Capstone - Ecommerce Customer churn, Final Report

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

A brief Introduction

An online retail (E commerce) company wants to know the customers who are going to churn, so accordingly they can approach customer to offer some promos

Problem Statement

1. To understand the dataset and check and establish data hygiene in the dataset for proper analysis
2. Analyze the dataset to understand the customer churn behavior and find causes as well as indicators of customer churn
3. Make clusters of the customers based on similar buying and churning properties.

Need of the Study

This study is needed to understand

1. Factors that are responsible for customer churn
2. Factors or variables that might be indicative of possible customer churn behaviour
3. What improvements can be done to reduce the customer churn

Business opportunity

The Average customer churn stands at 16.25%. Through this opportunity we have the ability to reduce the costs associated with customer churn by a significant margin by identifying the leaks and understanding the customer churn.

2. EDA and Business Implication

Data Report

1. **Data understanding in terms of time, frequency and methodology:** No such information is available in the dataset and hence so insight can be drawn in this regard.
2. **Visual inspection of data:**
    1. The Dataset comprises 5630 observations and 20 variables. Out of these 20 variables, 7 are categorical Variables and remaining 13 are numerical variables
    2. The data summary from the dataset is as follows:

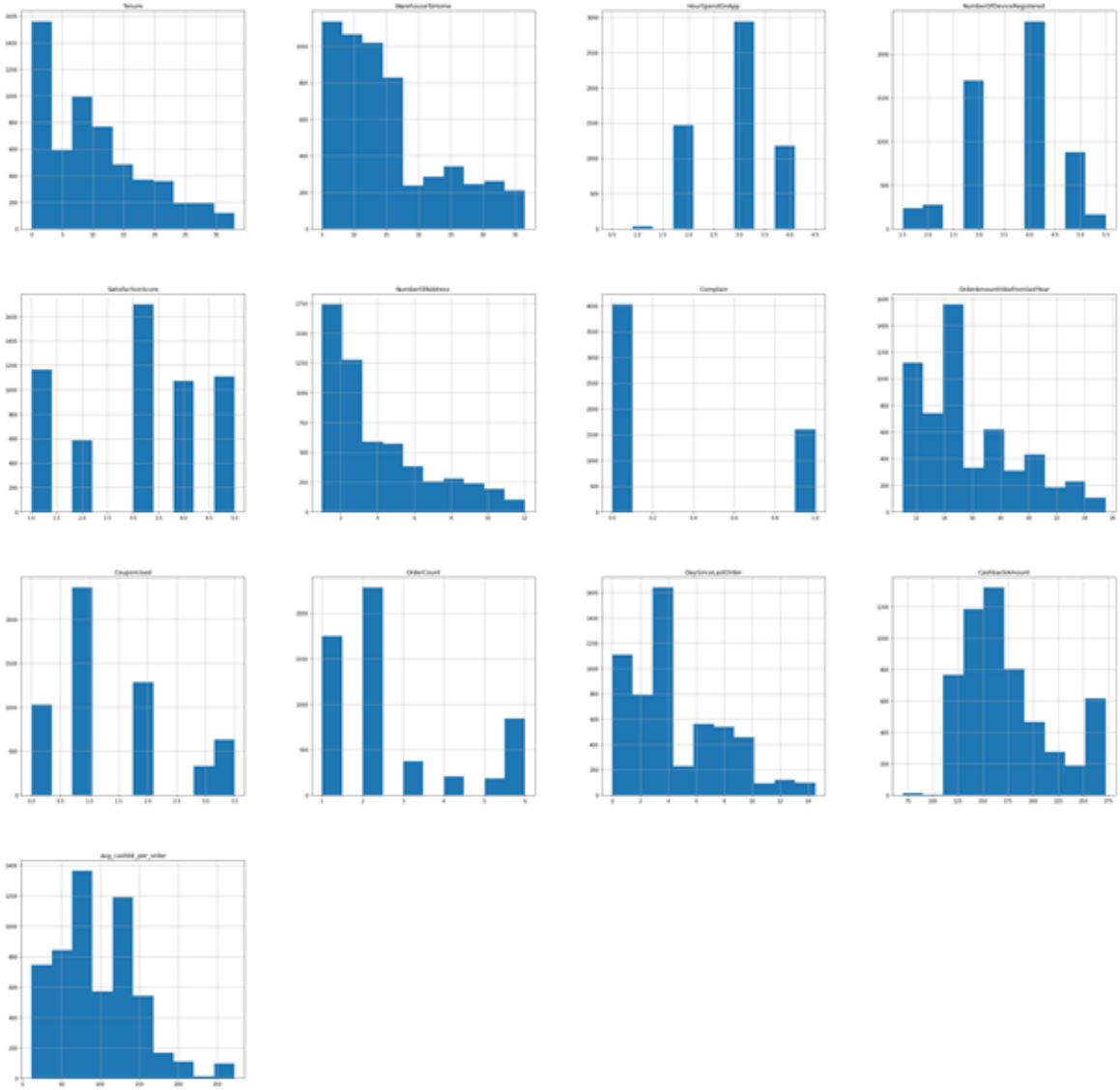
Aa Title	# count	# mean	# Median	# min Value	# max value
Tenure	5366	10.189899	9	0	61
WarehouseToHome	5379	15.639896	14	5	127
HourSpendOnApp	5375	2.931535	3	0	5
NumberOfDeviceRegistered	5630	3.688988	4	1	6
SatisfactionScore	5630	3.066785	3	1	5
NumberOfAddress	5630	4.214032	3	1	22
Complain	5630	0.284902	0	0	1
OrderAmountHikeFromlastYear	5365	15.707922	15	11	26
CouponUsed	5374	1.751023	1	0	16
OrderCount	5372	3.008004	2	1	16
DaySinceLastOrder	5323	4.543491	3	0	46

Aa Title	# count	# mean	# Median	# min Value	# max value
CashbackAmount	5630	177.22303	163.28	0	324.99

Exploratory Data Analysis

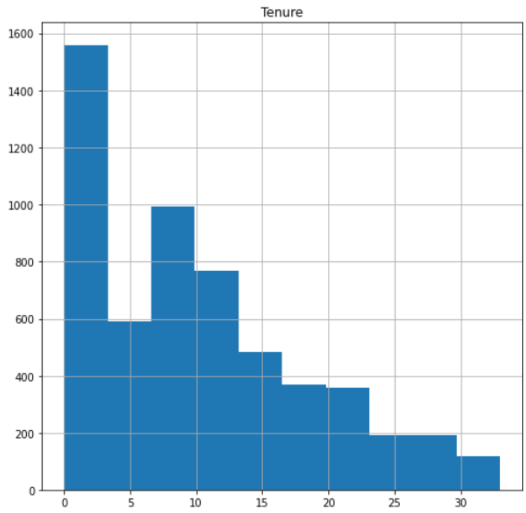
Univariate Analysis:

