mean         1.790142
std          1.969505
min          0.000000
25%          1.000000
50%          1.000000
75%          2.000000
max         16.000000
Name: coupon_used_for_payment, dtype: float64
```

```
df['cashback'].describe(include="all").T
```

```
count    11260.000000
mean       194.350178
std        175.107143
min          0.000000
25%        148.000000
50%        163.000000
75%        197.000000
max       1997.000000
Name: cashback, dtype: float64
```

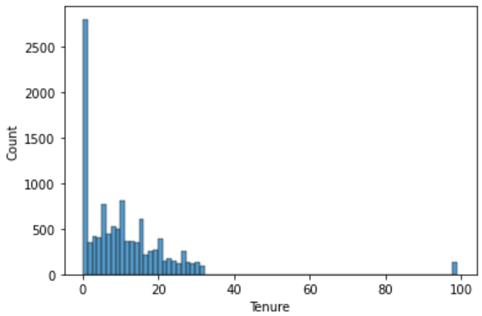
```
df['rev_growth_yoy'].describe(include="all").T
```

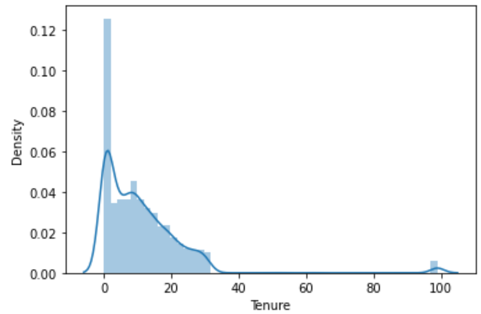
```
count    11260.000000
mean        16.189076
std          3.766505
min          0.000000
25%         13.000000
50%         15.000000
75%         19.000000
max         28.000000
Name: rev_growth_yoy, dtype: float64
```

3.2 Univariate anlysis using Pyton

The Histogram and boxplot are plotted to check the distribution of data

Tenure

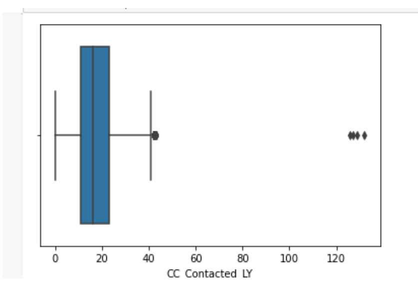




The Tenue of account holders in the rage 0 to 35 and presence of extreme values also present in the data

Here we can see the minimum is 0, the first quantile is 2 months. The median is 8 months, the q3 is 16 months and the highest is 99 months

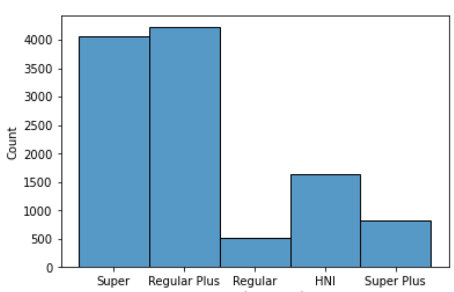
CC_Contacted_LY:




```
count    11260.000000
mean        17.705240
std          8.974194
min          0.000000
25%         11.000000
50%         16.000000
75%         23.000000
max        132.000000
Name: CC_Contacted_LY, dtype: float64
```

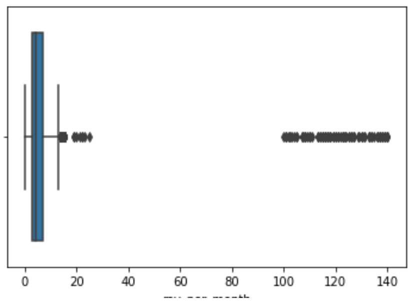
Here we can see the minimum is 0, the first quantile is 11 times. The median is 16 times, the q3 is 23 times and the highest is 132 times.

account_segment:



The maximum account user count is for super and regular + which are > 4000 counts.

rev_per_month:



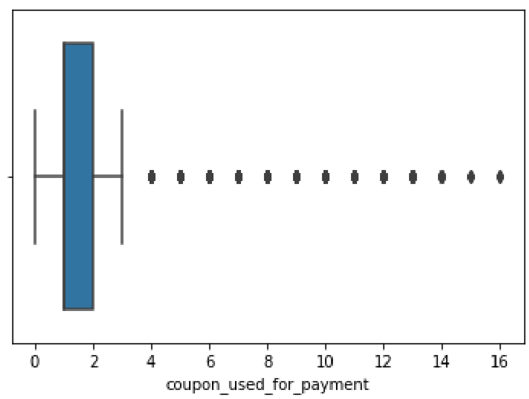

```
count    11260.000000
mean         5.915631
std         11.598273
min          0.000000
25%          3.000000
50%          4.000000
75%          7.000000
max        140.000000
Name: rev_per_month, dtype: float64
```

Here we can see the minimum is 0, the first quantile is 3. The median is 5, the q3 is 7 and the highest is 140

This gives us an insight that the spending capacity is highly diverse among the customer base.

