greatlearning

Learning for Life

CAPSTONE PROJECT NOTES 1

CUSTOMER CHURN (DTH PROJECT

Swarnava Das (Batch 3)

CUSTOMER CHURN DTH

TABLE OF CONTENTS

CUSTOMER CHURN DTH

1) Introduction of the business problem

- Problem statement.

An E Commerce company or DTH (you can choose either of these two domains) provider is facing a lot of competition in the current market and it has become a challenge to retain the existing customers in the current situation. Hence, the 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 1 account can have multiple customers. Hence by losing one account the company might be losing more than one customer.

- Need of the project.

To prevent the company from loosing its customers which may be because of a number of reasons by launching campaigns, giving offers etc. This way the company will retain most of its customers and get its revenue increased.

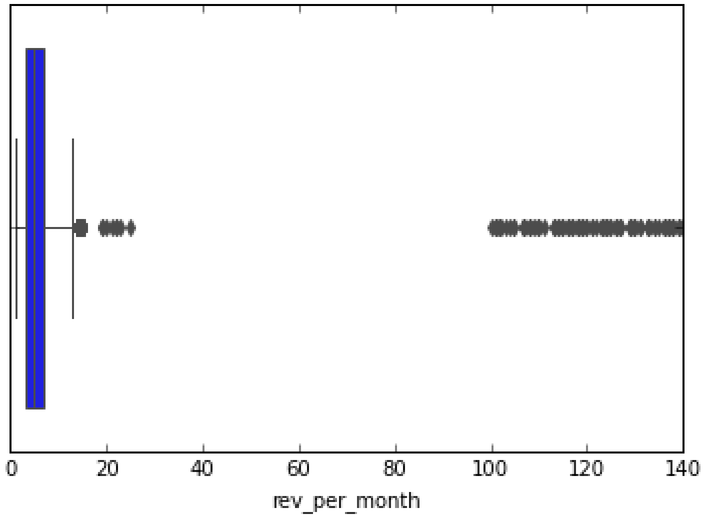
- Understanding business.

From the data we get to understand many factors, which may be of significant importance of leading a customer to churn from the DTH company.

3) Exploratory data analysis

Let us have look at how the continuous variables are distributed.

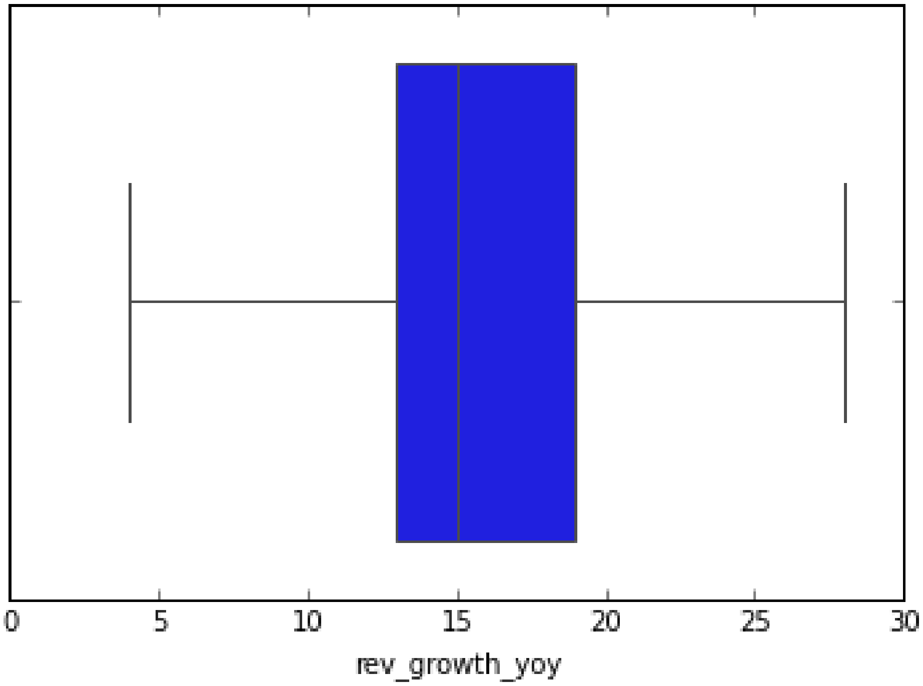
Revenue growth per month :



CUSTOMER CHURN DTH

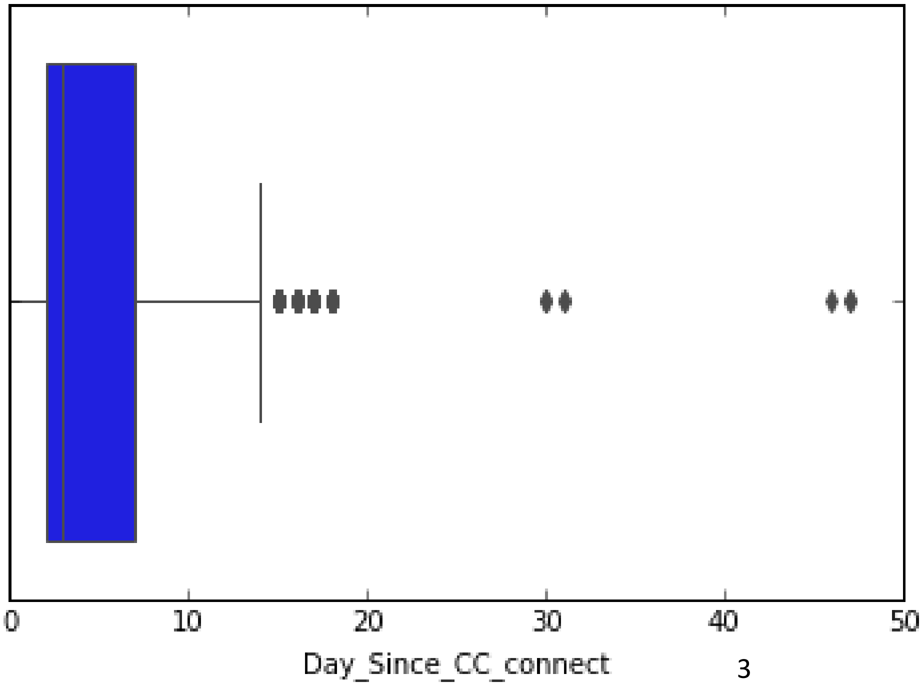
All the values are generally on the lower side if outliers are not considered.

Revenue growth over the past 2 years :



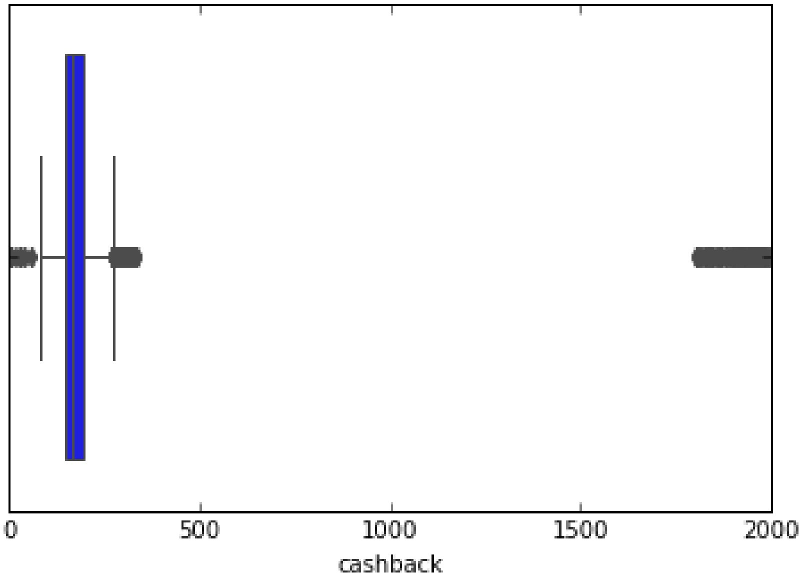
As seen it is almost normally spread.

Day since last customer contacted customer care :



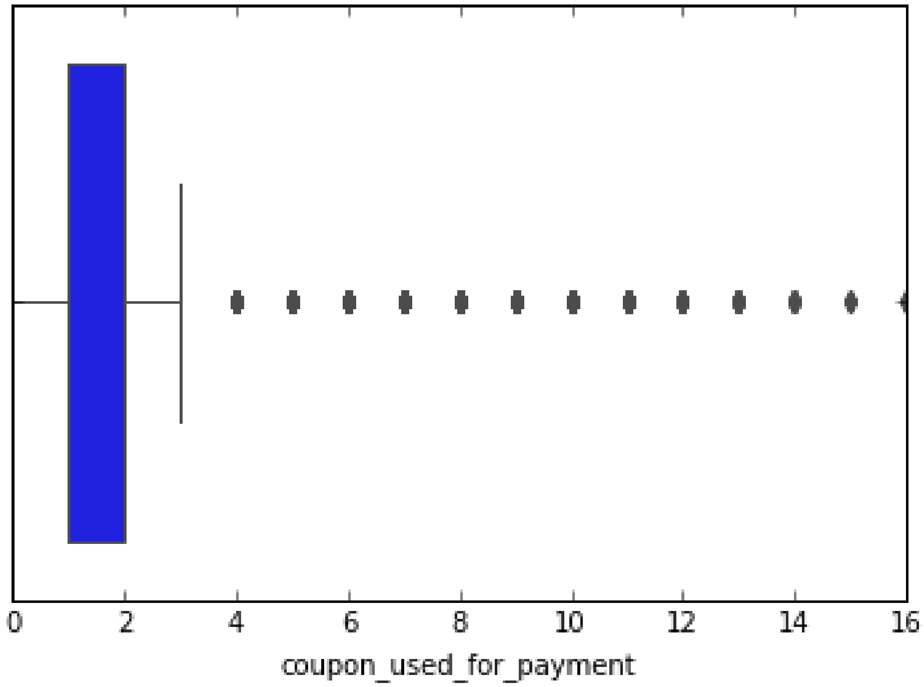
It seems to have mainly lower values and some outliers on the higher side.

