Customer Churn - E-Commerce

Capstone Project Report

Submitted by:

Karthiheswar M

Mentor guidance: Abhay Poddar

Batch: PGP-DSBA Sep'19

Date of Submission: 20 Sep'2020

Table of Contents

1 Introduction

E-Commerce (Electronic Commerce) is the activity of buying and selling of goods, products and online services over the internet. This also includes the sending and receiving of funds, inventory management and internet marketing. Business-to-Consumer (B2C) and Business-to-Business (B2B) are some of the important business transactions that can occur. E-Commerce is one of the hottest business over many industries like electronics, fashions, grocery, furniture, medicals, foods and etc.

1.1 Problem Statement

Since E-Commerce is one of the most important business that is happening, there few problems that can occur in this business. Among those problems customer Churn rate is one of the most important factor to be considered, as the major part of sales and profit depends on it. So in this problem, few necessary steps and precautions have been taken to predict the customer Churn rate.

1.2 Need of the Project study

Predicting the customer Churn rate helps the company to decide the right path to proceed as they can evaluate their feedback with the past Churn rate data. This also helps in identifying the reasons for the customer to Churn, also some indications that the customer may Churn.

1.3 Business/Social Opportunity

To predict the Churn rate of the E-Commerce company, first the company's dataset have to be explored to find the insights that are helpful in predicting customer Churn rate. To do so, initially the raw data has to be pre-processed with the required techniques. Ultimately this project predicts the customer Churn rate so that the company can turn up with some promos to offer to the customers and can do their marketing strategies accordingly.

2 Data Report

This report consists of the data from E-Commerce company, where the data are analyzed, explored and the insights are described with the necessary plots for the visualization of data. All these insights will be very supportive in predicting the customer Churn rate.

2.1 Collection of Data

The data is collected from a E-Commerce company which works on electronics, grocery, fashion and few others online shopping. This is the data source for this project with a customer with maximum tenure of 61 months and a customer with 0 value on days since last order which means the latest day.

2.2 Visual Inspection of Data

The data has 5630 of observations with 20 variables, where there are few null values in some of the variable's observations. The independent variables have both numerical and

categorical data where the Churn variable is considered as target variable. The below table describes the numerical data with few necessary details:

Variable name	*Count*	*Mean*	*Std*	*Min*	*25%*	*50%*	*75%*	*Max*
Churn	5630.0	0.168384	0.374240	0.0	0.00	0.00	0.0000	1.00
Tenure	5366.0	10.189899	8.557241	0.0	2.00	9.00	16.0000	61.00
CityTier	5630.0	1.654707	0.915389	1.0	1.00	1.00	3.0000	3.00
WarehouseToHome	5379.0	15.639896	8.531475	5.0	9.00	14.00	20.0000	127.00
HourSpendOnApp	5375.0	2.931535	0.721926	0.0	2.00	3.00	3.0000	5.00
NumberOfDeviceRegistered	5630.0	3.688988	1.023999	1.0	3.00	4.00	4.0000	6.00
SatisfactionScore	5630.0	3.066785	1.380194	1.0	2.00	3.00	4.0000	5.00
NumberOfAddress	5630.0	4.214032	2.583586	1.0	2.00	3.00	6.0000	22.00
Complain	5630.0	0.284902	0.451408	0.0	0.00	0.00	1.0000	1.00
OrderAmountHikeFromlastYear	5365.0	15.707922	3.675485	11.0	13.00	15.00	18.0000	26.00
CouponUsed	5374.0	1.751023	1.894621	0.0	1.00	1.00	2.0000	16.00
OrderCount	5372.0	3.008004	2.939680	1.0	1.00	2.00	3.0000	16.00
DaySinceLastOrder	5323.0	4.543491	3.654433	0.0	2.00	3.00	7.0000	46.00
CashbackAmount	5630.0	177.223030	49.207036	0.0	145.77	163.28	196.3925	324.99

2.3 Understanding of Attributes

The data has some of the details on customer's transaction history. On general observation, Tenure, Complain and DaySinceLastOrder are some of the important independent variables where the dependent variable usually depends on in such cases. Few observations under some of the variables are same with different names, so those entities have to be merged. Those are as follows:

PreferredLoginDevice: The Mobile Phone and Phone are same entities, so both are merged as Phone.

PreferredPaymentMode: The Cash on Delivery and COD are same entities, so both are merged as COD. Also CC and Credit Card are same entities, so both are merged as Credit Card.

PreferedOrderCat: The Mobile and Mobile Phone are same entities, so both are merged as Mobile.

There are 15 continues variables and 5 categorical variables present in the raw data. There are no duplicates present in the data. Also little amount of skewness are present in all continues variables as follows:

Variable name	*Skewness*
Churn	1.772843
Tenure	0.736513
CityTier	0.735326
WarehouseToHome	1.619154
HourSpendOnApp	-0.027213
NumberOfDeviceRegistered	-0.396969
SatisfactionScore	-0.142626
NumberOfAddress	1.088639
Complain	0.953347
OrderAmountHikeFromlastYear	0.790785
CouponUsed	2.545653
OrderCount	2.196414
DaySinceLastOrder	1.191000
CashbackAmount	1.149846

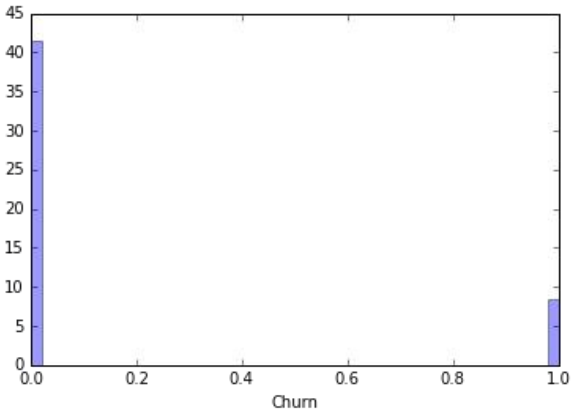
3 Exploratory Data Analysis

Exploratory data analysis is the important factor to find the insights in data. This can be done in various methods as following:

3.1 Univariate Analysis

It is the graphical representation of how a variable is distributed. The Univariate analysis is done on numerical data and categorical data separately. The below plots shows the distribution of data among the respective variables.