coupon_used_for_payment:

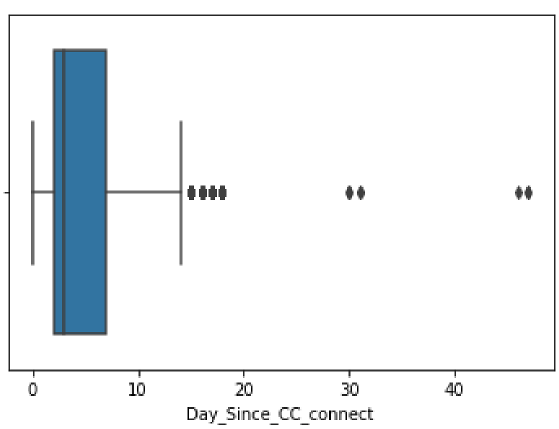
```
count    11260.000000
mean         1.790142
std          1.969505
min          0.000000
25%          1.000000
50%          1.000000
75%          2.000000
max         16.000000
Name: coupon_used_for_payment, c
```



Here we can see the minimum is 0, the first quantile is 1 The median is 1, the q3 is 2 and the highest is 16

Most customers have used 2 coupons. this seems to be an area of improvement; we would like this to be more balanced.

Day_Since_CC_connect:

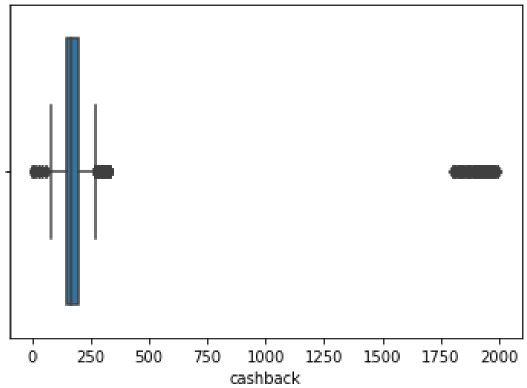



```
count    11260.000000
mean         4.580995
std          3.649867
min          0.000000
25%          2.000000
50%          3.000000
75%          7.000000
max         47.000000
Name: Day_Since_CC_connect, dtype: float64
```

Here we can see the minimum is 0, the first quantile is 2, The median is 3, the q3 is 7 and the highest is 40+

We would like lesser people contacting the customer care

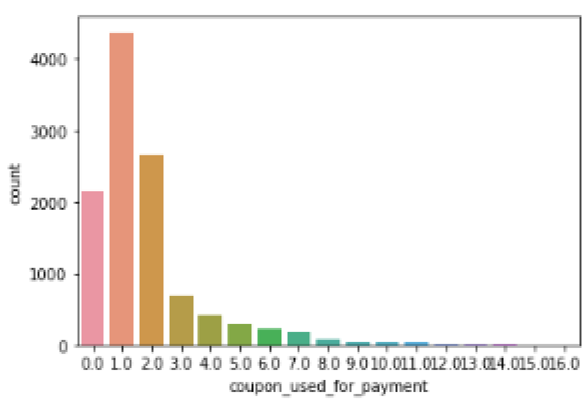
Cashback




```
count    11260.000000
mean       194.350178
std        175.107143
min          0.000000
25%        148.000000
50%        163.000000
75%        197.000000
max       1997.000000
Name: cashback, dtype: float64
```

customers having higher shopping frequency, who tend to earn more and more cashbacks on the platform

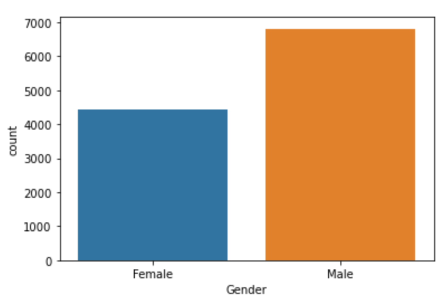
Coupons used for payment



Most no of customers used coupons of 1 & 2.