Cashback received :



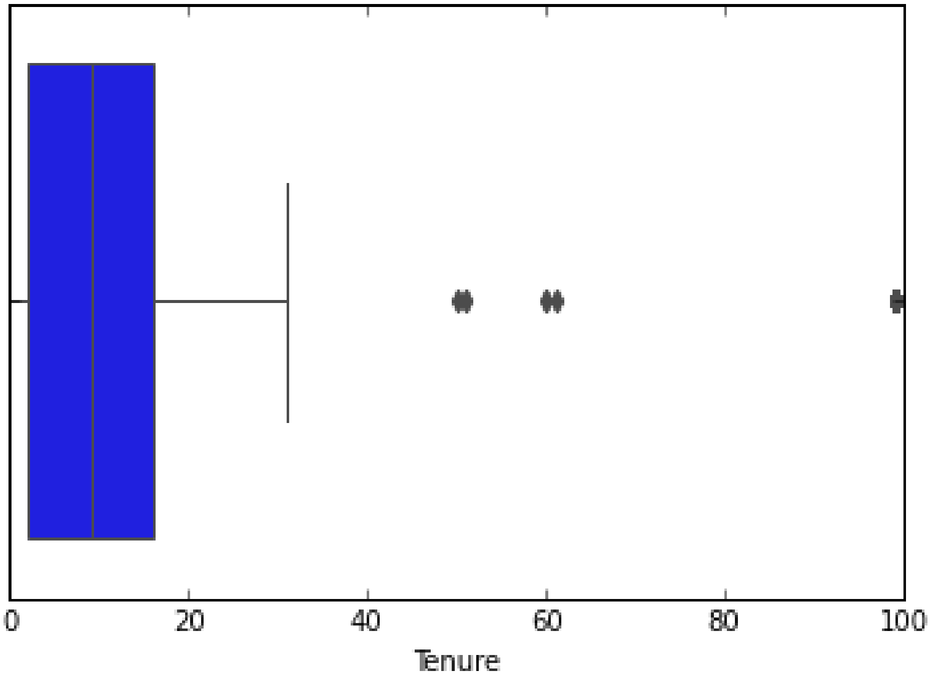
Cashback received by customers also have outliers and mainly have most of its values on the lower side.

Number of times coupon used by customer :

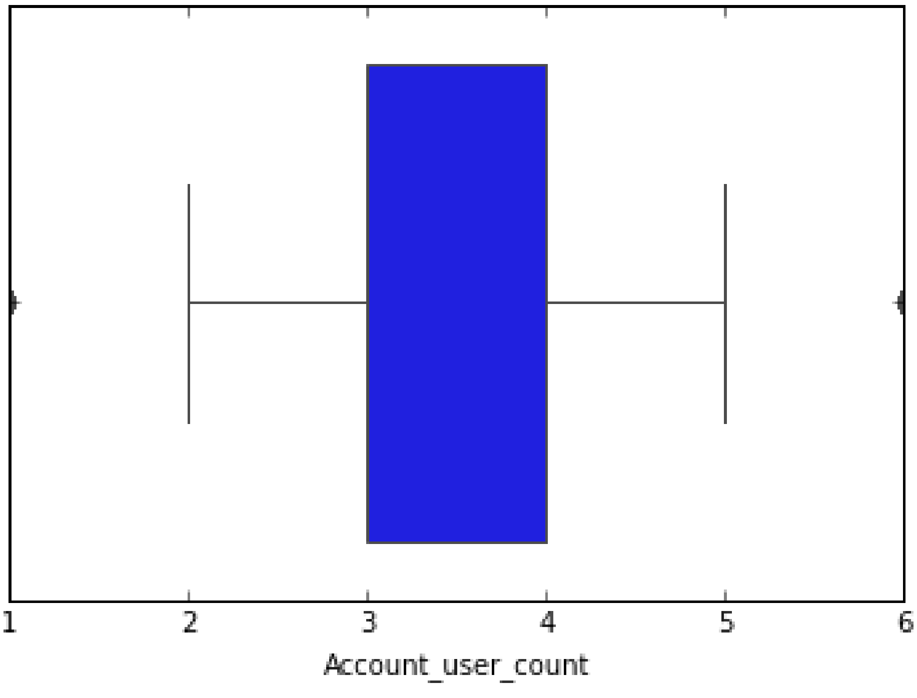


Tenure of customers :

CUSTOMER CHURN DTH

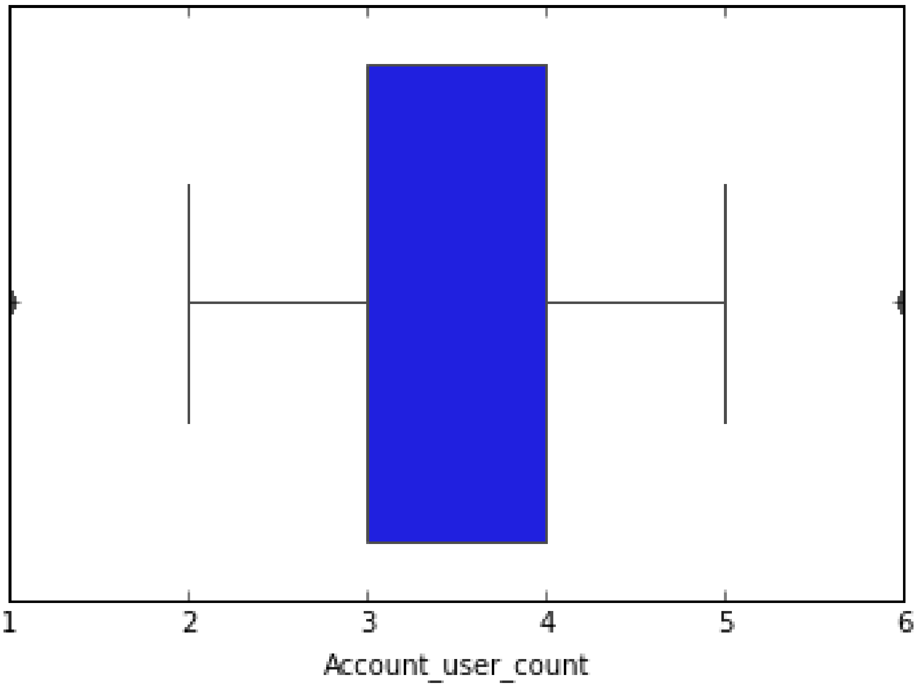


No of users tagged to an account :



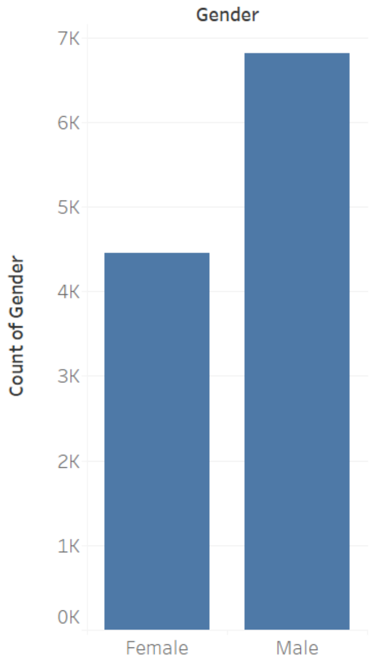
No outliers here and number of users tagged to each account is more or less normally distributed.

Number of times customers contacted customer support in the past year.



Lets see the distributions for categorical variables :

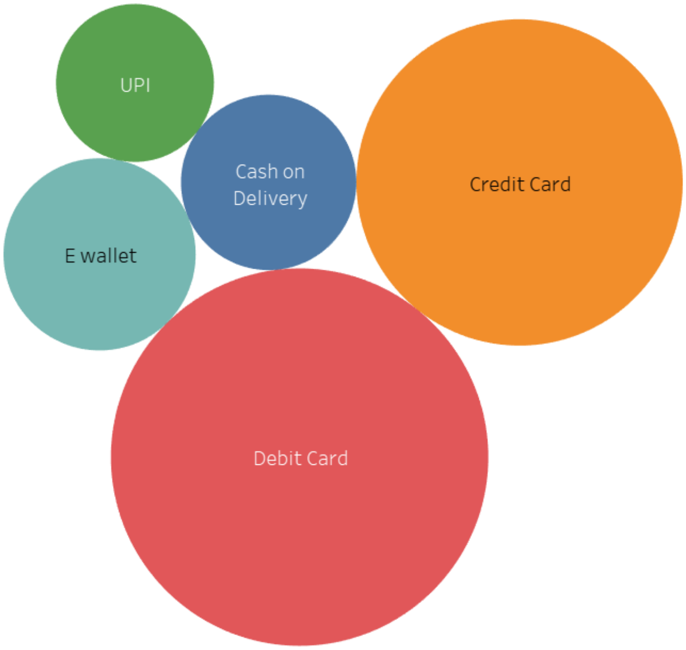
Gender distribution :



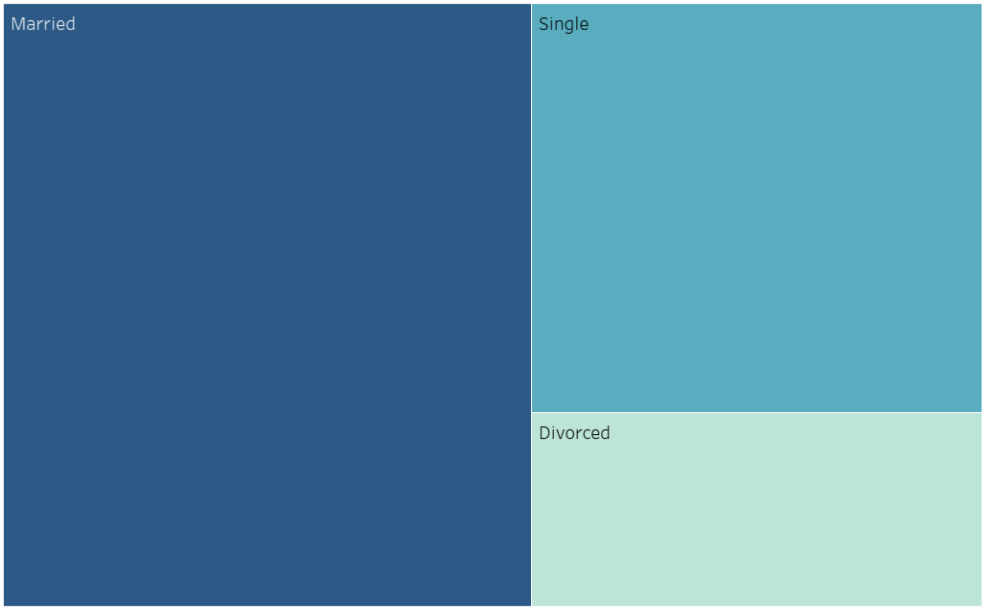
CUSTOMER CHURN DTH

It’s a male dominated customership of the DTH company.