Dependent Variable:

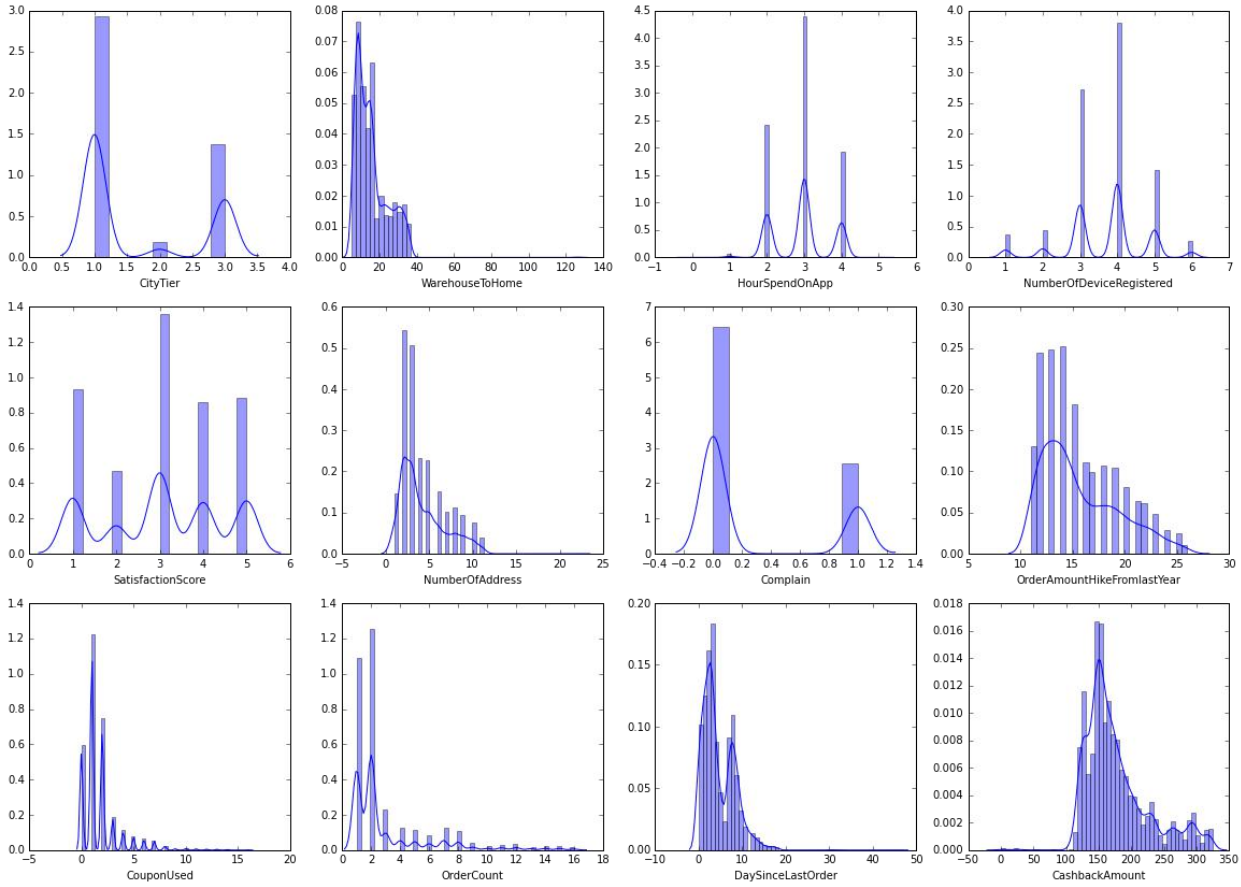


Insights

- The 0 value represents that customers not churned and 1 value represents the churned customers.
- Customers who churned are very less compared to customers who are not churned.

Independent Variables:

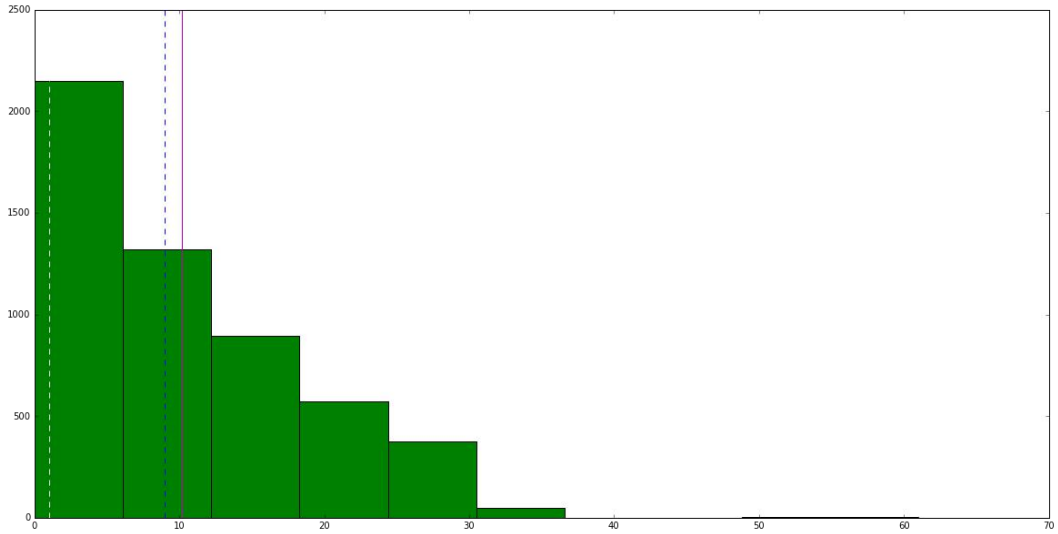
3.1.1 Numeric Variables:



Insights

- Most of the customers are from tier-1 cities where tier-2 cities have least customers.
- Many customers spent about 3 hours on company's app and also there few customers who spent 5 hours on company's app which is found as maximum hour spent.
- There are only few complaints raised in last month comparatively.
- The data are widely distributed in Satisfaction score, Order amount hike from last year, Days since last order and Cashback amount.

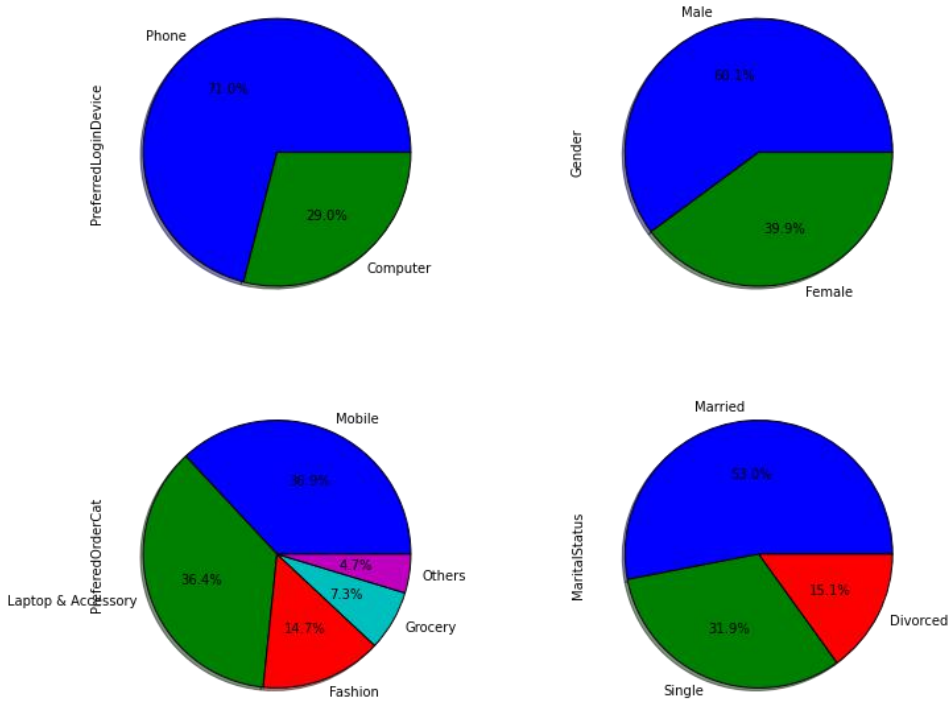
Tenure:



Insights

- The distribution of the variable is right skewed.
- The average tenure of the customer is around 10 months.
- Most of the customers are new as tenure is around 1 month which means recently joined.
- The maximum tenure that a customer has is 61 months.