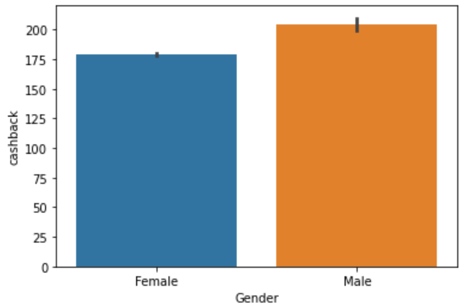
Bar plots

Gender:



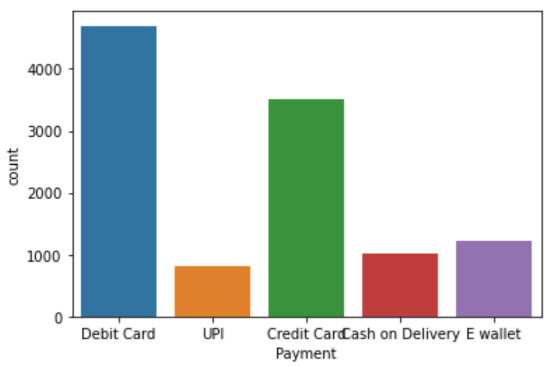
There is higher the male customers compare to female customers.

Cash back on Gender:



The female customers make more than half of the total customer base

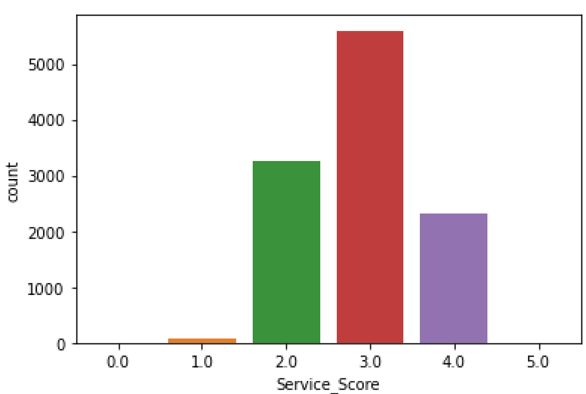
Payment bar plot

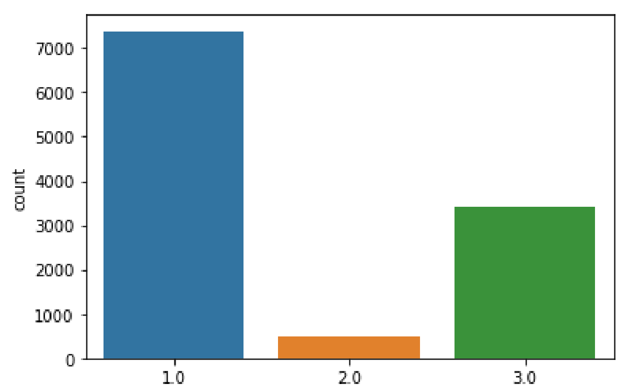


The debit card payment is high, next I credit card payment and the least is UPI

We can collaborate with banks and provide offers to customers on reward points and cashbacks

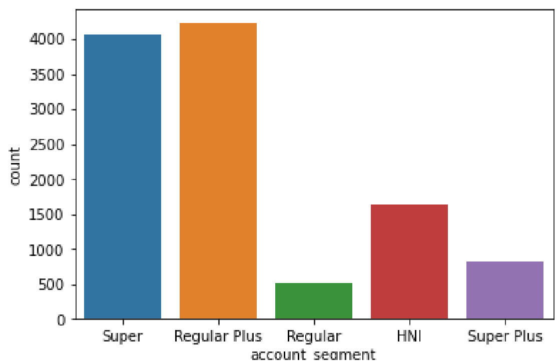
City_Tier: Bar plot





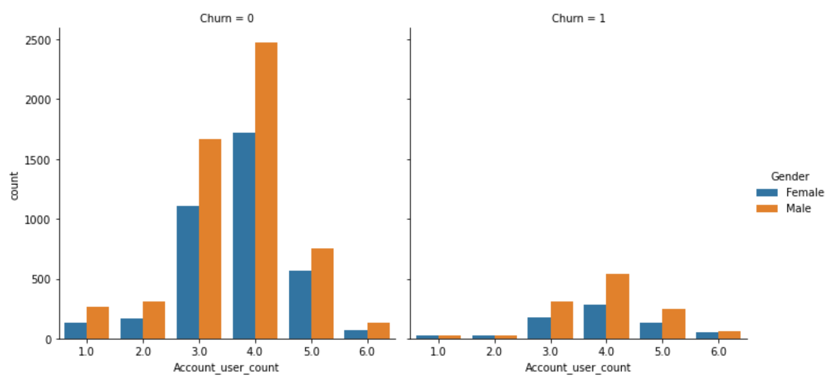
The most customer ate from tier 1 cities

The service score high for 2 ,3 and moderate for 4, and none for 5 account segment



we can see super and regular + having highest customer base

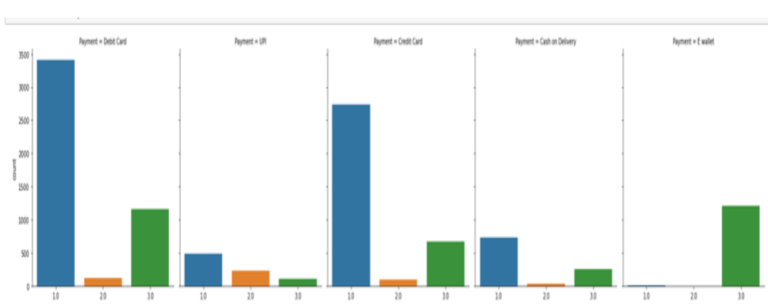
3.3 Bivariate anlysis using Pyton