Payment methods used :

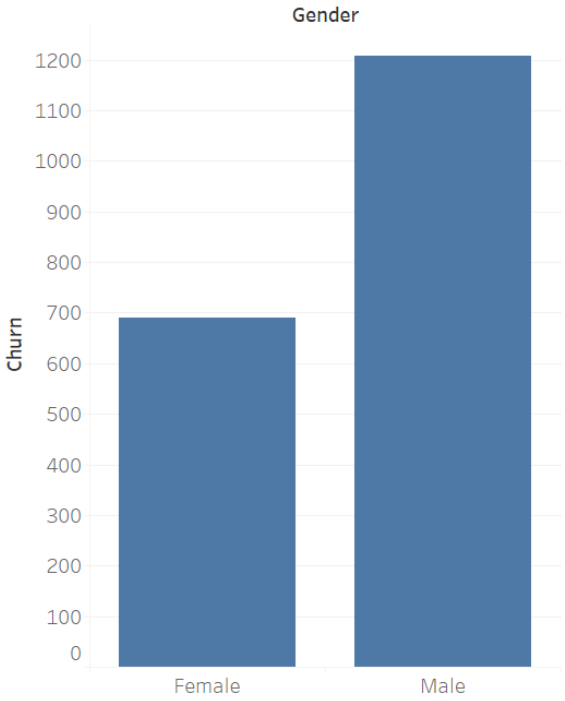


Marital status :

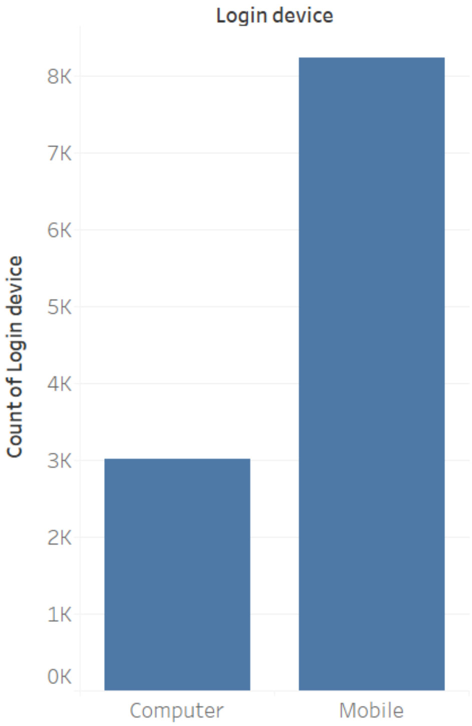


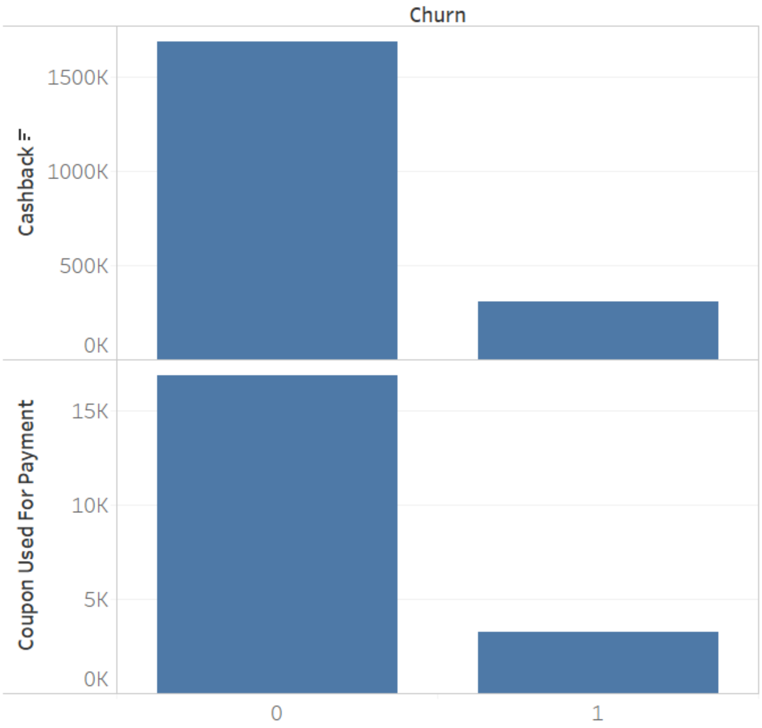
CUSTOMER CHURN DTH

Male customers seem to churn more.



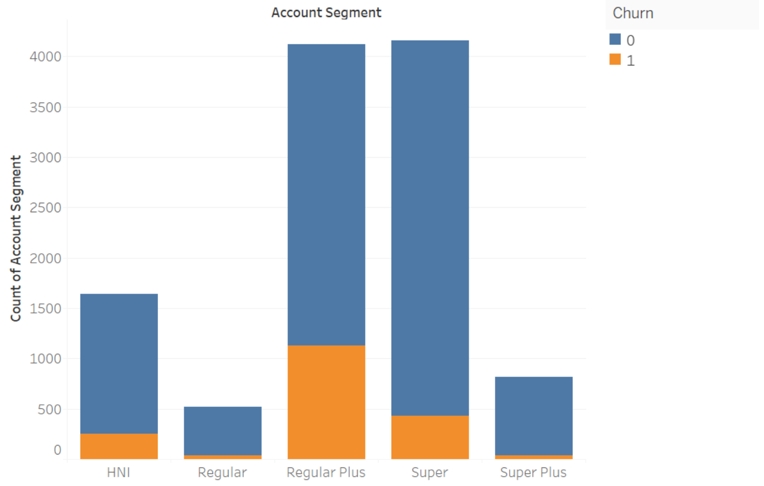
Most of the DTH company customers login by mobile.

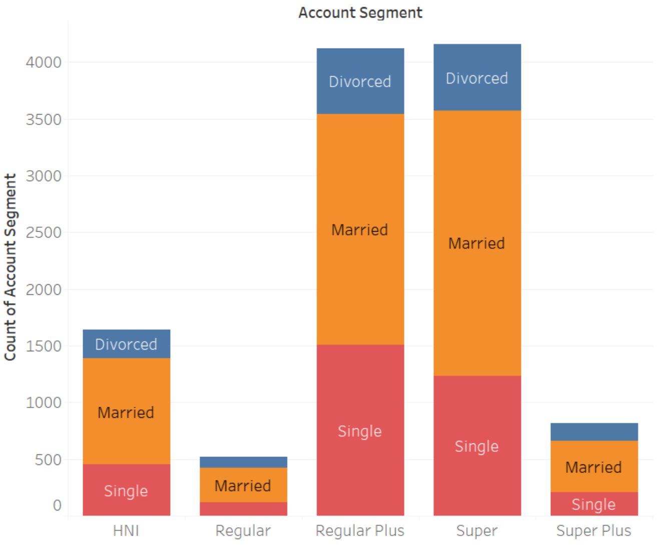




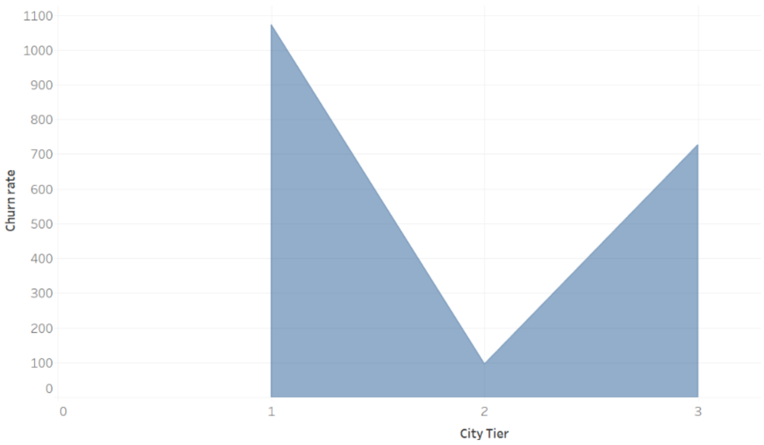
Customers availing offers and coupons tend to churn less.

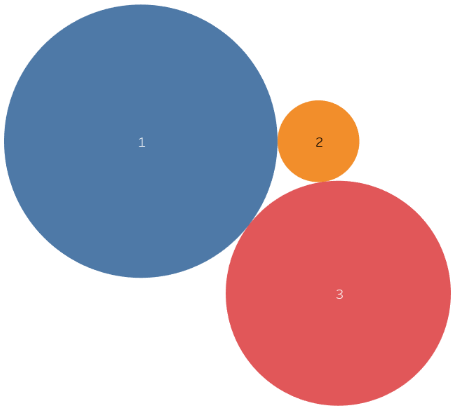
CUSTOMER CHURN DTH





Super plus segment customers churn the least and regular plus segment customers tend to churn the most. Married couple in all kinds of account segments churn the most.