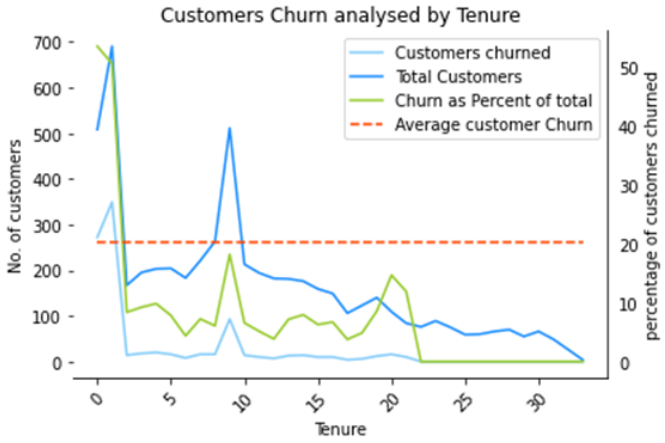
The Histograms:



From the above histograms we can draw the following inferences:

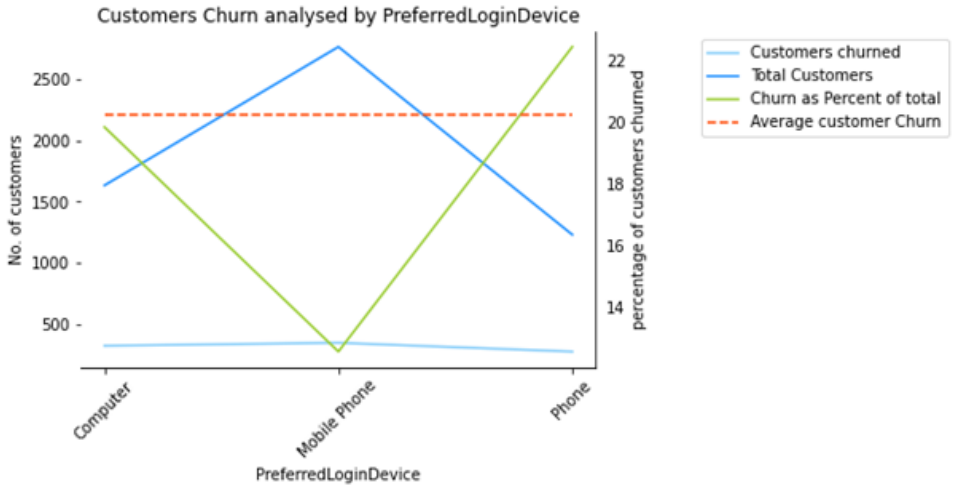
1. A majority of people have the tenure of less than 15, hence we can conclusively argue that there are very few people with considerably longer tenure.
2. **The people with very small tenure (<2) tend to have a very high attrition, but at the same time as the tenure increases, the churn rate decreases significantly.**
3. There is an increase in churn at the tenure level of around 7-8, which is also accompanied by increased customer. This could be due to a certain specific customer cohort, where acquisition was very high.
4. Also the customer churn is very high at the tenure level of 20. This may also be due to specific customer cohort or it could be that those who had a tenure of 20 have now shifted somewhere else.
5. How ever the focus should be on elongating the tenure of the customer since it significantly decreases the churn rate.
6. **Also we need to make some kind of intervention at the tenure of 2, and smoothen the sharp bend in the curve.**



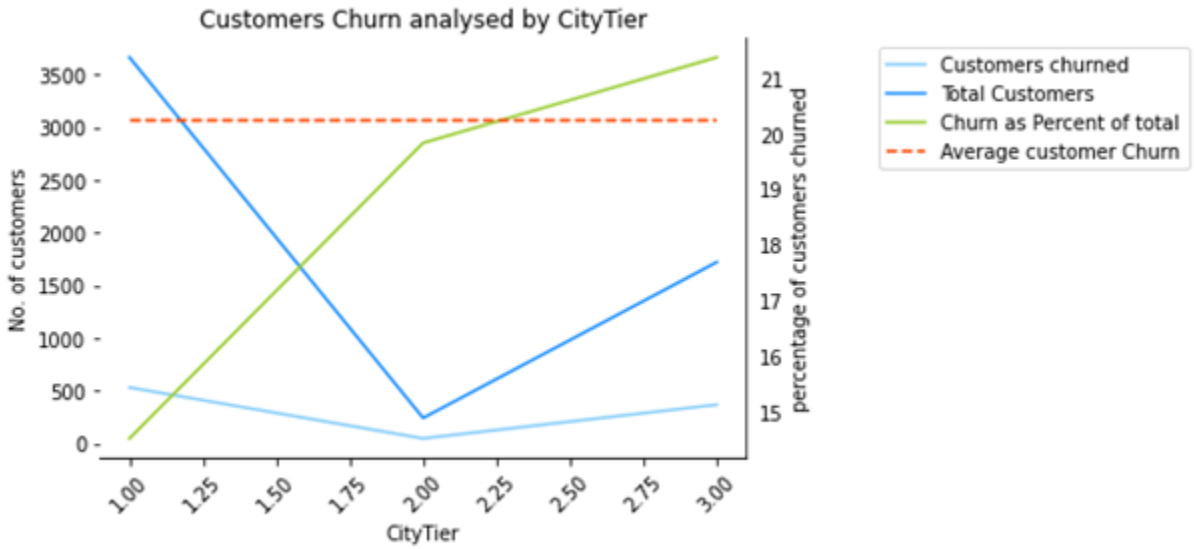


7. The customers who use the mobiles show the least churn and are present in high numbers.

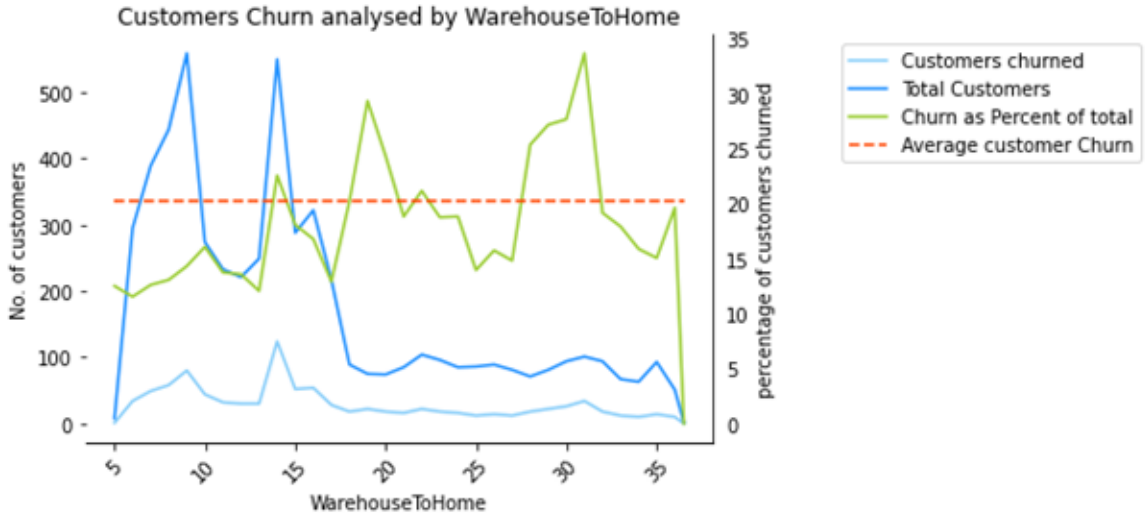
8. The customer using the phone are in the high churn area followed by the customers using the computer.