The customers are churning out have low tenure on the platform, the retention strategy is increasing the tenure f customer base

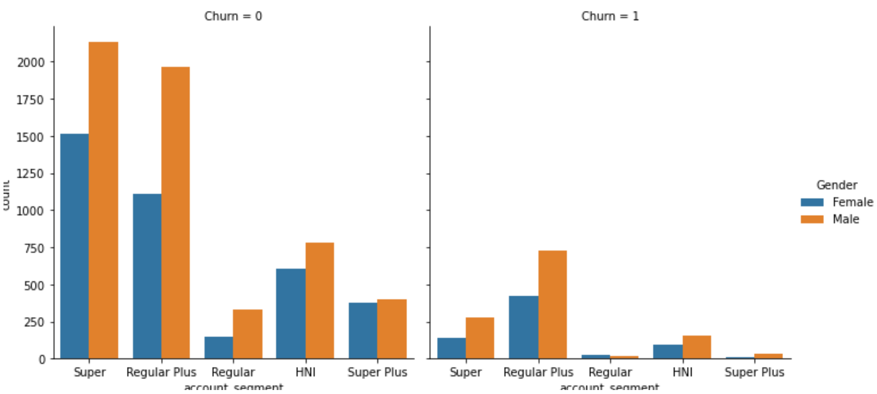
The customer cluster with low tenure must be identified and provided offers with their browsing history and order history

City, Payment's bar charts:



Highest is the tier 1 city with debit card payment, Least is tier 2 city with cash on delivery, and tier 1 & 2 with e wallet payment is null

Churns across subscription plans:



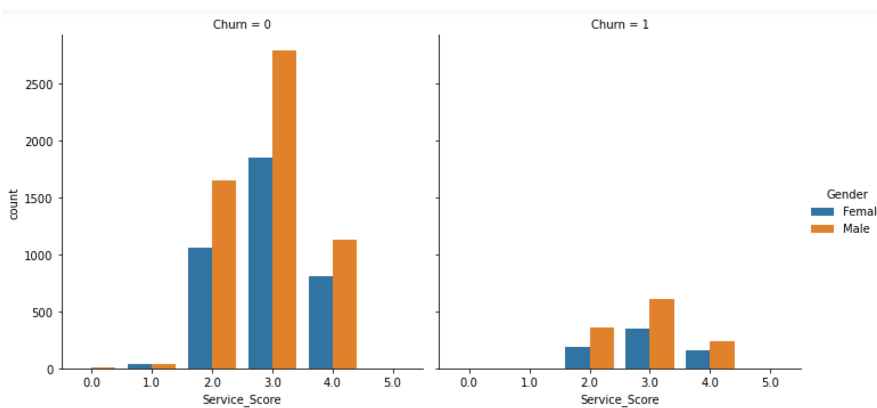
The debit card and debit card payment are high for our customer base, the churns for this payment method are also high

we can strategies such that the payment methods have good ROI Value, so they keep renewing the subscriptions

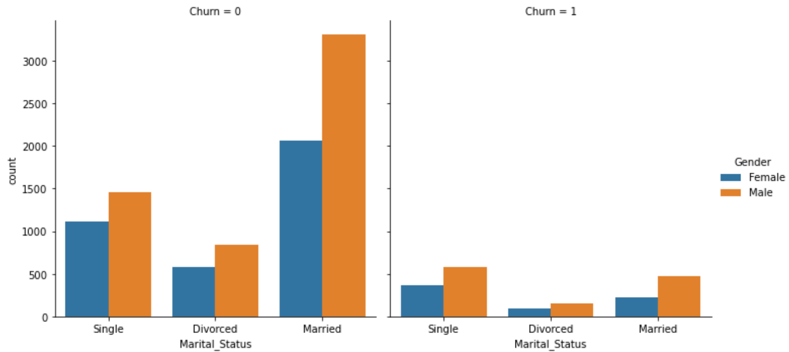
Churn across service score:

For all vales of service score we can see the churns, we can create a program where in customer care can highlight all al the area that the customer unhappy about