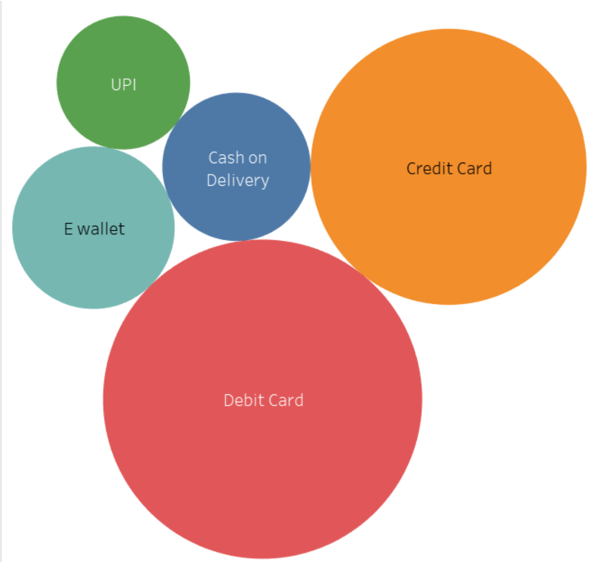


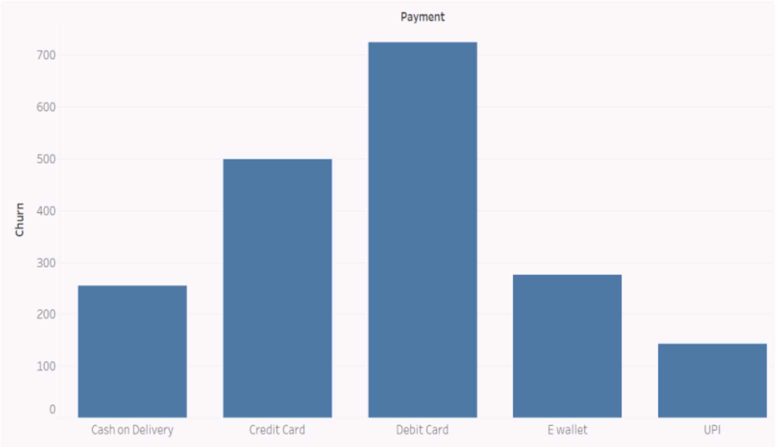


Tier 1 city has the most population and people from tier 1 cities churn the most

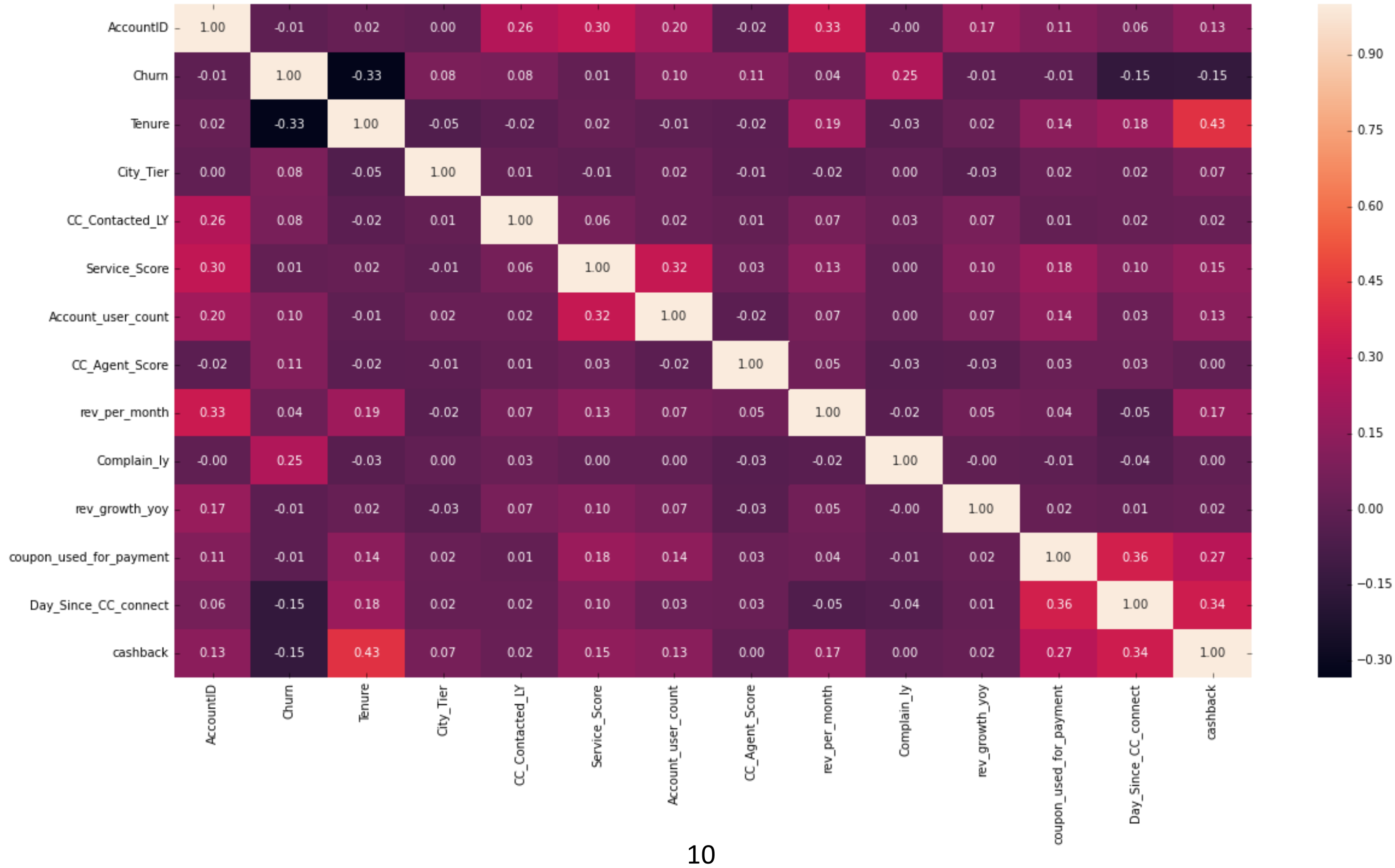
CUSTOMER CHURN DTH

Most of the customers use debit and credit cards as their payment methods. Customers using Debit card as their preferred mode of payment churns the most.



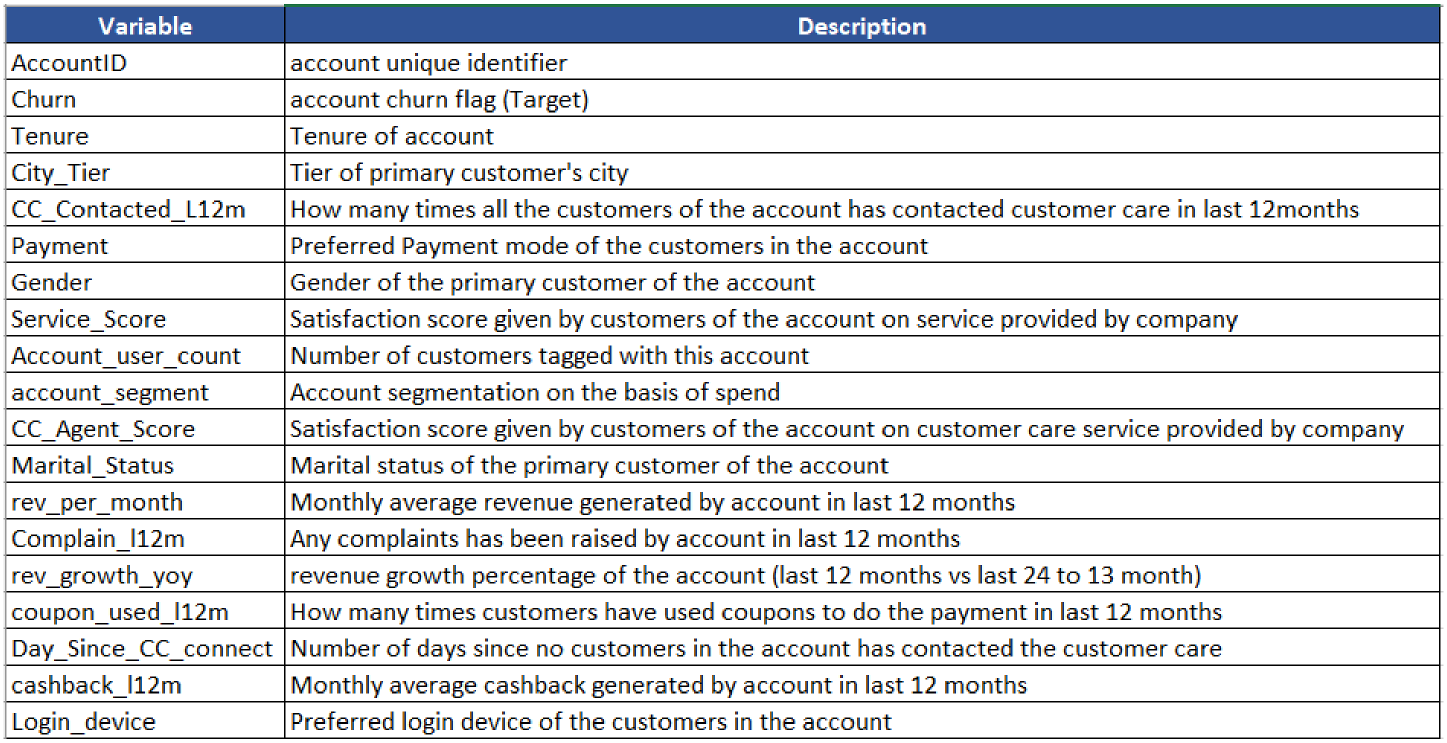


Lets look at the correlation heatmap which gives us an over all view of the linear relationship of all continuous variables .