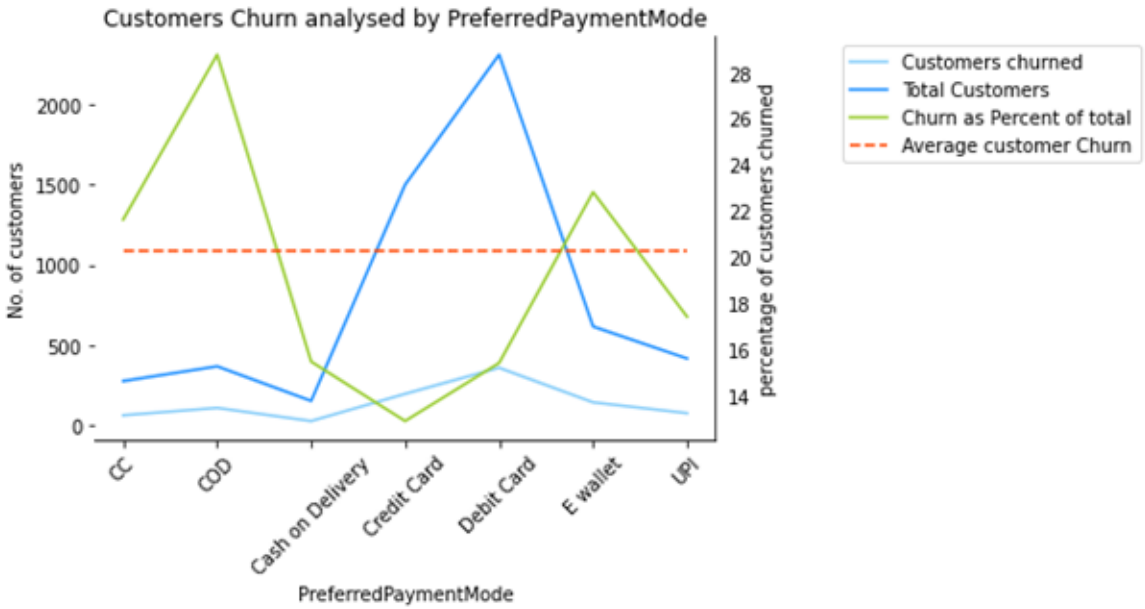
9. The customers in the tier 2 and tier 3 cities show the maximum churn which is almost close to the average churn level. But at the same time the customers from the tier 2 and tier 3 cities are very small in number. **Hence it follows that right targeting is needed in tier 2 and tier 3 cities to target the right customers and lower the customer churn.**



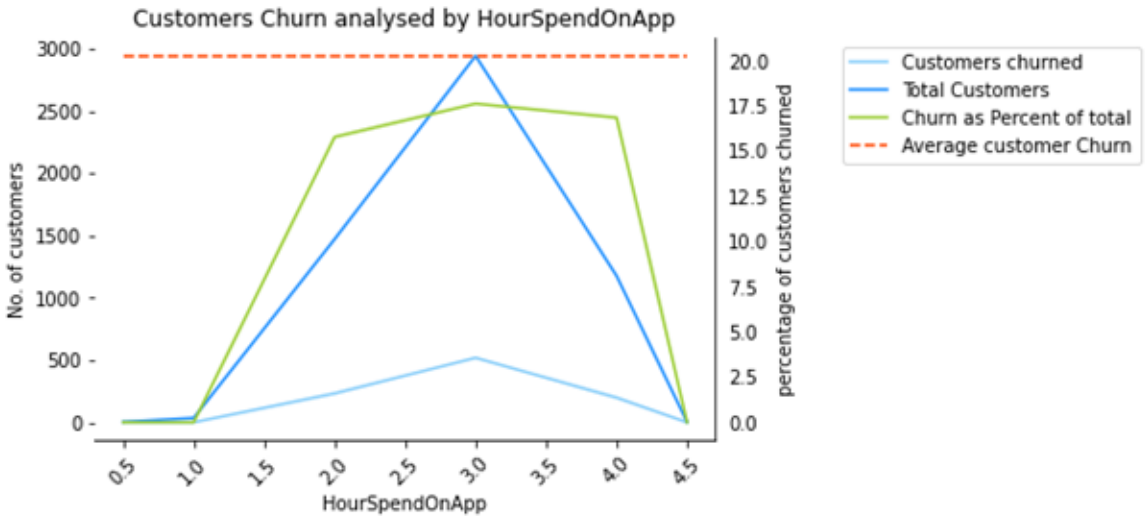
11. The customer churn seem directly influenced by the distance of the customer's house from the warehouse. The larger distances of the warehouse might be resulting in the poor customer service. Hence the company needs to rethink on the placement of the warehouses.



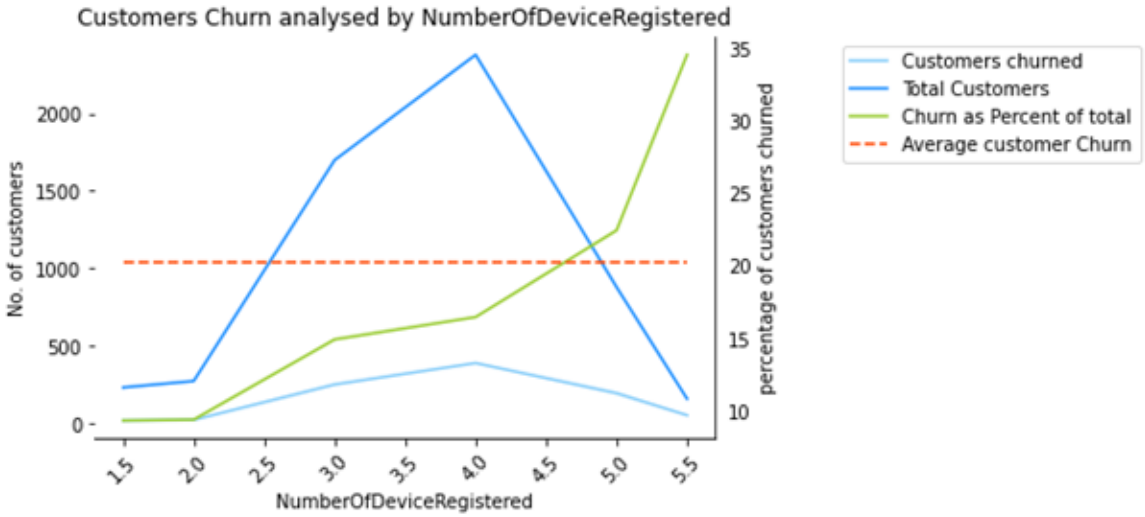
12. Here we can see that the customers with the CC, COD and E wallet methods have the highest churn. Hence there is a need to have a look at the order size of these customers and the quality of service being offered to them