3.1.2 Categorical Variables:

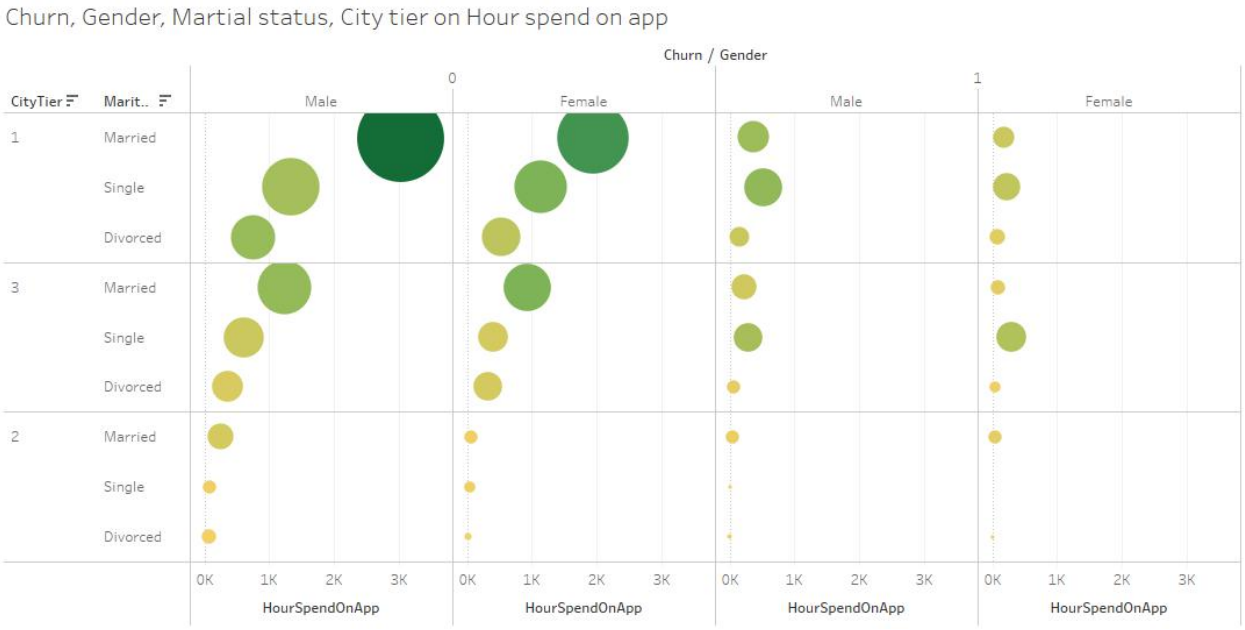


Insights

- Customer prefer to login in Phone than Computer as they may found easy to access.
- Male customers are more than female customers.
- We can see that frequency of post publishing increases daily from Monday, reaches its maximum point on Wednesday and then gradually declines.
- The base time frequency is showing similar patterns, it is maximum on Thursdays and then declining further.
- So, this may be an inference for business to think some other way to engage people during weekends rather than Facebook promotions.

3.2 Bivariate Analysis

This method of analysis describes the relationship between the variables.

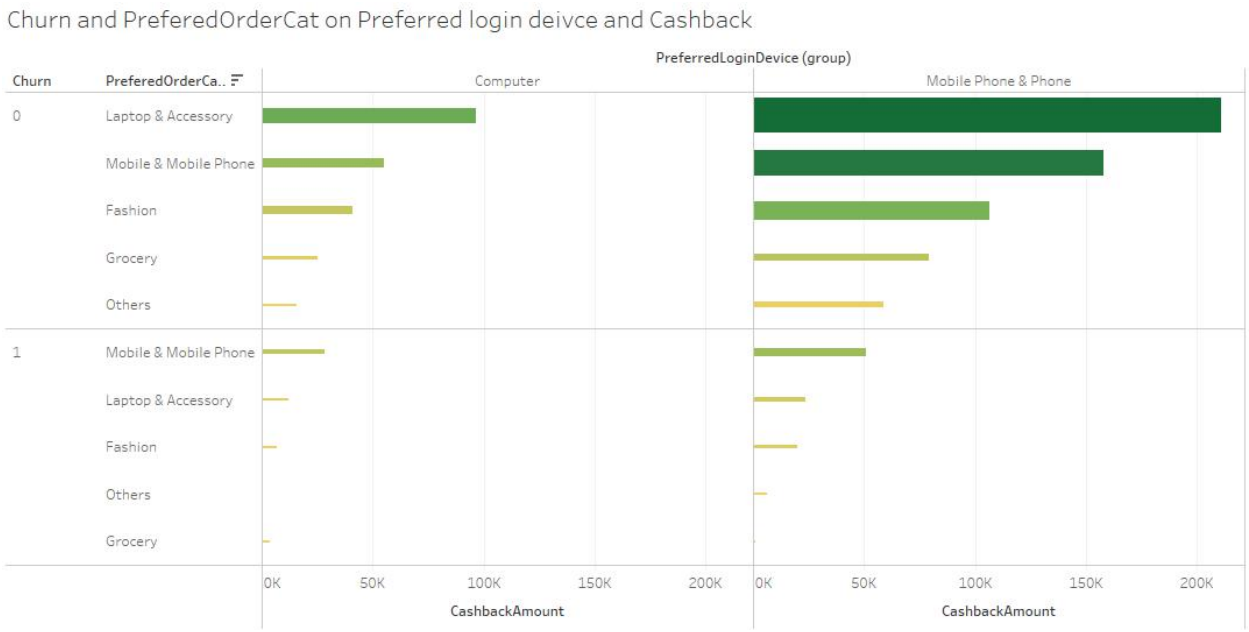


Description

- Darken the color, higher the number of complaints.
- Bigger the size, higher the satisfaction score.

Insights

- From above plot, customers from tier-1 city spent much hour on company's app, followed by tier-3 city.
- Also customers from tier-1 city have raised much complaints but also rated higher satisfaction score followed by tier-3 city.
- Customers from tier-1 cities are more active where the customers from tier-2 cities are least active as their number of complaints raised, satisfaction score and hours spent on app are very less.
- Customers who fail more complaints tends to churn.
- Customers who are married spend more hours on app, also they have raised much complaints and higher satisfaction score, also it compiles with the male customer.