2) Data cleansing / Data Report

The dataset contains the below variables along with the description of each.



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

As seen , the data has been collected over a period of a year to understand the patterns , revenue obtained from different accounts of customers.

Over here we will work on predicting the target variable : ‘Churn’

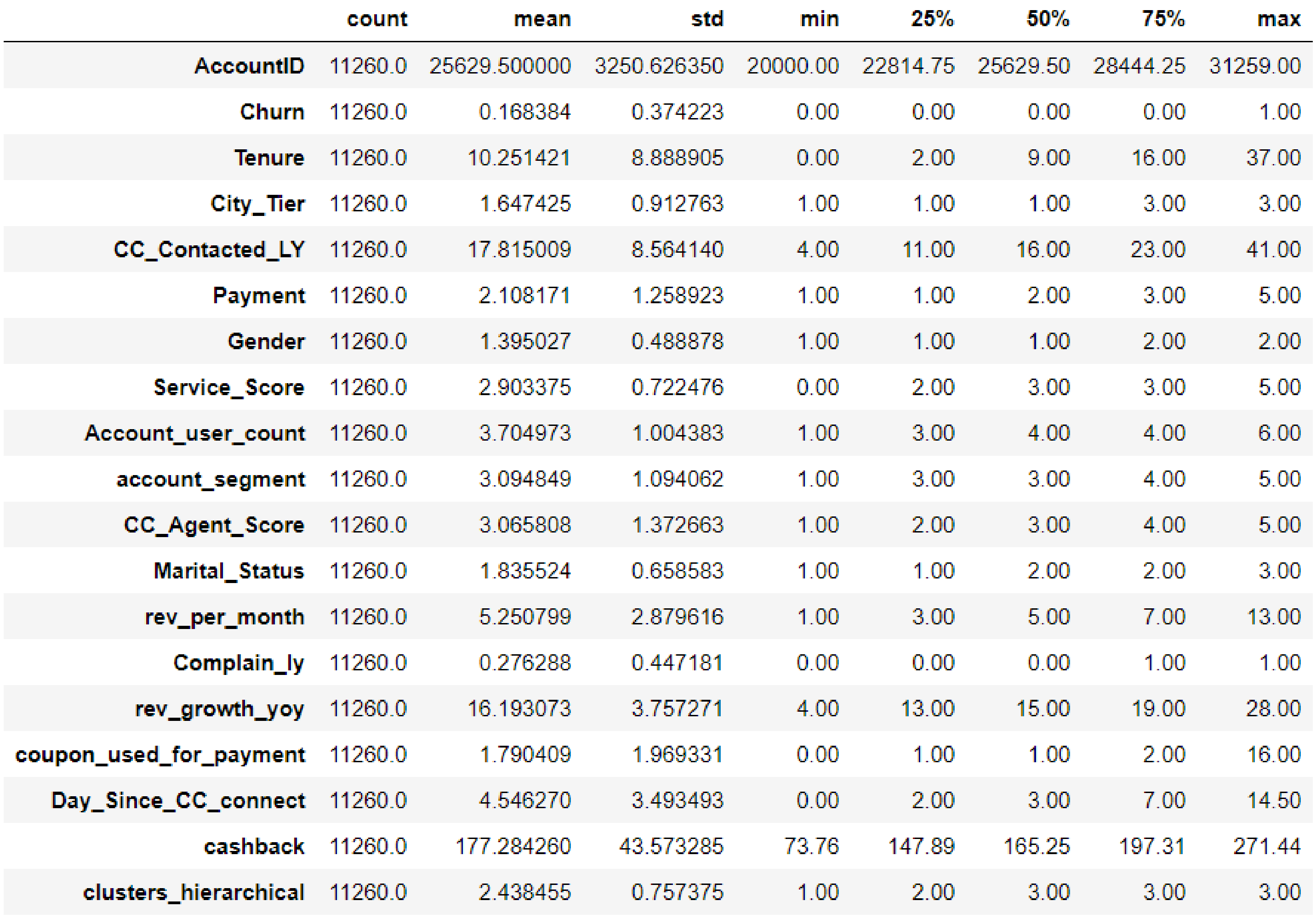
It can either be 0 or 1. 1 means the customer will churn and 0 means that the customer wont churn.

Variables which indicated the revenue generated by a particular account for example revenue generated monthly and the revenue growth over the past 2 years are also important.

Cashbacks and coupons are important attractive factors which will retain customers. Two of these important variables are no of times coupons have been used in the past one year and the amount of cashback earned.

Most of the columns has values with special characters which is to be removed/replaced. Columns ‘Gender’ and ‘Account segment’ has repetitive category names which is to be taken care of. For example in ‘Gender’ column, Male and M both categories are the same so M has to be replaced with Male in all the rows having so.

Statistical details of the data :



	count	mean	std	min	25%	50%	75%	max
AccountID	11260.0	25629.500000	3250.626350	20000.00	22814.75	25629.50	28444.25	31259.00
Churn	11260.0	0.168384	0.374223	0.00	0.00	0.00	0.00	1.00
Tenure	11260.0	10.251421	8.888905	0.00	2.00	9.00	16.00	37.00
City_Tier	11260.0	1.647425	0.912763	1.00	1.00	1.00	3.00	3.00
CC_Contacted_LY	11260.0	17.815009	8.564140	4.00	11.00	16.00	23.00	41.00
Payment	11260.0	2.108171	1.258923	1.00	1.00	2.00	3.00	5.00
Gender	11260.0	1.395027	0.488878	1.00	1.00	1.00	2.00	2.00
Service_Score	11260.0	2.903375	0.722476	0.00	2.00	3.00	3.00	5.00
Account_user_count	11260.0	3.704973	1.004383	1.00	3.00	4.00	4.00	6.00
account_segment	11260.0	3.094849	1.094062	1.00	3.00	3.00	4.00	5.00
CC_Agent_Score	11260.0	3.065808	1.372663	1.00	2.00	3.00	4.00	5.00
Marital_Status	11260.0	1.835524	0.658583	1.00	1.00	2.00	2.00	3.00
rev_per_month	11260.0	5.250799	2.879616	1.00	3.00	5.00	7.00	13.00
Complain_ly	11260.0	0.276288	0.447181	0.00	0.00	0.00	1.00	1.00
rev_growth_yoy	11260.0	16.193073	3.757271	4.00	13.00	15.00	19.00	28.00
coupon_used_for_payment	11260.0	1.790409	1.969331	0.00	1.00	1.00	2.00	16.00
Day_Since_CC_connect	11260.0	4.546270	3.493493	0.00	2.00	3.00	7.00	14.50
cashback	11260.0	177.284260	43.573285	73.76	147.89	165.25	197.31	271.44
clusters_hierarchical	11260.0	2.438455	0.757375	1.00	2.00	3.00	3.00	3.00

There were quite a few missing values in the data as below . For missing values in continuous variables, it was imputed with median and for those of categorical variables, they were treated with mode. Number of missing values after imputation . Outliers were present in the few variables as already stated above, hence we remove the outliers and look back at the boxplots. Charts after outlier removal has been shown in the **APPENDIX**. Categorical columns namely Payment, Account Segment, Marital Status and login device are transformed into numerical columns by label encoding. Here is how the data looks after encoding.

The data fetched for building the churn prediction model is an imbalanced dataset in terms of the class (Churn). Lets have a look below :

Number of people who doesn't churn (class 0) = 9364 (83.16 %)

Number of people who churn (class 1) = 1896 (16.83 %)

In case of any imbalanced data set the model tends to predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class.

From business perspective in this case the model will have high chances of predicting wrongly a customers tendency to churn. This means that the model wont recognize a customer who will churn and wrongly predict as a customer who will stay with the company. This is a serious issue for the company as it will unknowingly loose customers instead of having a churn prediction model and its revenue will get affected.

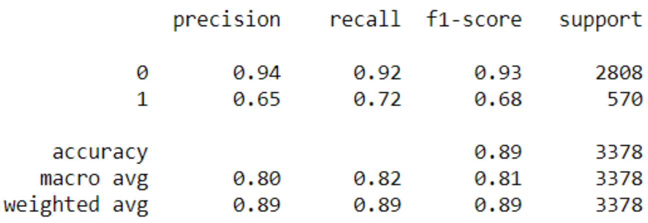
This can be treated by implementing SMOTE (Synthetic Minority Over-sampling Technique) in our dataset. It will do an over sampling in the data by increasing the number of minority classes and reduce over fitting of the model.

4) Model Building and interpretation

Random Forest :

After applying the best parameters from Grid Search on the Random Forest model we get the below results. Model accuracy for test data : 0.89

Classification report :

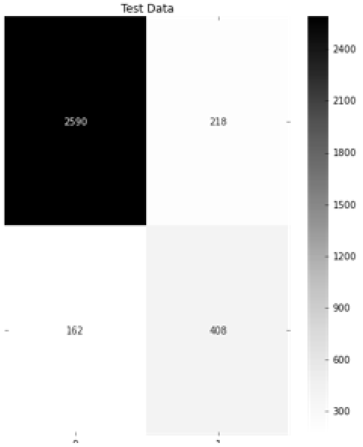


```
              precision    recall  f1-score   support

           0       0.94      0.92      0.93      2808
           1       0.65      0.72      0.68       570

    accuracy                           0.89      3378
   macro avg       0.80      0.82      0.81      3378
weighted avg       0.89      0.89      0.89      3378
```

Confusion matrix :

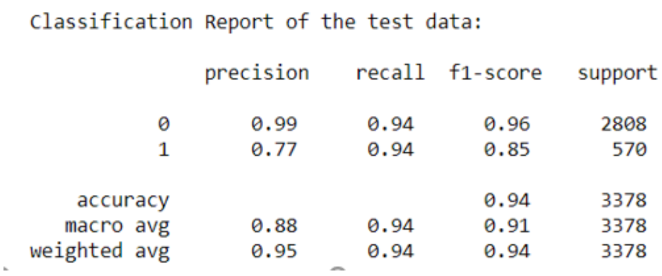


Surprisingly after tuning the random forest model, the performance decreases significantly on the test set. The recall has become 0.72 and will perform average in predicting which customers will churn. We wont recommend using this model however the normal untuned default model can be used .Please refer to the Appendix for ROC curve.