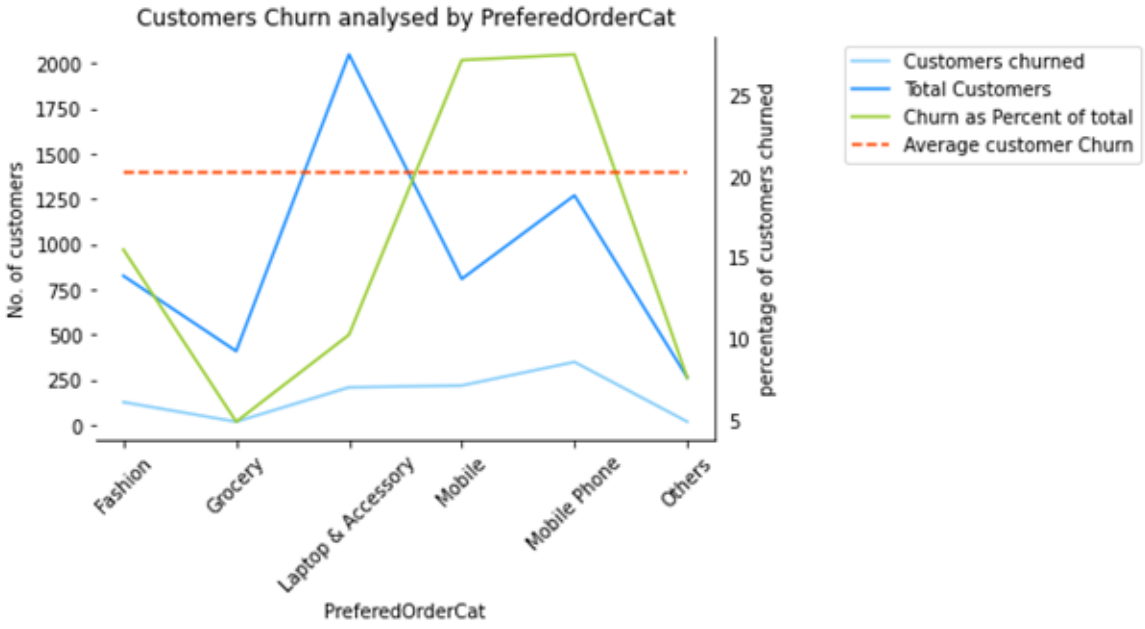
13. Here we can see that most customers spend about 3 -4 hours on the app. Hence the churn comes out to be higher. The churn varies only marginally as the number increases from 2 hrs to 4.



14. Here we can see that most customers use form 2-4 devices. The churn increases as the number of devices increase. This could be that more than one person uses any paid account. However, more information such as subscription / pricing information of the customer will be needed to understand the reasons thoroughly.

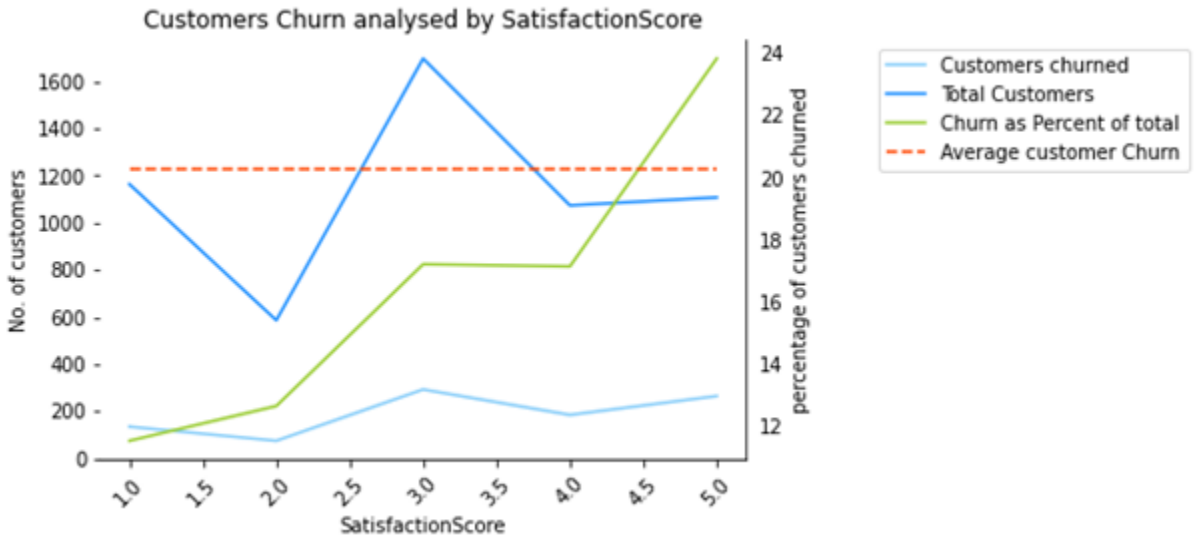


15. The order category of Mobile, mobile phones have the max churn where as the categories of necessity such as Grocery have a lower churn.

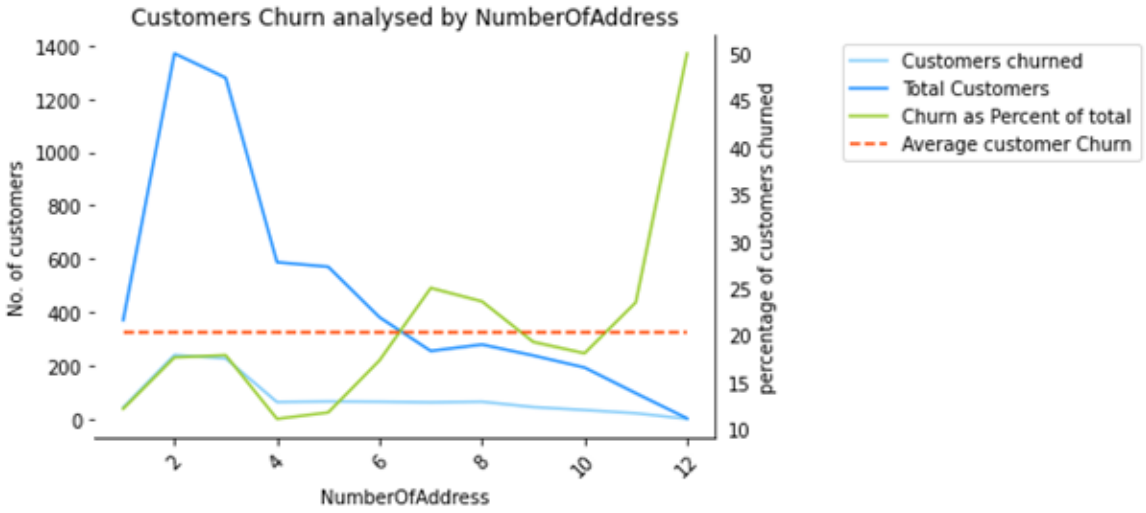


16. People who have given a lower satisfaction score show a lower churn compared to the people who gave a higher satisfaction score. This could be