and we can prioritize and solve the problem in sequential manner



Churn across Marital Status:



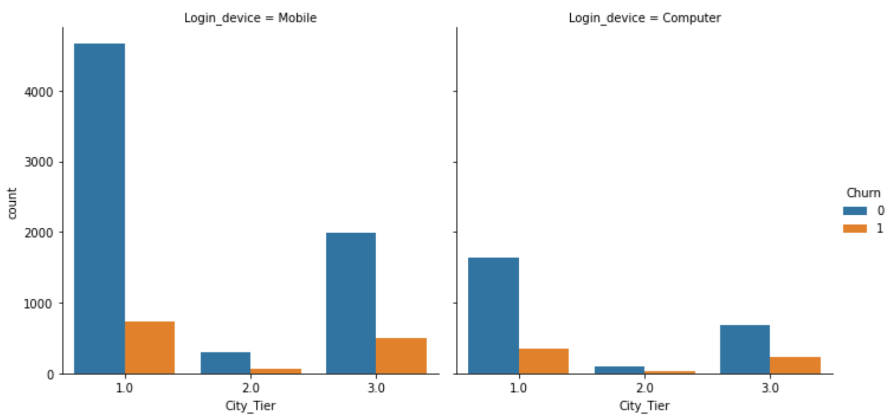
The chur is high for single customers, we need to understand their expectations and fulfill them

We need to focus on those group and next is married customers where the churn seems to be high,

Strategy can be devised to specially focus among these groups.

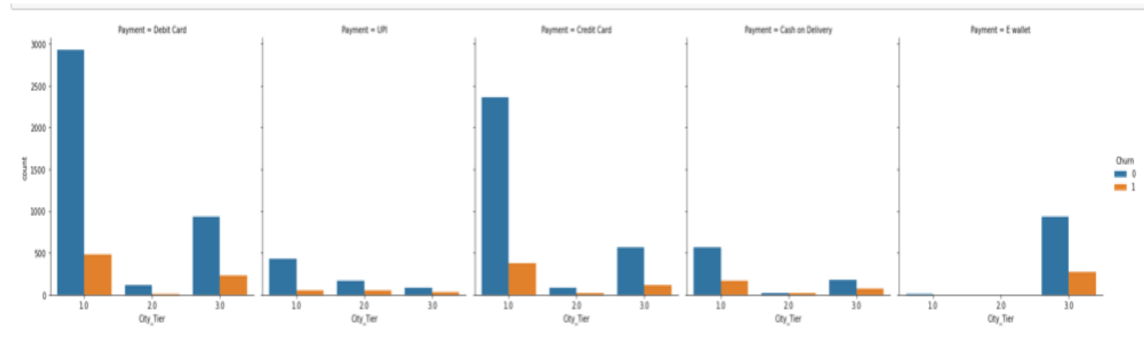
3.4 Multivariate anlysis using Pyton

Login device across city Tier with hue as Churn:



The login device is dependent on the population density of the city, for the mobile platform the preference is higher compared to computer

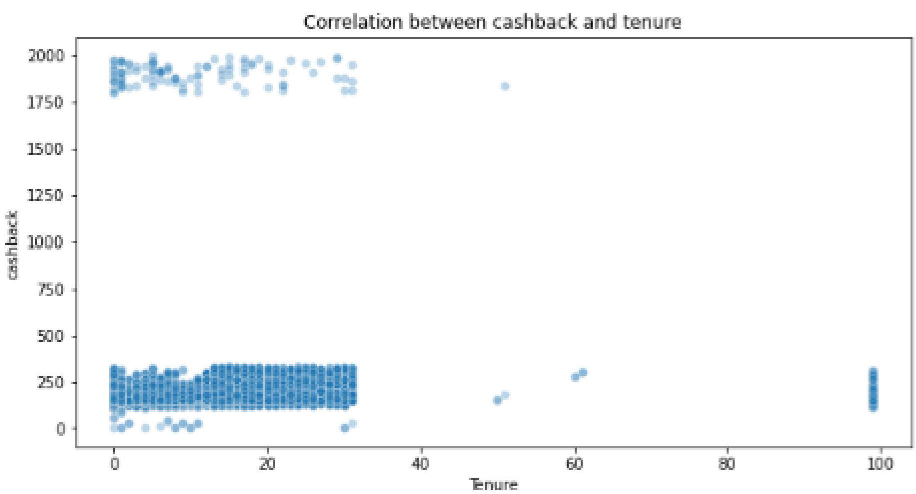
Payment method across city among with churn values:



For tier 2 city the payment is high on UPI method payment. the churn is highest for this type of payment methos

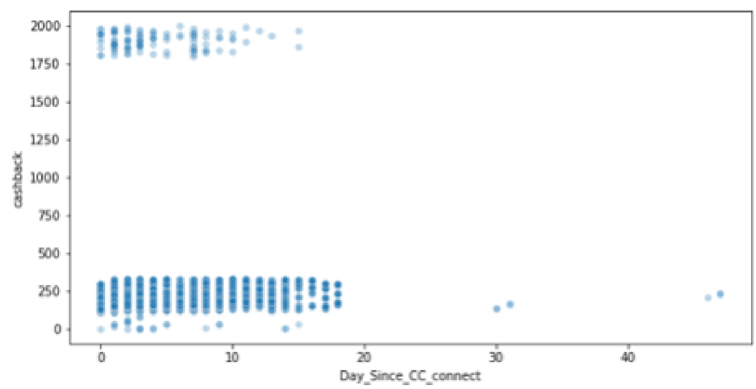