KNN :

After applying the best parameters from Grid Search on the KNN model we get the below results. Model accuracy for test data : 0.94

Classification report :



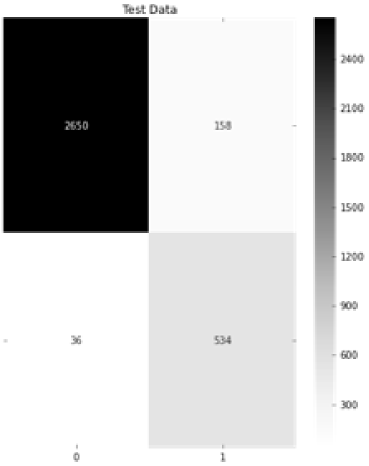
```
Classification Report of the test data:

              precision    recall  f1-score   support

           0       0.99      0.94      0.96      2808
           1       0.77      0.94      0.85       570

    accuracy                           0.94      3378
   macro avg       0.88      0.94      0.91      3378
weighted avg       0.95      0.94      0.94      3378
```

Confusion matrix :

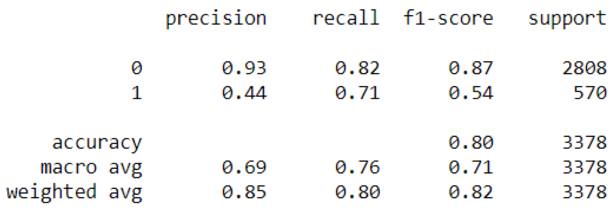


Tuning the KNN model increases all accuracy, precision and recall for both the train and test sets. It is giving a good recall of 0.94 means that the model was able to catch 94% of the actual Churn cases. This is the measure we really care about, because we want to miss as few of the true Churn cases as possible. Please refer to the Appendix for ROC curve.

Logistic Regression :

After tuning the logistic regression model we find that model score for test data : 0.799

Classification report :

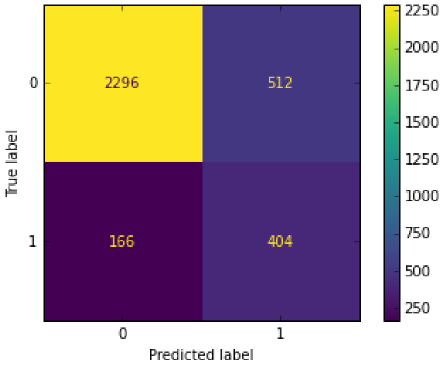


```
              precision    recall  f1-score   support

           0       0.93      0.82      0.87      2808
           1       0.44      0.71      0.54       570

    accuracy                           0.80      3378
   macro avg       0.69      0.76      0.71      3378
weighted avg       0.85      0.80      0.82      3378
```

Confusion matrix :

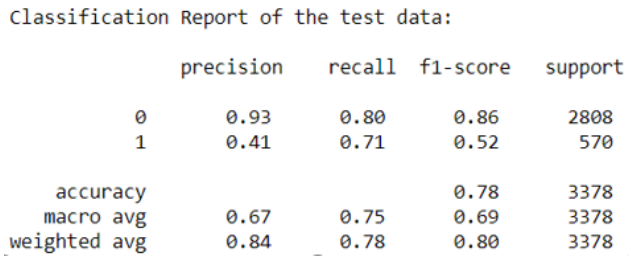


The Logistic regression model performs well on the training set. It had good accuracy and recall. In the test set, the model also works descent as recall and accuracy is fairly descent. A Churn class Recall of 0.71 means that the model was able to catch 71% of the actual Churn cases. This is the measure we really care about, because we want to miss as few of the true Churn cases as possible. Please refer to the Appendix for ROC curve.

Linear Discriminant Analysis :

After tuning the LDA model, we find that the accuracy score on the test data is 0.78

Classification report :



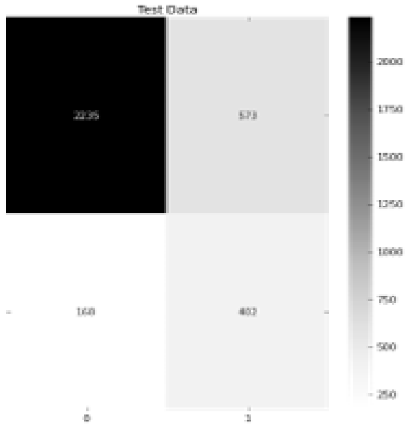
```
Classification Report of the test data:

              precision    recall  f1-score   support

           0       0.93      0.80      0.86      2808
           1       0.41      0.71      0.52       570

    accuracy                           0.78      3378
   macro avg       0.67      0.75      0.69      3378
weighted avg       0.84      0.78      0.80      3378
```

Confusion matrix :



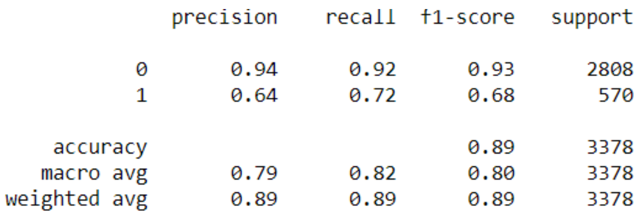
The Linear Discriminant Analysis prediction model performs well on the training set. It had good accuracy and recall. In the test set, the model also works descent as recall and accuracy is fairly descent. A Churn class Recall of 0.71 means that the model was able to catch 71% of the actual Churn cases. This is the measure we really care about, because we want to miss as few of the true Churn cases as possible. Please refer to the Appendix for ROC curve.

Bagging :

Using Random forest in bagging classifier, we get the following results :

Model accuracy for test data : 0.89

Classification report :

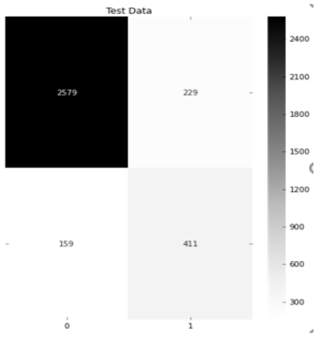


```
              precision    recall  f1-score   support

           0       0.94      0.92      0.93      2808
           1       0.64      0.72      0.68       570

    accuracy                           0.89      3378
   macro avg       0.79      0.82      0.80      3378
weighted avg       0.89      0.89      0.89      3378
```

Confusion matrix :



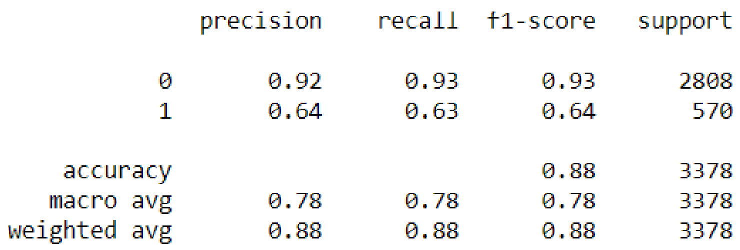
Bagging technique with Random forest although works well in the train data however fails to perform good in the test data. Recall is also less as compared to Random forest or K nearest neighbors. Please refer to the Appendix for ROC curve.

Adaptive Boosting :

After applying adaptive boosting, we get the below results on the test data :

Model accuracy for test data : 0.88

Classification report :

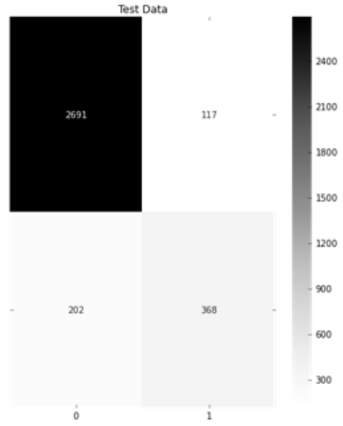


```
              precision    recall  f1-score   support

           0       0.92      0.93      0.93      2808
           1       0.64      0.63      0.64       570

    accuracy                           0.88      3378
   macro avg       0.78      0.78      0.78      3378
weighted avg       0.88      0.88      0.88      3378
```

Confusion matrix :



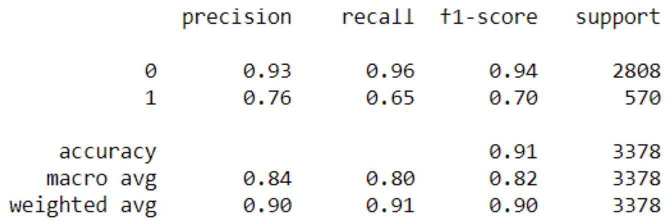
Adaptive boosting performs well on the train data but on the test data it performs poorly . Precision and Recall as compared to the other models are below average and it predicts 63% of the total cases of customers who will actually churn . Hence we will disregard this model. Please refer to the Appendix for ROC curve.