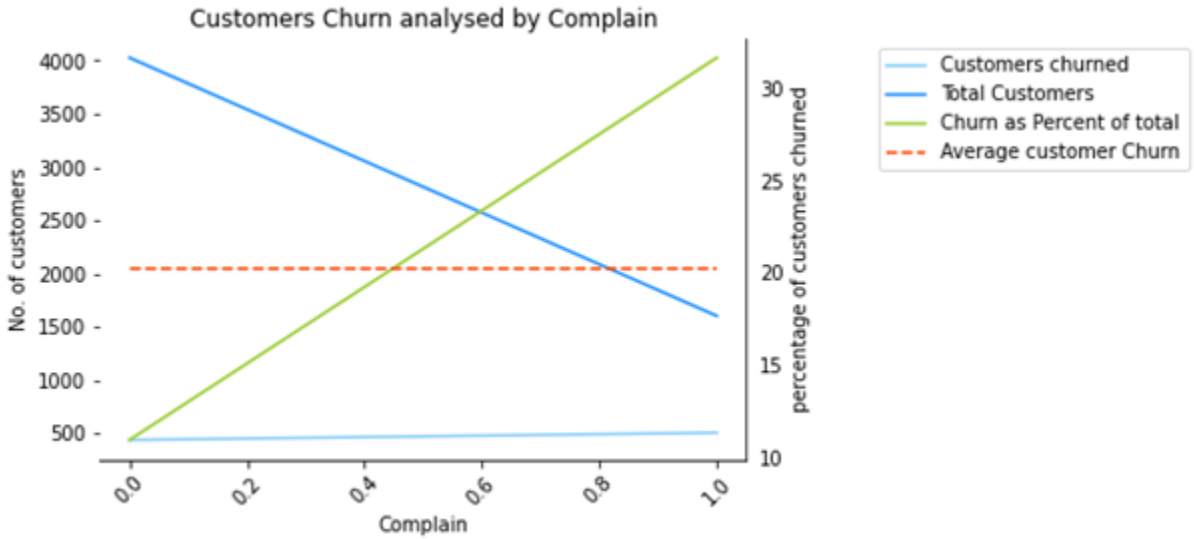
due to the fact that 1 Has been rated as the highest and 5 as the lowest satisfaction. If this is the case then the satisfaction score turns out to be the best predictor of the customer churn.



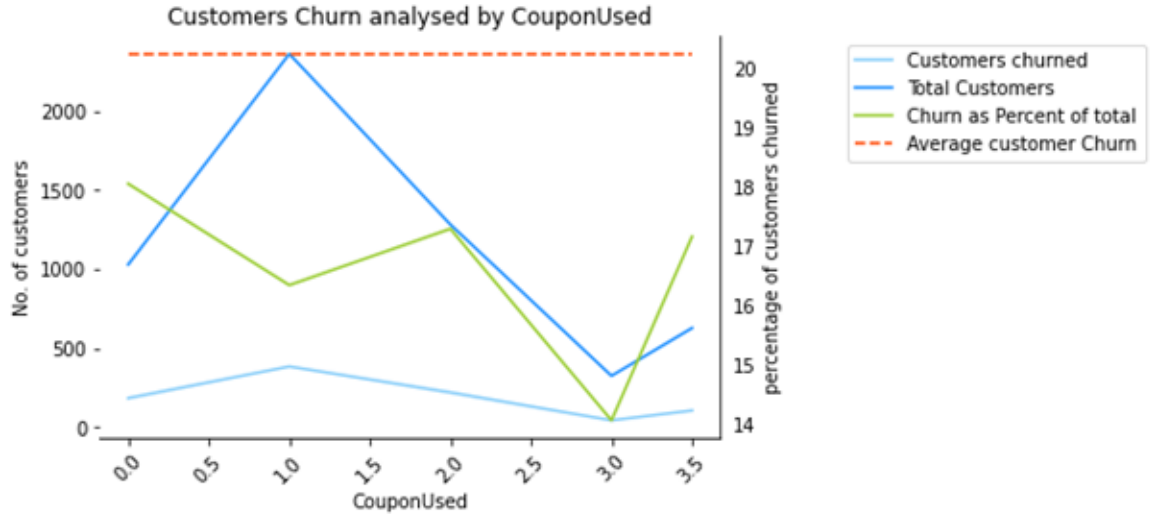
17. We can see below that the customers with the lower number of addresses show the lowest churn and as the number of addresses increase, the churn also increases.



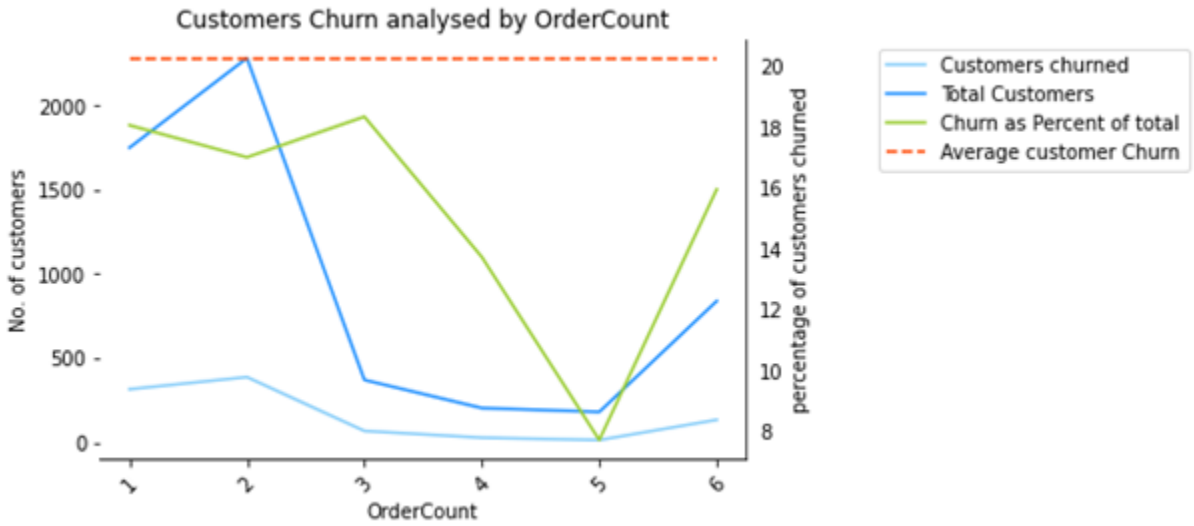
18. As the number of complaints increase, the customer churn increases significantly