Also, it has to be duly noted and insights section we have suggested strategies which are built on these insights

Correlation among Tenure and cash back



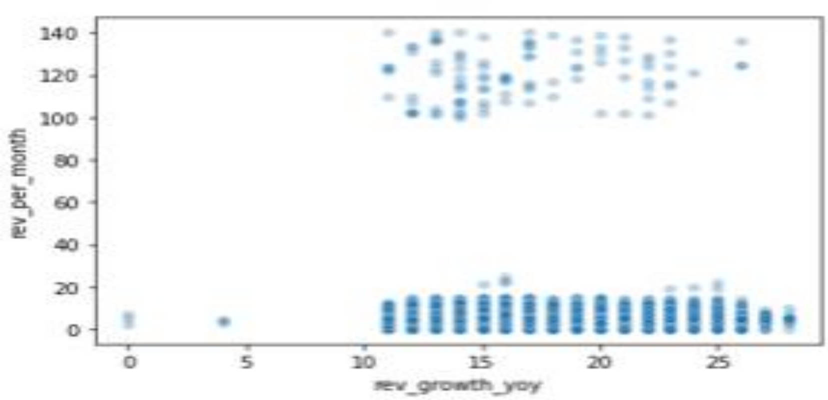
The customers in the range 0 to 30 have taken higher cash backs: Strong positive correlation

Correlation among day since connect and cash back



The customers in the range 0 to 20 contacted have taken higher cash backs: Strong positive correlation

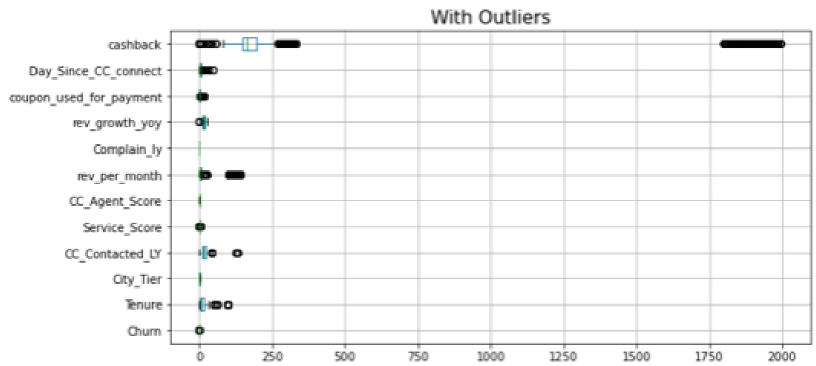
Correlation among revenue growth and revenue per month of customers:



It's a high positive correlation

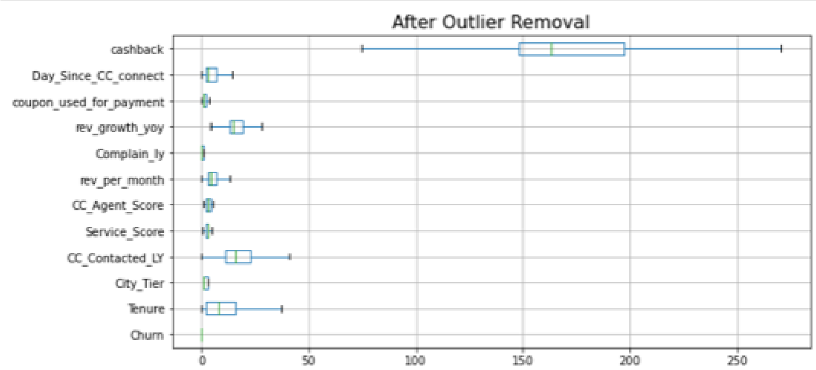
Outliner treatment for continuous variables:

Before outline treatment:



Outliners present in cashbacks and in revenue per month at the maximum

After outline treatment:



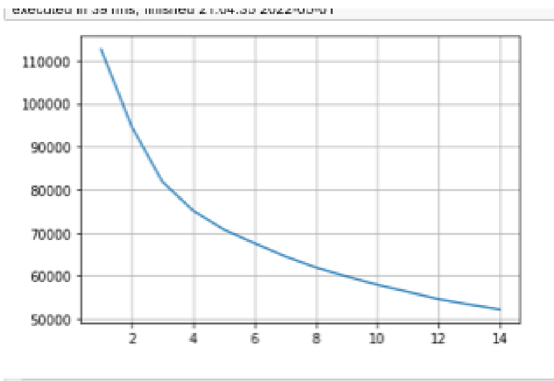
Correlation among continuous variables:

We can see higher correlation among cash backs and tenure, higher the cashbacks the customer tenure time increases.