Gradient Boosting :

After applying Gradient Boosting, we get the below results :

Model accuracy for test data : 0.91

Classification report :

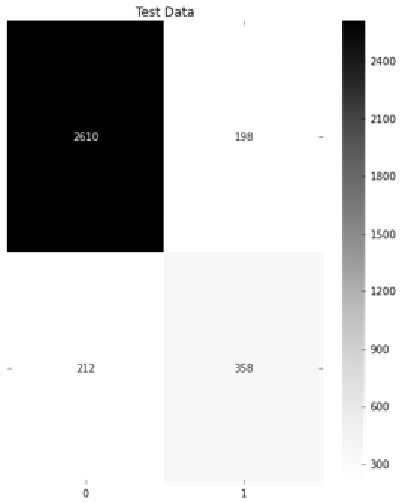


```
              precision    recall  f1-score   support

           0       0.93      0.96      0.94      2808
           1       0.76      0.65      0.70       570

    accuracy                           0.91      3378
   macro avg       0.84      0.80      0.82      3378
weighted avg       0.90      0.91      0.90      3378
```

Confusion matrix :

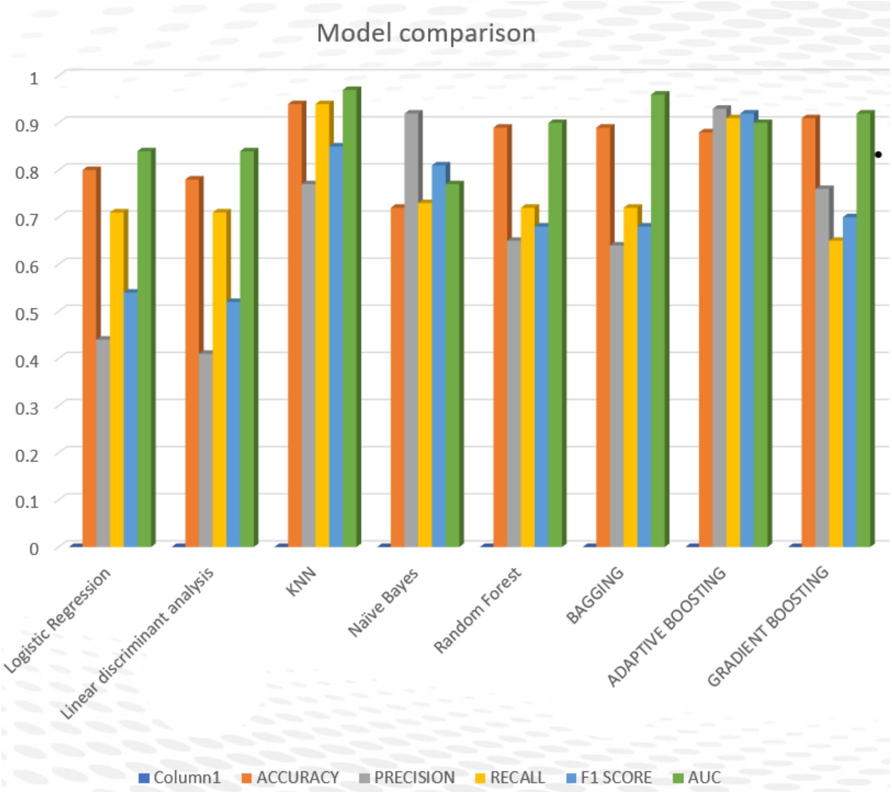


Model Validation

Lets look at the summarized table below which contains the scores for all the tuned models on the test data.

	ACCURACY	PRECISION	RECALL	F1 SCORE	AUC
Logistic Regression	0.80	0.44	0.71	0.54	0.84
Linear Discriminant Analysis	0.78	0.41	0.71	0.52	0.84
KNN	0.94	0.77	0.94	0.85	0.97
Naïve Bayes	0.72	0.92	0.73	0.81	0.77
Random Forest	0.89	0.65	0.72	0.68	0.90
BAGGING	0.89	0.64	0.72	0.68	0.96
ADAPTIVE BOOSTING	0.88	0.93	0.91	0.92	0.90
GRADIENT BOOSTING	0.91	0.76	0.65	0.70	0.92

Lets also visualize which model does the best in terms of accuracy and recall.



Besides accuracy, recall is the most important factor when validating a churn prediction model. Lets look at what does precision and recall actually mean in churn prediction model.

- **Precision – Of all the users that the algorithm predicts will churn, how many of them do actually churn?**
- **Recall – What percentage of users that end up churning does the algorithm successfully find?**

Hence, in this case, recall is useful in validating the models.

Comparing the performances of the models, KNN was selected as the most optimum model for this business problem. Since it maximum number of times it catches the actual churn cases or in other words have a good recall. This is the measure we really care about, because we want to miss as few of the true Churn cases as possible.

CUSTOMER CHURN DTH

After building and tuning couple of models , lets select the K Nearest Neighbor tuned model to solve our business purpose as it maximum number of times it catches the actual churn cases or in other words have a

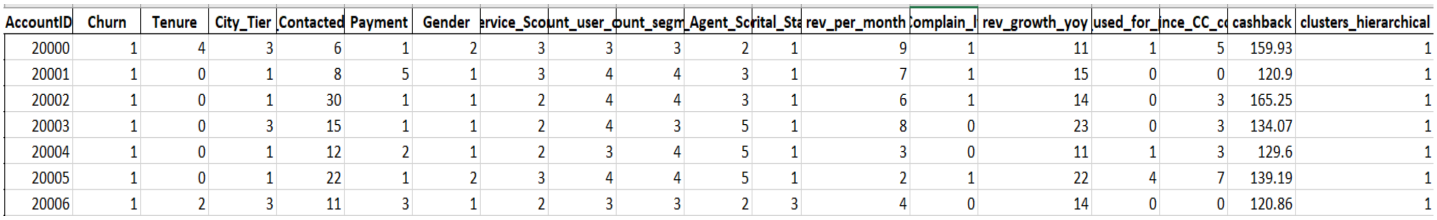
good recall. his is the measure we really care about, because we want to miss as few of the true Churn cases as possible.

94 % of the actual churn cases was caught by the model in the testing set, also the model performs pretty similarly in both the train and test sets.

Final interpretation / recommendation

Clustering was done on the dataset and based on the same we have three clusters . Lets have a look.

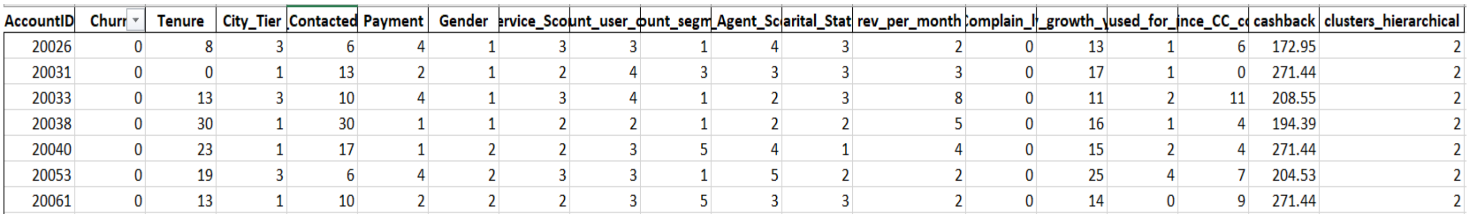
Cluster 1 :



AccountID	Churn	Tenure	City_Tier	Contacted	Payment	Gender	ervice_Sco	unt_user_	ount_segm	Agent_Sc	rital_Sta	rev_per_month	omplain_l	rev_growth_yoy	used_for_	ince_CC_c	cashback	clusters_hierarchical
20000	1	4	3	6	1	2	3	3	3	2	1	9	1	11	1	5	159.93	1
20001	1	0	1	8	5	1	3	4	4	3	1	7	1	15	0	0	120.9	1
20002	1	0	1	30	1	1	2	4	4	3	1	6	1	14	0	3	165.25	1
20003	1	0	3	15	1	1	2	4	3	5	1	8	0	23	0	3	134.07	1
20004	1	0	1	12	2	1	2	3	4	5	1	3	0	11	1	3	129.6	1
20005	1	0	1	22	1	2	3	4	4	5	1	2	1	22	4	7	139.19	1
20006	1	2	3	11	3	1	2	3	3	2	3	4	0	14	0	0	120.86	1

In this segment of accounts customers are more likely to churn . Accounts for this cluster are new or are having less tenure . Most of them are from Tier 1 cities and having an average or low revenue growth in months and growth in the past two years. Customers in this segment are less connected with the customer support and have used less coupons/offers.

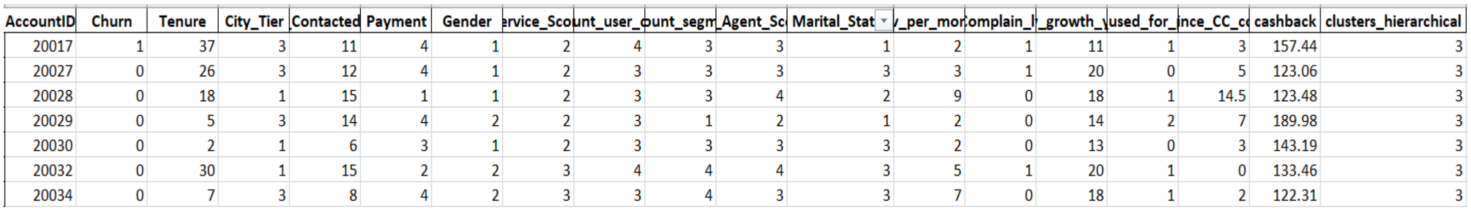
Cluster 2:



AccountID	Churr	Tenure	City_Tier	Contacted	Payment	Gender	ervice_Sco	unt_user_	ount_segm	Agent_Sc	arital_Stat	rev_per_month	omplain_l	_growth_y	used_for_	ince_CC_c	cashback	clusters_hierarchical
20026	0	8	3	6	4	1	3	3	1	4	3	2	0	13	1	6	172.95	2
20031	0	0	1	13	2	1	2	4	3	3	3	3	0	17	1	0	271.44	2
20033	0	13	3	10	4	1	3	4	1	2	3	8	0	11	2	11	208.55	2
20038	0	30	1	30	1	1	2	2	1	2	2	5	0	16	1	4	194.39	2
20040	0	23	1	17	1	2	2	3	5	4	1	4	0	15	2	4	271.44	2
20053	0	19	3	6	4	2	3	3	1	5	2	2	0	25	4	7	204.53	2
20061	0	13	1	10	2	2	2	3	5	3	3	2	0	14	0	9	271.44	2

Customers in the accounts for this cluster are less likely to churn. They have been old and loyal customers . They have been frequent in contacting the customer care and generated descent to average revenue for the company. They are using coupons and availing offers to get more than average cashbacks, most of them belonging to tier 3 city.