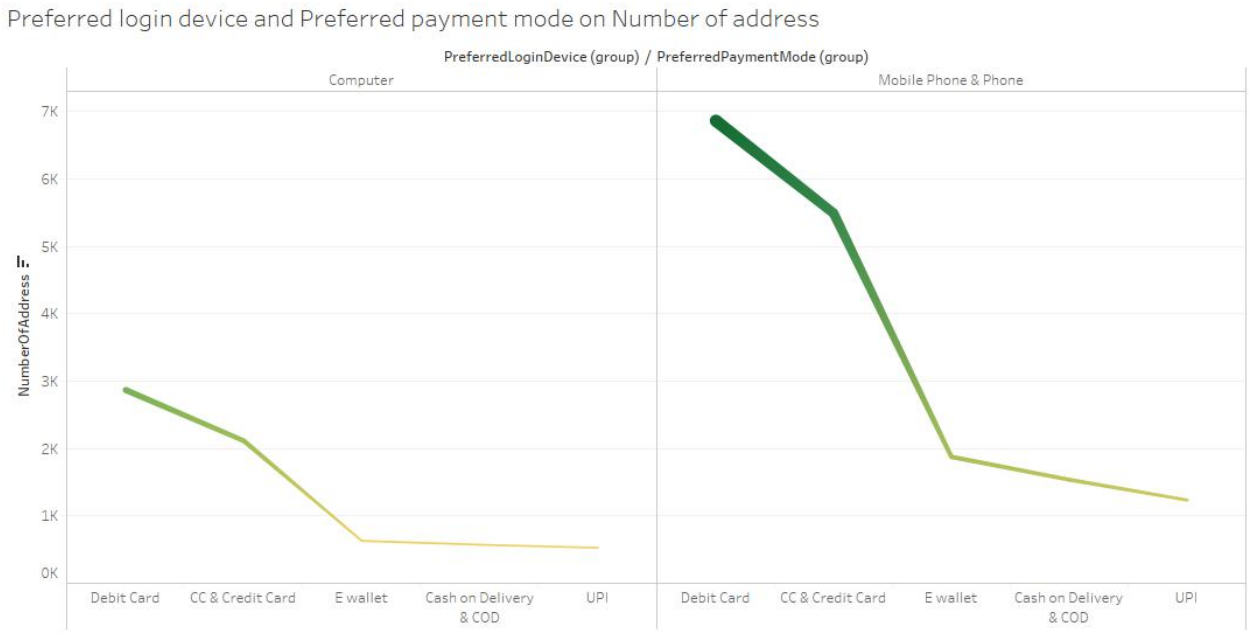


Description

- Darken the color, higher the Order amount hike from last year.
- Bigger the size, more number of Coupons used.

Insights

- Laptop and accessories are the most Preferred order category followed by Mobile Phones, where groceries and others are least Preferred order category.
- Also more number of Coupons used on Laptop and accessories followed by Mobile Phones, where groceries and others are least on which Coupons used.
- Since Laptop and accessories are most Preferred order category, the Cashback amount is huge on Laptop and accessories followed by Mobile Phones.



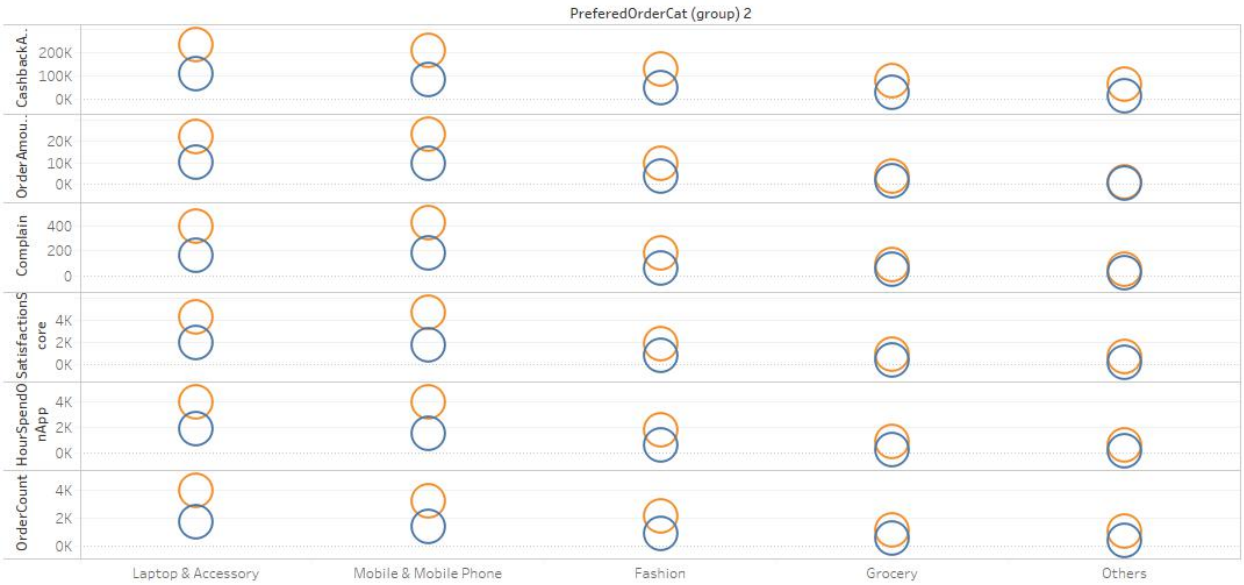
Description

- Darken the color, higher the Number of devices registered.
- Bigger the size, more the number of Days since last order.

Insights

- Customers with more Number of address tends to register in more Number of devices, also these customers have higher number of Days since last order and tend to pay in Debit card.
- Customers with less number of Days since last order tends to pay through UPI as they have less Number of devices registered and have less Number of address.

PreferedOrderCat on Cashback, Order amount hike, Complain, Satisfaction score and Hour spend on app