Service score and coupon used- the service level of company increases as more offers and coupon provided to the c ustomers

3.5 Clustering: After removal of categorical variables:

WSS plots is made with 3 Clusters



3.6 Scaling is done by standard scaler

	Tenure	CC_Contacted_LY	Account_user_count	rev_per_month	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback	rev_growth_y
0	4.0	6.0	3.0	9.0	11.0	1.0	5.0	160.0	
1	0.0	8.0	4.0	7.0	15.0	0.0	0.0	121.0	
2	0.0	30.0	4.0	6.0	14.0	0.0	3.0	152.0	
3	0.0	15.0	4.0	8.0	23.0	0.0	3.0	134.0	
4	0.0	12.0	3.0	3.0	11.0	1.0	3.0	130.0	

	Tenure	CC_Contacted_LY	Account_user_count	rev_per_month	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback	rev_growth
0	-0.675327	-1.337224	-0.769057	1.290208	-1.379506	-0.430905	0.129952	-0.381191	-1.37
1	-1.119827	-1.108052	0.312917	0.660907	-0.316332	-1.337988	-1.301249	-1.274835	-0.31
2	-1.119827	1.412834	0.312917	0.346256	-0.582125	-1.337988	-0.442528	-0.564502	-0.58
3	-1.119827	-0.305952	0.312917	0.975558	1.810015	-1.337988	-0.442528	-0.976953	1.80
4	-1.119827	-0.649709	-0.769057	-0.597697	-1.379506	-0.430905	-0.442528	-1.068609	-1.37

]:

unt	rev_per_month	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback	rev_growth_yoy	coupon_used_for_payment	Kmeans_clusters
3.0	9.0	11.0	1.0	5.0	160.0	11	1	1
4.0	7.0	15.0	0.0	0.0	121.0	15	0	1
4.0	6.0	14.0	0.0	3.0	152.0	14	0	1
4.0	8.0	23.0	0.0	3.0	134.0	23	0	2
3.0	3.0	11.0	1.0	3.0	130.0	11	1	1

K means (1-3) cluster is added at the last columns

4.Model selection:

The objective of the model development is to predict the churn rate accounts for the given data set

As the target variable is churn, which is a categorical data, we can decide to create a classification model to predict

the churn rates, we will be implementing all the possible classification algorithm to identify the best fit model, belo w is list of possible algorithms which we may apply

- 1.Logistic regression:
- 2.CART
- 3.Random Forest
- 4. Linear Discriminant Analysis
- 5. K Nearest Neighbors
- 6. Gaussian Naive Bayes
- 7. Gradient Boosting
- 8)MLP Classifier (Artificial Neural Network)

Ensemble Model:

- Bagging model
- Random forest:

4.1 Evaluation parameters

- Accuracy: The measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset
- AUC: It's a statistical measure that we can use to evaluate the model predictions using a probabilistic framework.
- PRECISON: the quality of a positive prediction made by the model; Precision refers to the number of true positives divided by the total number of positive predictions
- RECALL: Recall is dependent on positive samples and independent of negative samples
- F1 SCORE: The F1 score conveys the balance between the precision and the recall.