Cluster 3:

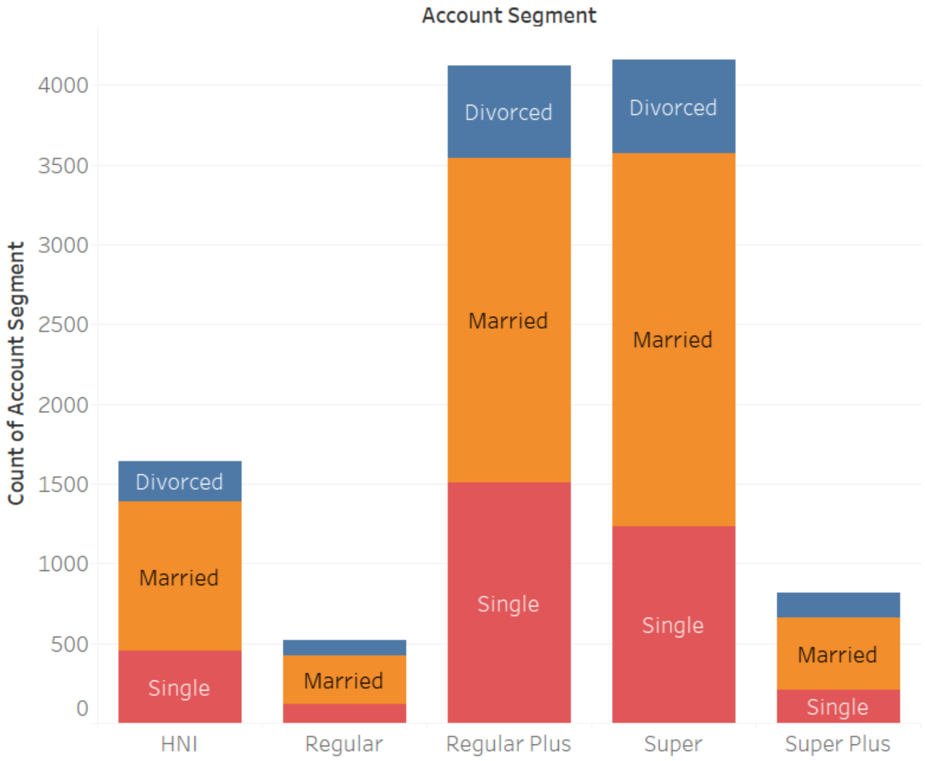


AccountID	Churn	Tenure	City_Tier	Contacted	Payment	Gender	ervice_Sco	unt_user_c	ount_segm	Agent_Sc	Marital_Stat	v_per_mor	omplain_l	_growth_y	used_for_	ince_CC_c	cashback	clusters_hierarchical
20017	1	37	3	11	4	1	2	4	3	3	1	2	1	11	1	3	157.44	3
20027	0	26	3	12	4	1	2	3	3	3	3	3	1	20	0	5	123.06	3
20028	0	18	1	15	1	1	2	3	3	4	2	9	0	18	1	14.5	123.48	3
20029	0	5	3	14	4	2	2	3	1	2	1	2	0	14	2	7	189.98	3
20030	0	2	1	6	3	1	2	3	3	3	3	2	0	13	0	3	143.19	3
20032	0	30	1	15	2	2	3	4	4	4	3	5	1	20	1	0	133.46	3
20034	0	7	3	8	4	2	3	3	4	3	3	7	0	18	1	2	122.31	3

Customers in the accounts for this cluster are also less likely to churn but we can say that they are less active customers in the segment. With an average tenure of 10 years, although they have stayed with the company, however plan upgrades, availing offers and coupons have been on a low for this cluster.

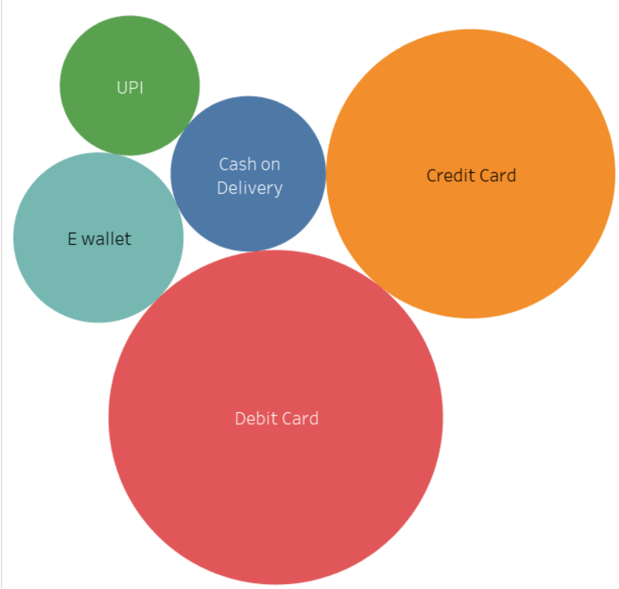
Lets have a look at some recommendations :

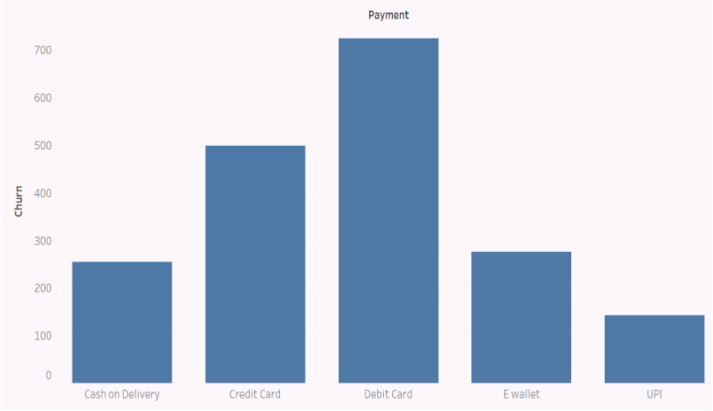
- Married couples should be mostly targeted to get attractive offers/discounts as they are more likely to churn



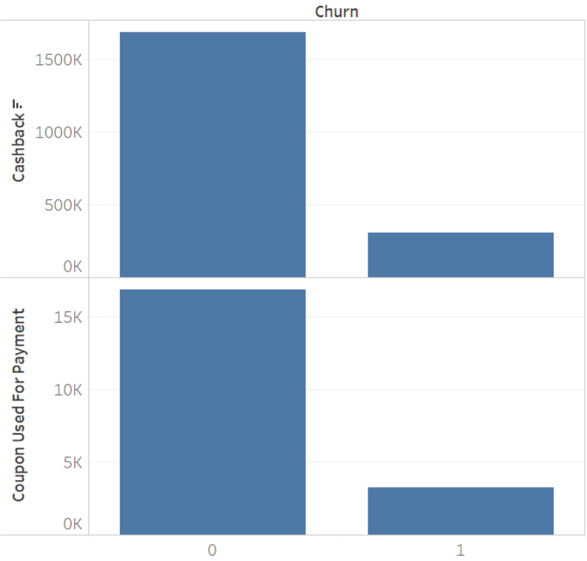
CUSTOMER CHURN DTH

- Companies should host discount campaigns in festive seasons in Tier 1 cities as churn rate for customer accounts in these tier cities is the highest.
- Customer accounts who are into using debit cards for the payment methods should be targeted for receiving cashbacks from Debit card transactions.



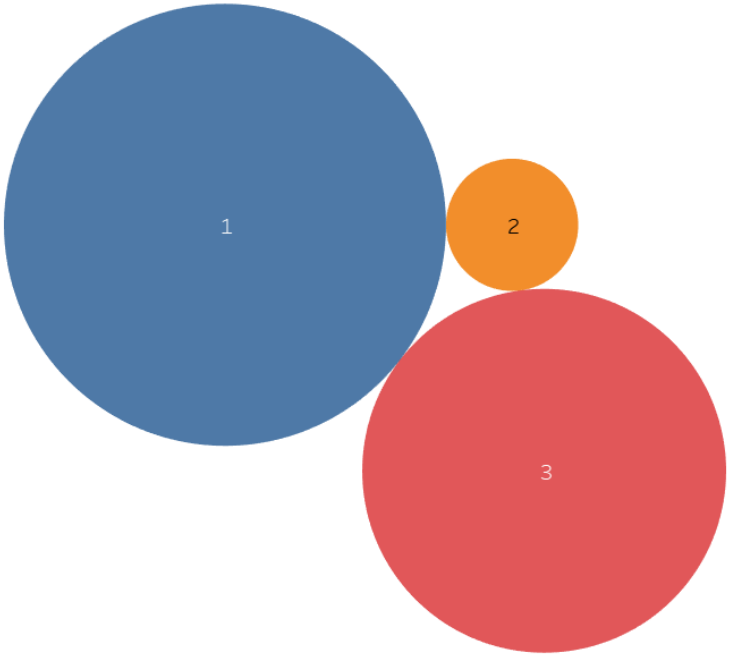


- Customer accounts where cashbacks and coupons used are frequent are loyal customers and are less likely to churn. Customers who are not availing discounts/offers maybe asked for feedback based on which the company can take necessary actions.



CUSTOMER CHURN DTH

- The DTH company can organize upgrade campaigns in the Tier 1 cities as accounts from these areas churn the most.



APPENDIX

Appendix.xlsx