19. The more coupons a customer uses, the lower is the churn rate.



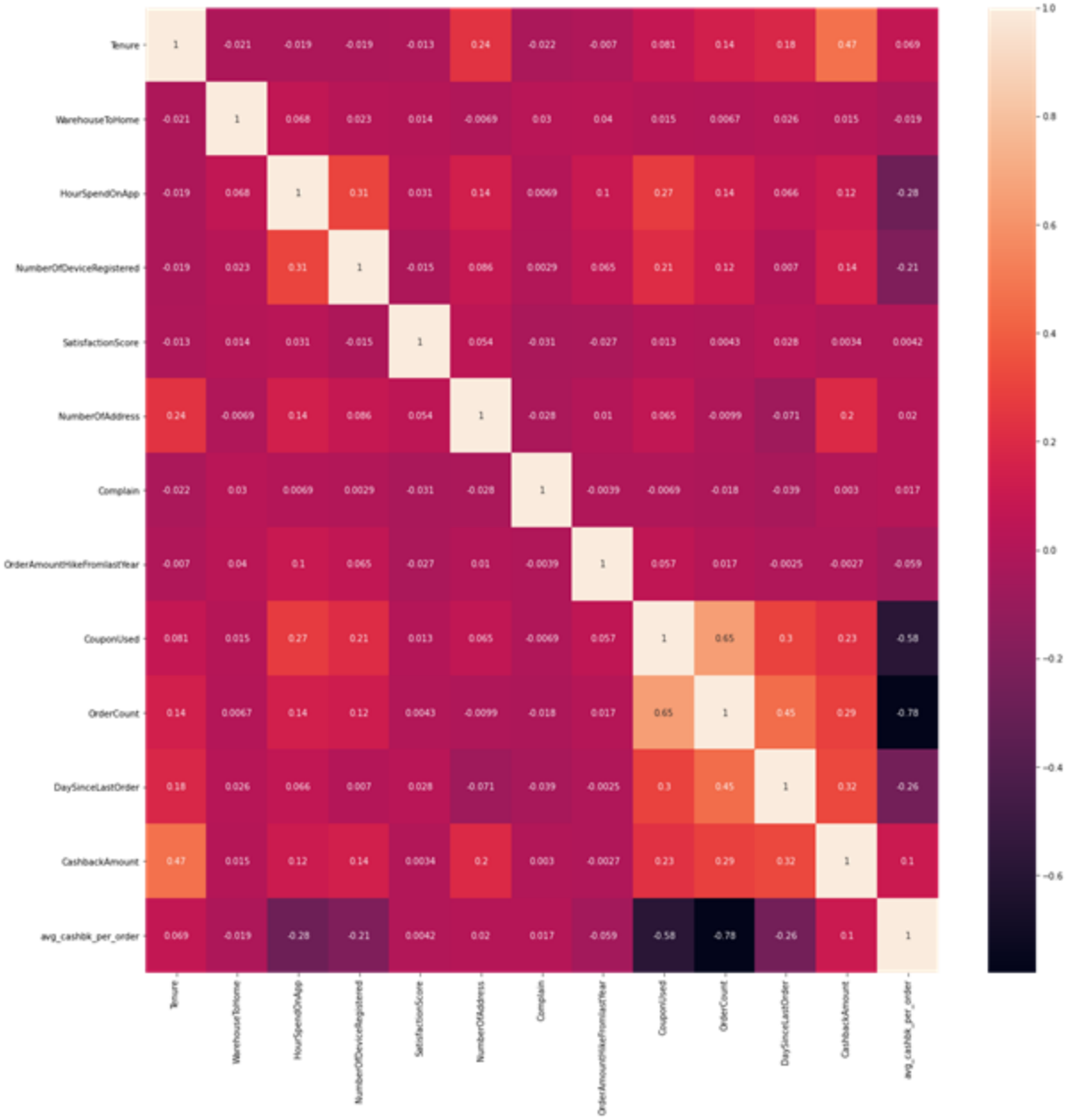
20. The higher is the order count the lower is the churn rate. Majority of the order count is 2.



Bivariate analysis

From the correlation heat map we can notice the follwing:

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

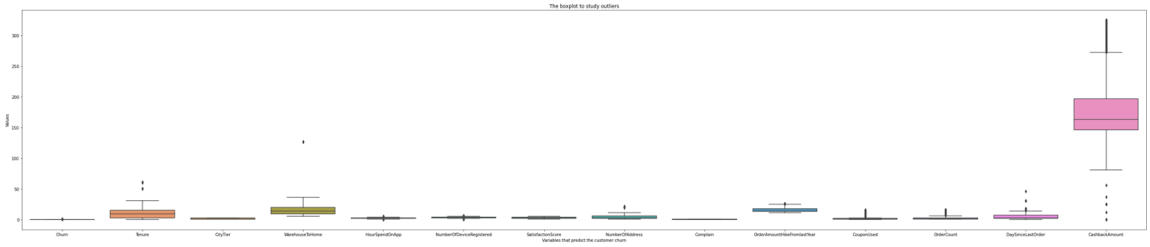


3. Data Cleaning and Pre-processing

1. There are missing values in the dataset. The information is as follows:

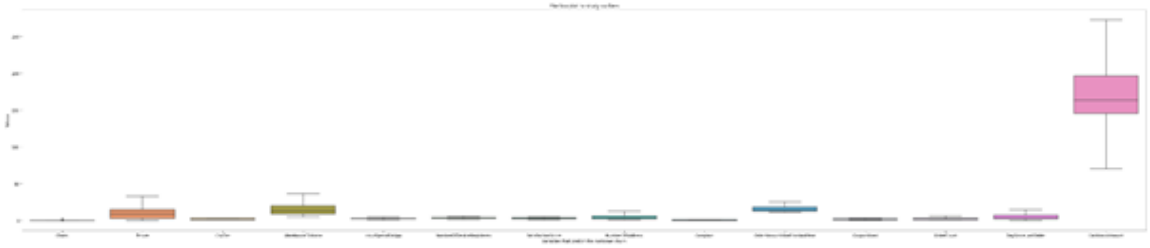
Tenure	264 null values
WarehouseToHome	251 null values
HourSpendOnApp	255 null values
OrderAmountHikeFromlastYear	265 null values
CouponUsed	256 null values
OrderCount	258 null values
DaySinceLastOrder	307 null values

2. **Treatment of Missing Values:**
    1. The missing values are filled with the median values of the of the relevant columns
3. **Outlier Treatment:**
    1. There were a few outliers in the data



2. The outliers were required to be treated since otherwise they would have created distortions in the Univariate and Bivariate analysis.

After treatment the outliers are now replaced with their corresponding upper range or lower range values.



New Variables Created & Variables dropped:

- **Average Cashback per order:**

    - Average Cashback per order = Cashback Amount / OrderCount.
    - This new variable was created to understand if amount of cashback per order had any effect on the customer churn or not.
- OrderCount and coupons used were dropped from the modelling since they showed a high positive correlation with each other and high negative correlation with the Average Cashback per order.
    - Also since effect of both the variables was already counted in the final model in the form of Average cashback per order, they were removed.

4. Model building and interpretation

Various models built:

- Logistic regression
- Linear Discriminant Analysis
- Decision Tree
- Random Forest
- KNN Model
- XGBoost

Why these models?