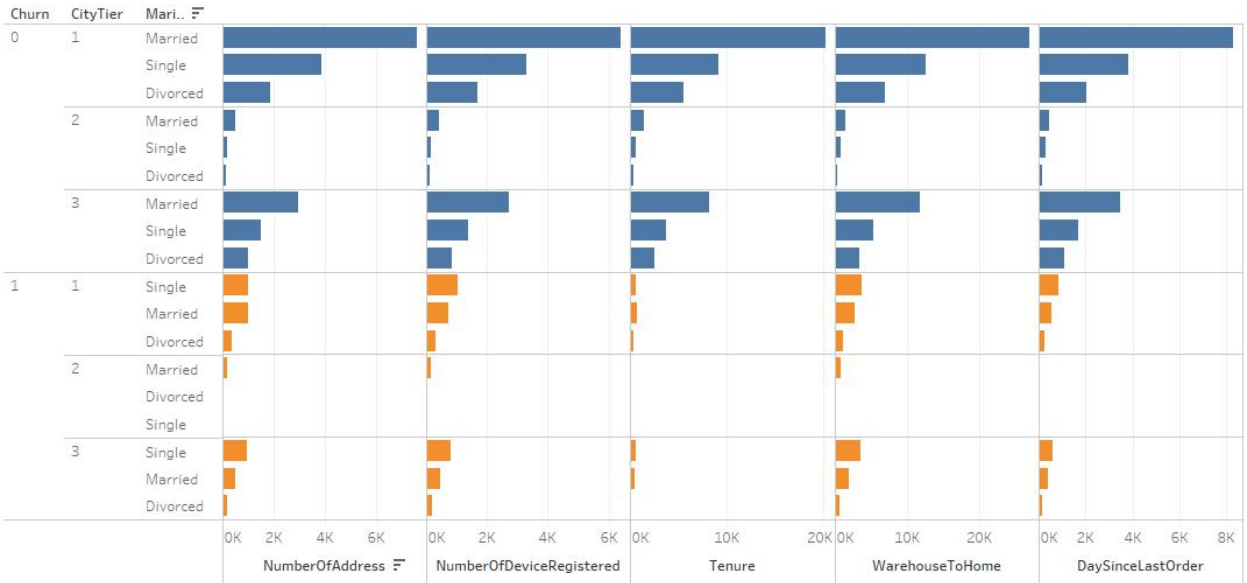
Description

- Orange color indicates Mobile phone and Blue color indicates Computer in Preferred login device.

Insights

- Laptop and accessories are having higher values on almost all variables like Order count, Satisfaction score, Complain, Order amount hike from last year and Cashback amount followed by Mobile Phones, where groceries and others have least values.

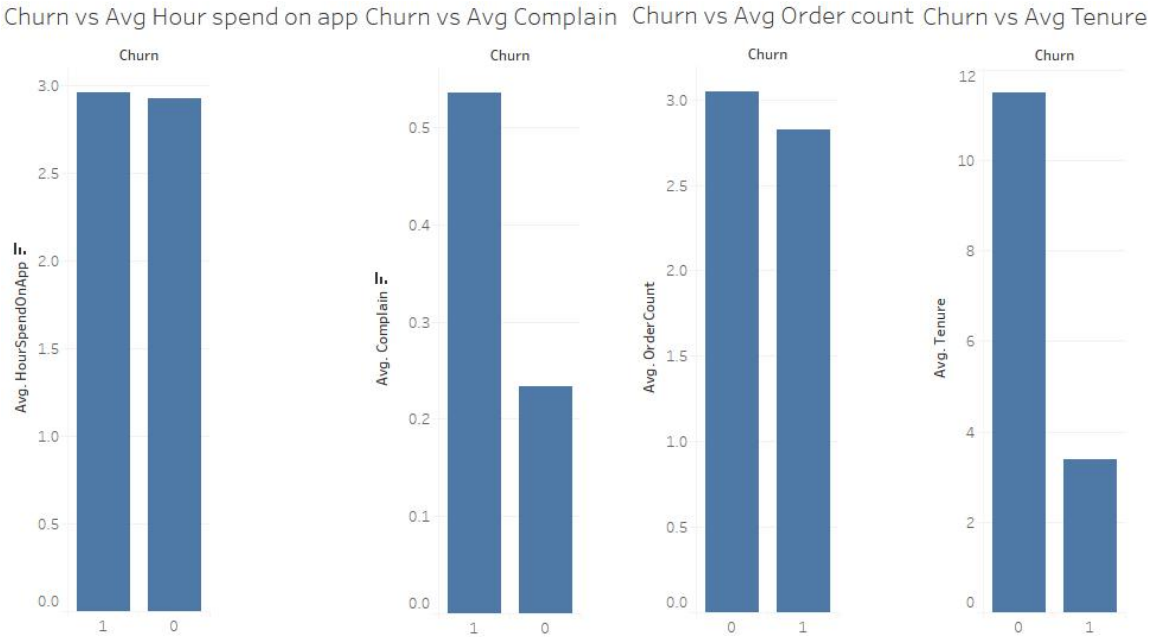
Churn on No of address, No of device registered, Tenure, Wharehouse to home and Day since last order

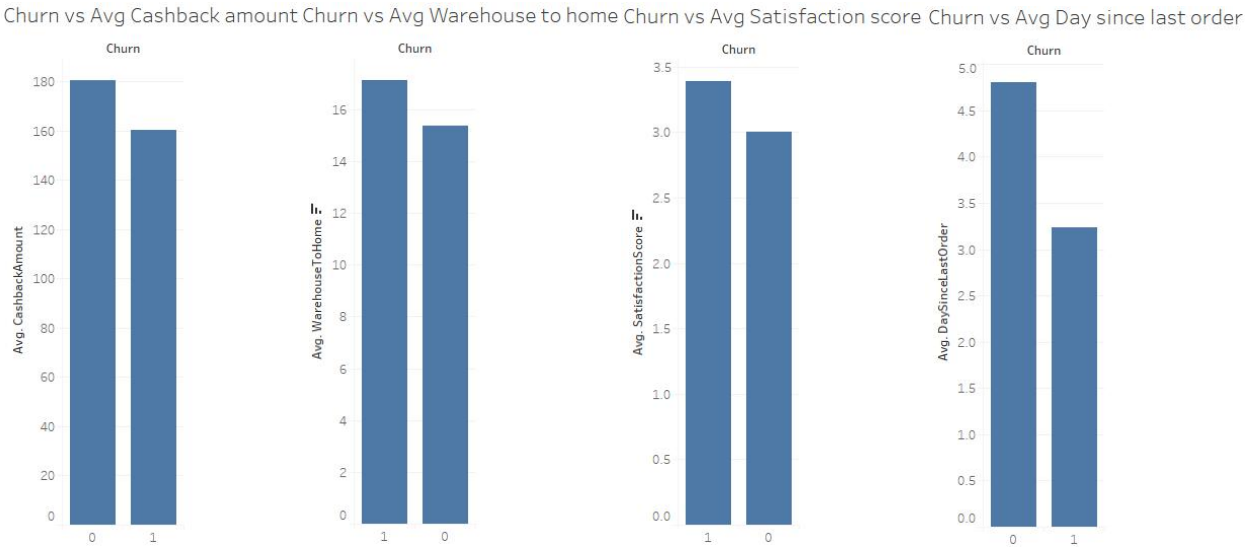


Insights

- Customers from tier-1 cities who are married have more Number of address, Number of device registered, Tenure, distance from Warehouse to home and Days since last order followed by customers from tier-3 cities.