4.2 Input into the model:

Prepared our predictor set to create input for our classification models

Removed following columns that are less significant /Gives duplicate info

Login device, service score, account user count, cash back rev growth yoy

Converted categorical data into factors

Converted categorical predictors into continuous predictors

Dummy Encoding: since target encoding lead to over fitting, we decide to perform dummy encoding on the rest of columns to convert them

Performed correlation lot to check multi collinearity

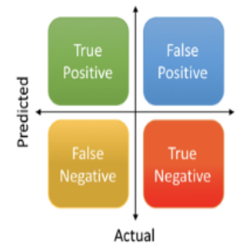
Split the data into train and test using 70:30 ratio

Performed scaling on predictors to reduce the variation

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



4.3: Train date Metrics:

executed in 197ms, finished 20:53:42 2022-05-15

	CART train	RF Train	Log Train	LDA Train	KNN Train	NB Train	ANN train	Gr.Boost Train
Accuracy	0.92	0.92	0.87	0.86	0.92	0.86	0.98	0.91
AUC	0.96	0.96	0.85	0.82	0.97	0.81	1.00	0.95
Recall	0.75	0.69	0.35	0.27	0.60	0.28	0.95	0.62
Precision	0.78	0.83	0.76	0.74	0.87	0.72	0.94	0.81
F1 Score	0.76	0.75	0.48	0.40	0.71	0.41	0.94	0.70

4.4: Test data metrics:

executed in 7 ms, finished 20:53:46 2022-05-15

	CART test	RF test	Log Test	LDA test	KNN test	NB test	ANN test	Gr.Boost test
Accuracy	0.90	0.90	0.87	0.87	0.90	0.86	0.95	0.90
AUC	0.94	0.94	0.85	0.82	0.93	0.81	0.97	0.94
Recall	0.70	0.63	0.36	0.98	0.52	0.29	0.82	0.58
Precision	0.70	0.79	0.77	0.87	0.80	0.67	0.87	0.77
F1 Score	0.70	0.70	0.49	0.92	0.63	0.41	0.85	0.66

4.5: Ensemble models

1)Bagging model

2)Random Forest

Random forest: Test data

```
0.9662522202486679
[[2777   32]
 [  82  487]]
              precision    recall  f1-score   support

           0       0.97      0.99      0.98      2809
           1       0.94      0.86      0.90       569

    accuracy                           0.97      3378
   macro avg       0.95      0.92      0.94      3378
weighted avg       0.97      0.97      0.97      3378
```

Bagging model: Test data

```
0.9683244523386619
[[2767   42]
 [  65  504]]
              precision    recall  f1-score   support

           0       0.98      0.99      0.98      2809
           1       0.92      0.89      0.90       569

    accuracy                           0.97      3378
   macro avg       0.95      0.94      0.94      3378
weighted avg       0.97      0.97      0.97      3378
```

4.6: ANN & Decision Tree: After Outliner treatment Ann & Decision Tree: After Outliner treatment

F1 SCORE RF: 98

ANN & Decision Tree: After Outliner treatment

Models	Test set score accu	train test score accu	test auc score	train auc score
Decion tree	0.897	0.921	0.937	0.958
Decion tree-Outliner treated/Scaled	0.897	0.921	0.935	0.958
Neural Networks(ANN)	0.949	0.981	0.973	0.997
Neural Networks(ANN)-Outliner treated/Scaled	0.953	0.986	0.975	0.998

After outliner treatment We can see a significant raise in the accuracy and in auc score for test and train

Out of the 8 models we have designed, recommend the Machine learning model Artificial neural network (ANN)

The model has responded well to our data and performed well with a score of F1 Score=.98(test), AUC- .97(test)

The ANN model is adaptive and with high accuracy, and it keeps on learning, in our case we can't be dependent on a static Situation because to be accurate all the time as the customer behavior keeps on changing

4.7: Conclusion & Recommendations

Columns	Silver	Gold	Platinum
	Cluster 2	Cluster 1	Cluster 3
revenue growth percentage	13.98	15	20
Coupon used for payment	0.96	2.96	1.4
Monthly average revenue	4.5	5	5.12
Tenure of account	8.6	14.6	10
CC Contacted LY 12 months	17	17	18

Recommendation cluster 1: Silver (New onboards)

The revenue growth of this base must be increased by proving more offers and benefits,

They must be moved to next level of cluster gold

They are new customers with high probability of Churn and with low tenure for them attractive offers must be provided

coupons usage must be increased with coupons floating delas with other

companies in the market

Recommendation cluster 2: Gold (Customers to be moved to platinum base)

- The revenue growth of this base must be increased by proving more offers and benefits
- They must be moved to next level of cluster platinum
- They have high tenure period, it must be sustained

Recommendation cluster 3: platinum (premiums customers to be protected from competitions)

- They have high revenue growth, which should be further increased
- They are using other type of payment hence coupon usage by them should be increased
- They have been in system for moderate period, which should be maintained and increased by promo offers and other benefit
- Their customer care complains must be addressed immediately and problems must be solved immediately

2)From accessing tenure, we could see a lot of new customers are not the platform

- This Is a good opportunity for the company to increase the customers
- The focus rear is retaining them with in the platform by understanding what the customer feels is valuable

3) the people contacting customer care is high

- This is important insight; we need to start accessing areas where complaint arising and investigate it and closing it & permanently help in more customer retention

Thank you