- The above models were build as the problem was that of classification.
- Apart from prediction of the customers who would churn, we also wanted to know which of the features are more important in predicting the customer churn and hence which of features should be focused upon to reduce the churn.
- All the models above provide the customer prediction as well as feature importance.
- Further the models should be easy to understand. Hence Logistic Regression, Linear Discriminant Analysis and Decision trees are the easiest to understand and transparent.

- Ensemble methods such as Random Forest and XGBoost are not transparent but are more reliable and avoid overfitting and hence are used.

Some facts involved in model building

- The categorical data was encoded using one hot encoding technique
- The continuous variables in the data were Scaled using standardscaler.
- The data was imbalanced with the label 1 being the 16.8% of the total data
  - Before OverSampling, counts of label '1': 948
  - Before OverSampling, counts of label '0': 4682
- **The data was later balanced for some of the iterations to improve the models. The data was balanced using SMOTE**
  - After OverSampling, counts of label '1': 4682
  - After OverSampling, counts of label '0': 4682
- There were 2 versions of the datasets
  1. Models trained with imbalanced data
  2. Models trained with balanced data

5. Model validation

Comparison of Various models

Models were compared based on the following metrics:

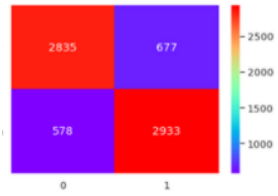
1. Specificity (ability to correctly predict the churn)
2. Accuracy (Denoted by Model Score)
3. AUC-Score and area under ROC Curves.
4. Feature importance as validated in the EDA

The Specificity and accuracy were calculated from the Confusion Matrices as given below.