<u>Target variable vs Numerical variables:</u>



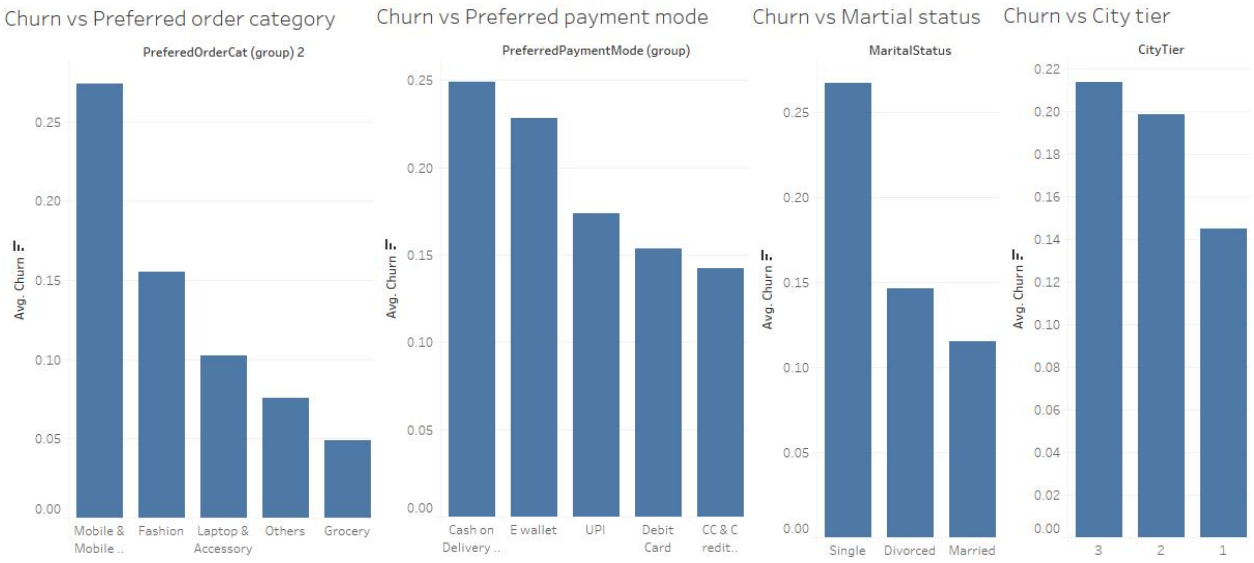


Insights

- Customers spending hours on app doesn't decide the Churn rate and Orders count also doesn't shows much difference on Churn rate.
- Customers who raised more complaints tends to churn, also customers with less Tenure Churn's lot and on other hand the churned customers have high Satisfaction score.

- The customers with comparatively less Cashback amount, more distance from Warehouse to home and surprisingly with more Satisfaction score and less number of Days since last order are tending to Churn.

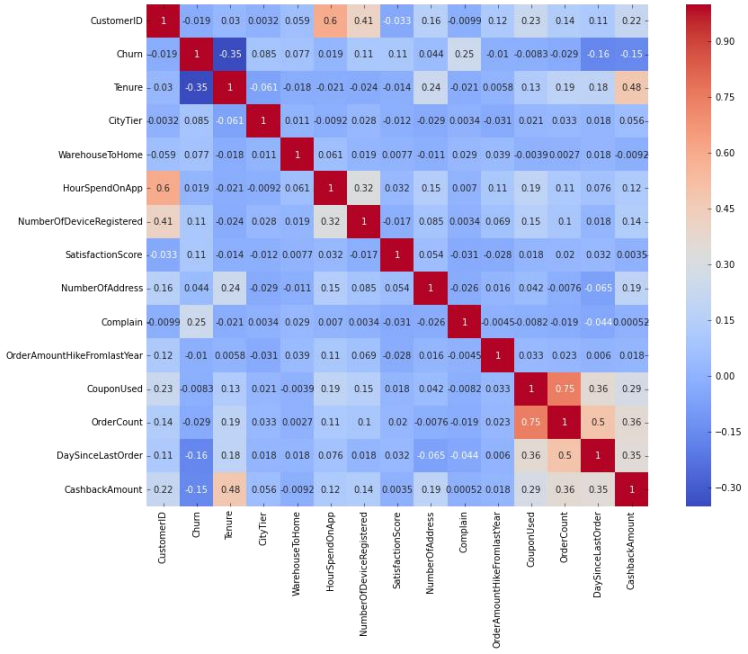
Target variable vs Categorical variables:



Insights

- Customers who are single Churns more followed by divorced customers.
- Customers from from tier-1 cities churns less comparatively also they pay through Cash on Delivery.

Correlation plot:



Insights

- Churn variable is highly correlated with Tenure and Complain variable and least correlated with Coupon used.
- Also Churn is decently correlated with Days since last order, Cashback amount, Satisfaction score and Number of device registered.
- Some independent variables are highly correlated with each other.
- Order count is highly correlated with Coupon used and Days since last order.
- Also Cashback amount is having a decent positive correlation with Tenure, Coupon used, Order count and Days since last order which means higher the Tenure, Coupon used, Order count and Days since last order higher the Cashback amount.

3.3 Removal of Unwanted Variables

The variable CouponUsed has some good correlation with OrderCount and also very less correlated with the dependent variable, so CouponUsed variable can be dropped from the dataset. Also CustomerID variable has to be dropped at model building stage and for outlier treatment.

3.4 Missing Value Treatment

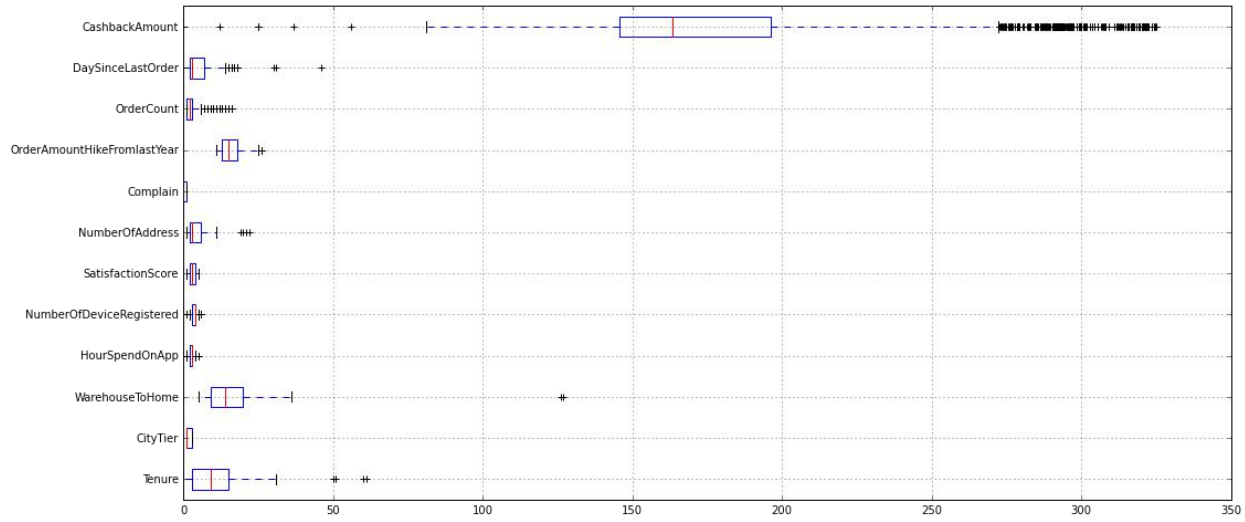
Missing values or null values are a common occurrence in a dataset which cause a significant effect and also it will be a barrier to build a good model, so these missing values have to be treated accordingly. There are 1600 missing values present in the dataset. The following table list the number of missing values with respect to their variables:

Variable name	*Missing values*
Tenure	264
WarehouseToHome	251
HourSpendOnApp	255
OrderAmountHikeFromlastYear	265
OrderCount	258
DaySinceLastOrder	307

These missing values are treated with their median values as all null values are present in the continues variables.