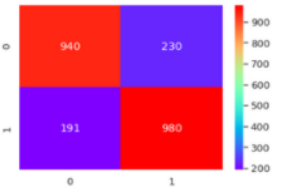
Confusion Matrices

Logistic regression - Smoted

Train

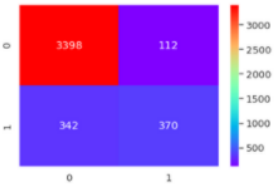


Test

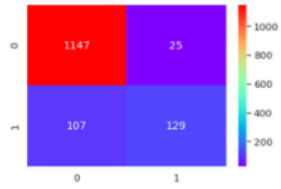


Logistic regression - Non Smoted

Train

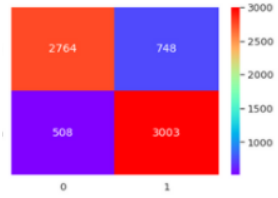


Test

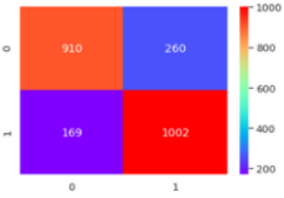


LDA - Smoted

Train



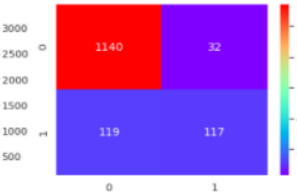
Test



LDA - Non Smoted

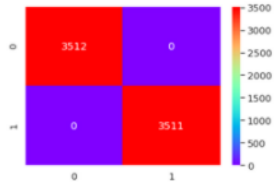
Train

Test

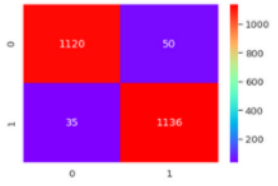


Decision Tree - Smoted

Train

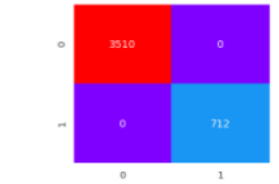


Test

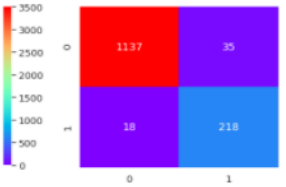


Decision Tree- Non Smoted

Train

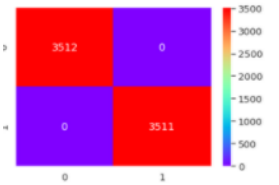


Test



Random Forest - Smoted

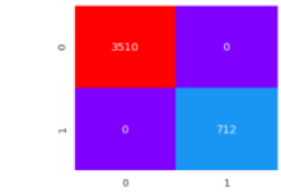
Train



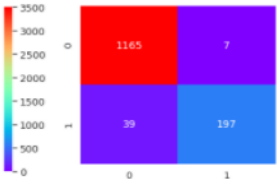
Test

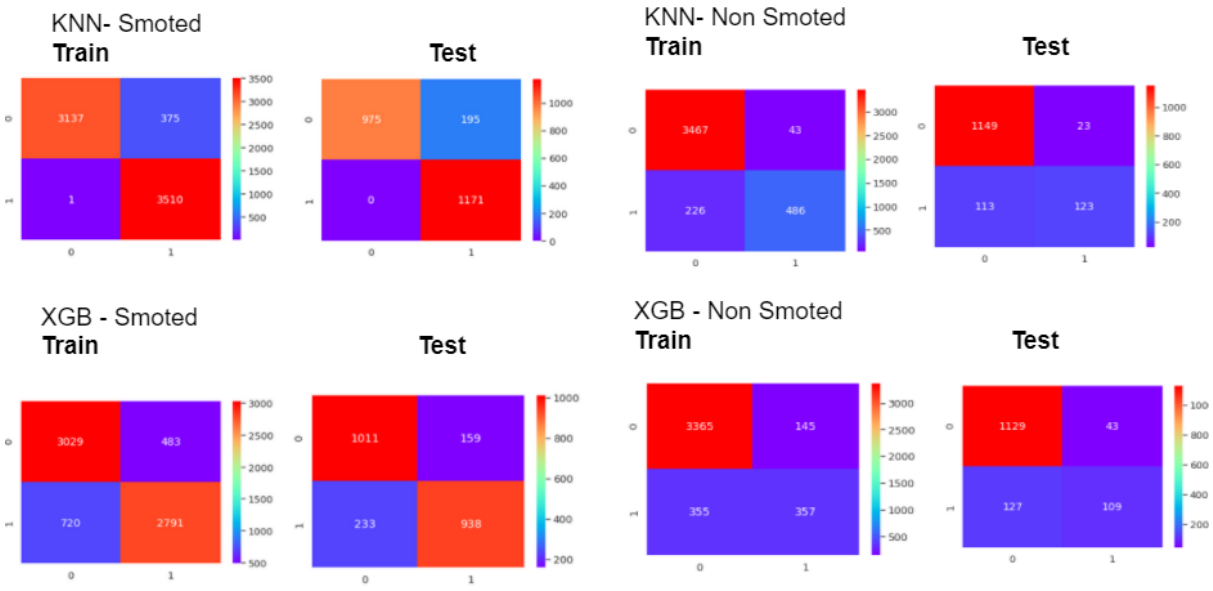
Random Forest - Non Smoted

Train



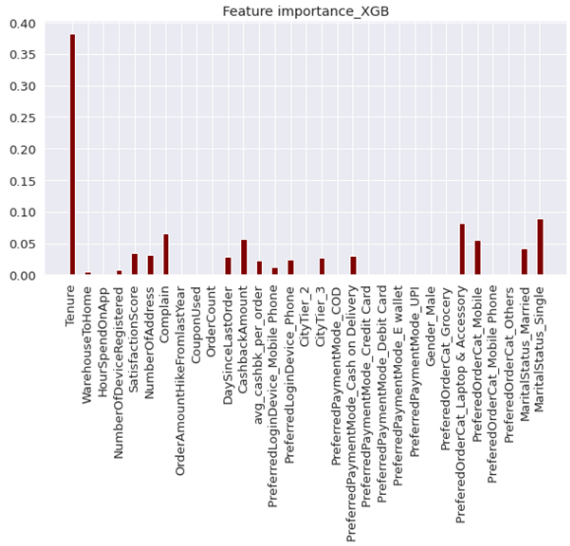
Test

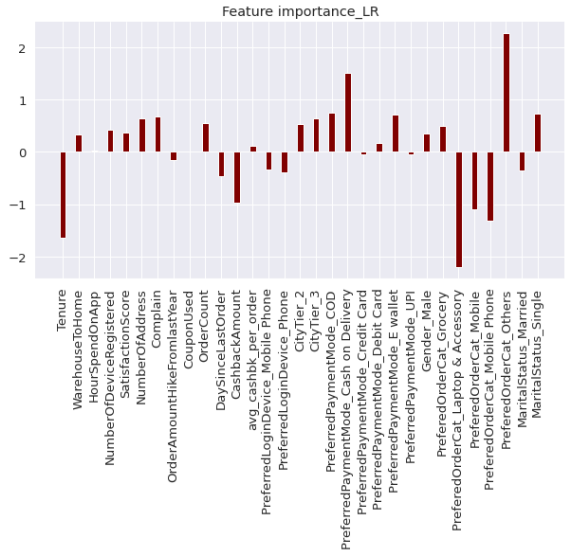


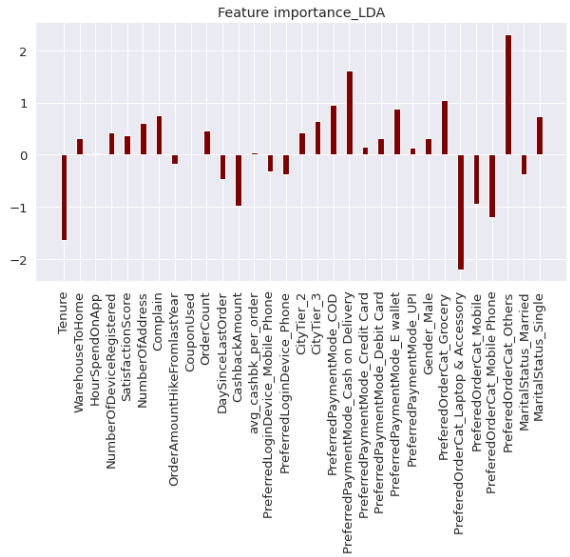


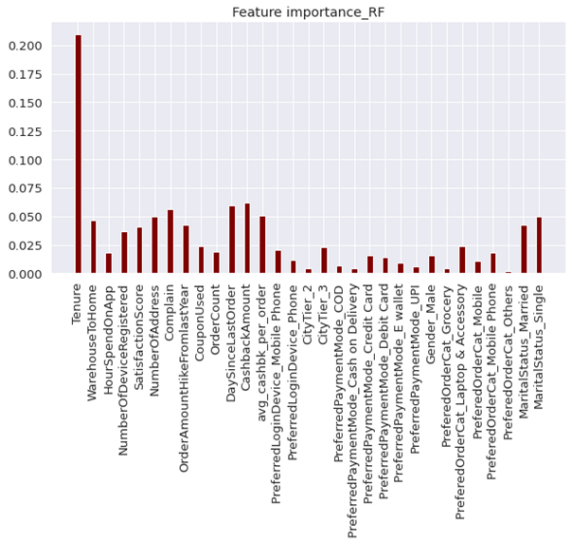
Feature Importance:

1. Below there are importance of features listed for the following models:
    1. XGBoost
    2. Logistic Regression
    3. LDA
    4. Random Forest
2. For XGBoost and Random Forest, the Feature Importance do not correlate with what was observed in the EDA. For instance, the as the Tenure increases, the churn rate as a percentage of the total decreases. Hence the Churn should show a high negative value which is depicted in the Logistic regression and LDA models.









A summary of the scores of the respective models:

1. Below we can see the scores of the various models built.
2. Based on the Senstivity and specificity scores, the AUC - RoC score and overall scores,
    - Random Forest, Decision Trees and KNN model seem to be highly overfit and hence might be a unreliable.
    - Logistic Regression, LDA and XGBoost seem to be doing well. However upon looking at the feature importance, XGBoost, doesn't look like a reliable one.

- **Between Logistic Regression and LDA models, Logistic is more accurate.**

3. **Logistic Regression Turns out to be the best models owing to the following reasons:**
    1. High AUC score
    2. Better Specificity
    3. Correct prediction of the feature importance.

.

Non Smoted Models	Total Number of observations:			5360				
Model	Model Score		AUC score		Senstivity	Specificity	Senstivity	Specificity
	Train		Train		Train		Test	
Logistic Regression	0.892	0.906	0.897	0.897	0.909	0.768	0.915	0.838
LDA	0.885	0.893	0.890	0.890	0.903	0.742	0.905	0.785
Descision Tree	1.000	0.962	1.000	1.000	1.000	1.000	0.984	0.862
Random Forest	1.000	0.967	1.000	1.000	1.000	1.000	0.968	0.966
KNN	0.936	0.903	0.979	0.979	0.939	0.919	0.910	0.842
XGBoost	0.882	0.879	0.877	0.877	0.905	0.711	0.899	0.717
Smoted Models	Total Number of observations:			9364				
Model	Model Score		AUC score		Senstivity	Specificity	Senstivity	Specificity
	Train		Train		Train		Test	
Logistic Regression	0.821	0.820	0.903	0.903	0.831	0.812	0.831	0.810
LDA	0.821	0.817	0.902	0.902	0.845	0.801	0.843	0.794
Descision Tree	1.000	0.964	1.000	1.000	1.000	1.000	0.967	0.960
Random Forest	1.000	0.985	1.000	1.000	1.000	1.000	0.990	0.981
KNN	0.946	0.917	1.000	1.000	1.000	0.903	1.000	0.857
XGBoost	0.829	0.833	0.911	0.911	0.808	0.852	0.813	0.855

6. Recommendations

1. Most important features are the following as per the models and the EDA are as follows:
    1. Tenure
    2. Payment methods - COD
    3. City tier2 and tier 3
    4. Complaints

Hence we can arrive at the following recommendations:

1. Focus should be on increasing the Tenure of the customer
2. We should try to reduce the complaints

3. The locations of warehouses should be reconsidered so as to keep the customer service intact
4. Right targeting of the customers in Tier 2 and 3 cities
5. More focus on customers who use CC, COD and E wallet mode of payment, since this is a very critical segment.
6. The Best possible time is to give some sort of promos at T=2 (2 units of tenure)