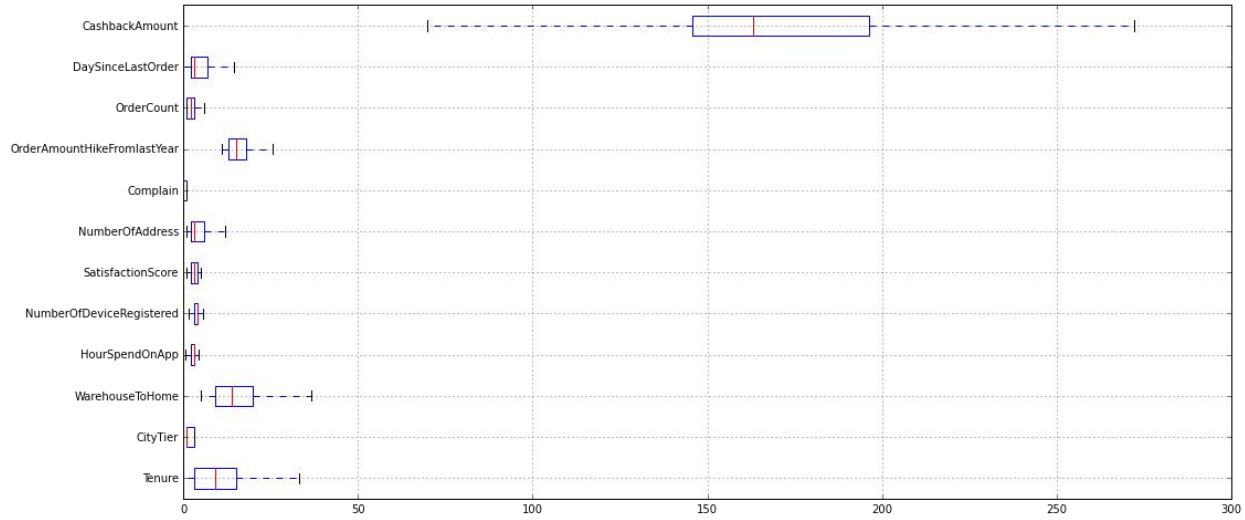
3.5 Outliers Treatment

Outliers are the values that are present far from the remaining observations and it also causes significant difference at model building. So these outliers have to be treated to build a better model.



The above plot clearly indicates there are too much of outliers present in the dataset and outliers are present in almost every continues variables. Removing of outliers can cause some huge loss in data, so instead these outliers are imputed.

These outliers are treated by finding Inter Quartile Range (IQR) of upper range value and lower range value for all continues variables. So the observations above Upper range of IQR are replaced with upper range value and the observations below Lower range of IQR are replaced with lower range value. The below plot is the plot after treated from the outliers:



3.6 Variable Transformation

The dataset doesn't require any scaling and normalization as there many categorical variables present and the continues variables are of nearly in same magnitude.

The CityTier variable is shown as numerical variable but it should be converted as categorical variable as it describes the type of the city.

One hot encoding is not done as it creates lot more variables, instead all the categorical variables are label encoded which means the entities are turned up with the codes and

made as numerical variable for model building. The below table are the top 5 observation of dataset after label encoding:

	0	1	2	3	4
CustomerID	50001.00	50002.0	50003.00	50004.00	50005.0
Churn	1.00	1.0	1.00	1.00	1.0
Tenure	4.00	9.0	9.00	0.00	0.0
PreferredLoginDevice	1.00	1.0	1.00	1.00	1.0
CityTier	2.00	2.0	2.00	2.00	2.0
WarehouseToHome	6.00	8.0	30.00	15.00	12.0
PreferredPaymentMode	2.00	4.0	2.00	2.00	1.0
Gender	0.00	1.0	1.00	1.00	1.0
HourSpendOnApp	3.00	3.0	2.00	2.00	3.0
NumberOfDeviceRegistered	3.00	4.0	4.00	4.00	3.0
PreferedOrderCat	2.00	3.0	3.00	2.00	3.0
SatisfactionScore	2.00	3.0	3.00	5.00	5.0
MaritalStatus	2.00	2.0	2.00	2.00	2.0
NumberOfAddress	9.00	7.0	6.00	8.00	3.0
Complain	1.00	1.0	1.00	0.00	0.0
OrderAmountHikeFromlastYear	11.00	15.0	14.00	23.00	11.0
OrderCount	1.00	1.0	1.00	1.00	1.0
DaySinceLastOrder	5.00	0.0	3.00	3.00	3.0
CashbackAmount	159.93	120.9	120.28	134.07	129.6

4 Business insights from EDA

Out of all the analysis done with dataset, the ideas and information that drives the business are the insights.

4.1 Checking whether the data is balanced

In the given dataset there are totally 4682 customers not churned and 948 customers churned. So among the given dataset 16.838% of customers are churned. So there is no required of SMOTE technique.

4.2 Clustering

The customers based on their behaviour, they are divided into 5 groups as following table:

Variable name	*Row1*	*Row2*	*Row3*	*Row4*	*Row5*
Cluster	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
CustomerID (Count)	1313	1059	1913	646	699
Churn (Unique concatenate with count)	1(162), 0(1151)	1(288), 0(771)	1(360), 0(1553)	1(32), 0(614)	1(106), 0(593)
Tenure (Mean)	9.713632902	6.780925401	7.792472556	19.87925697	13.27753934
PreferredLoginDevice (Unique concatenate with count)	Phone(927), Computer(386)	Phone(749), Computer(310)	Phone(1328), Computer(585)	Phone(496), Computer(150)	Phone(496), Computer(203)

CityTier (Unique concatenate with count)	3(601), 1(678), 2(34)	1(825), 3(171), 2(63)	3(513), 1(1311), 2(89)	1(452), 3(156), 2(38)	1(400), 3(281), 2(18)
WarehouseToHome (Mean)	16.12947449	14.43862134	15.6589127	14.78328173	16.43347639
PreferredPaymentMode (Unique concatenate with count)	Credit Card(365), E wallet(216), Debit Card(537), COD(108), UPI(87)	UPI(86), Debit Card(454), Credit Card(352), COD(125), E wallet(42)	Debit Card(782), E wallet(177), Credit Card(617), COD(174), UPI(163)	COD(37), Debit Card(267), Credit Card(231), E wallet(62), UPI(49)	Credit Card(209), Debit Card(274), E wallet(117), UPI(29), COD(70)
Gender (Unique concatenate with count)	Male(761), Female(552)	Male(674), Female(385)	Female(744), Male(1169)	Female(275), Male(371)	Female(290), Male(409)
HourSpendOnApp (Mean)	3.135186596	2.617563739	2.948248824	2.925696594	3.009298999
NumberOfDeviceRegistered (Mean)	3.891850724	3.260623229	3.705697857	3.748452012	3.908440629
PreferedOrderCat (Unique concatenate with count)	Fashion(235), Laptop & Accessory(1021), Mobile(57)	Mobile(992), Laptop & Accessory(65), Grocery(2)	Laptop & Accessory(877), Mobile(1031), Fashion(5)	Others(264), Grocery(334), Fashion(48)	Fashion(538), Grocery(74), Laptop & Accessory(87)
SatisfactionScore (Mean)	3.038080731	3.058545798	3.07997909	3.080495356	3.084406295
MaritalStatus (Unique concatenate with count)	Single(427), Divorced(187), Married(699)	Single(396), Divorced(148), Married(515)	Single(635), Divorced(277), Married(1001)	Divorced(124), Single(154), Married(368)	Divorced(112), Single(184), Married(403)
NumberOfAddress (Mean)	4.4843869	3.248347498	4.11134344	4.944272446	4.726752504
Complain (Mean)	0.275704494	0.282341832	0.286983795	0.275541796	0.309012876
OrderAmountHikeFromlastYear (Mean)	15.81188119	15.37110482	15.75352849	15.37616099	15.91273247
OrderCount (Mean)	2.753236862	1.634560907	2.349189754	3.383900929	3.097281831
DaySinceLastOrder (Mean)	5.17136329	2.551463645	4.029534762	6.723684211	4.852646638
CashbackAmount (Mean)	179.6992003	126.3645881	152.1284527	268.3160043	218.821731

The following table groups the customers based on their Churn rate shows their behaviour on every variables and clearly defines on every aspect:

Variable name	***Row1***	***Row2***
Churn	Churn_0	Churn_1
CustomerID (Count)	4682	948

Tenure (Mean)	11.38530543	3.859704641
PreferredLoginDevice (Unique concatenate with count)	Phone(3372), Computer(1310)	Phone(624), Computer(324)
CityTier (Unique concatenate with count)	3(1354), 1(3134), 2(194)	3(368), 1(532), 2(48)
WarehouseToHome (Mean)	15.26719351	16.85654008
PreferredPaymentMode (Unique concatenate with count)	E wallet(474), Debit Card(1958), COD(386), Credit Card(1522), UPI(342)	Debit Card(356), UPI(72), Credit Card(252), COD(128), E wallet(140)
Gender (Unique concatenate with count)	Male(2784), Female(1898)	Female(348), Male(600)
HourSpendOnApp (Mean)	2.928662965	2.964135021
NumberOfDeviceRegistered (Mean)	3.650683469	3.916666667
PreferedOrderCat (Unique concatenate with count)	Fashion(698), Laptop & Accessory(1840), Mobile(1510), Others(244), Grocery(390)	Laptop & Accessory(210), Mobile(570), Others(20), Fashion(128), Grocery(20)
SatisfactionScore (Mean)	3.001281504	3.390295359
MaritalStatus (Unique concatenate with count)	Divorced(724), Married(2642), Single(1316)	Single(480), Divorced(124), Married(344)
NumberOfAddress (Mean)	4.15890645	4.450421941
Complain (Mean)	0.234087997	0.535864979
OrderAmountHikeFromlastYear (Mean)	15.68272106	15.61708861
OrderCount (Mean)	2.543571123	2.407172996
DaySinceLastOrder (Mean)	4.68058522	3.187236287
CashbackAmount (Mean)	178.5006413	159.6365032

4.3 Other Business Insights

- From all above data analysis, it's clear that customers with less Tenure churns more.
- Majority of the customers are Male and in general the majority of the customer's Martial status are Married.
- From the cluster table, Satisfaction score doesn't vary much with churned and not churned customers.
- Most customers prefer to shop Mobile Phones and Laptop and accessories where Groceries and Others sells least.
- Customers from tier-2 cities are the least number of customers, so there is no enough reach in tier-2 cities.

5 Model Building and Interpretation

Once the insights are derived from Exploratory Data Analysis the next step is to build the models for the Churn prediction from the given dataset, models result are interpreted to find the best suited model.

5.1 Building Various Models

Various models are built using various machine learning algorithms.

5.1.1 Data Split

Initially the processed dataset is splitted into 2 following subsets by dropping CustomerID variable:

- Training Data: This subset has 70% of data which is for model building using machine learning algorithm.
- Testing data: This subset has remaining 30% of data where the built models are tested on test data.

5.1.2 Machine Learning Algorithms

The target variable taken is Churn and since it is binary variable, many classification algorithms are considered to the built the models. The following algorithm techniques are:

1. **Logistic Regression:** This algorithm is the basic machine learning algorithm of classification technique, using regression technique it establishes the relation between independent variable and dependent variable.
2. **Linear Discriminant Analysis:** LDA uses linear combinations of independent variables to predict the class in the response variable of a given observation.
3. **K-Nearest Neighbors:** KNN works based on feature similarity. It calculates the similarities or distance of test query from each point in train subset.
4. **Naive Bayes:** It is based on the principle of probability where probability of an event which is actually based on the preceding values of the event. It assumes that the input features are independent from each other.
5. **Support Vector Machine:** The principle of SVM is to find an hyperplane which, can classify the training data points in to labeled categories. The input of SVM is the training data and use this training sample point to predict class of test point.
6. **Artificial Neural Network:** ANN works based on number of neurons and number of hidden layers assigned. It calculates the weightage of independent variables on neurons and hidden layers assigned.

5.2 Performance Metrics

The predictive models are built out of training data by applying various machine learning techniques, these predictive models are tested against testing data and their performance are determined with some metrics.

5.2.1 Logistic Regression:

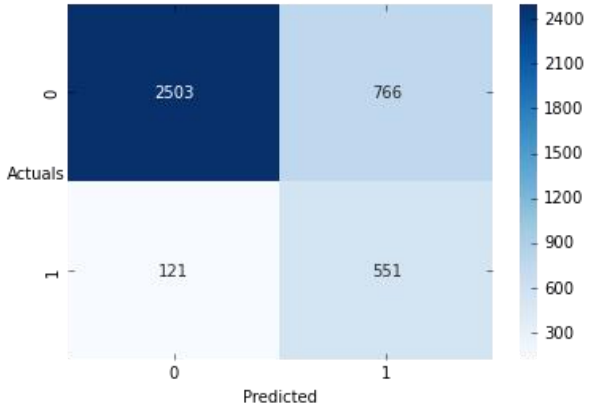
1. **Accuracy of train data:**

   0.7749

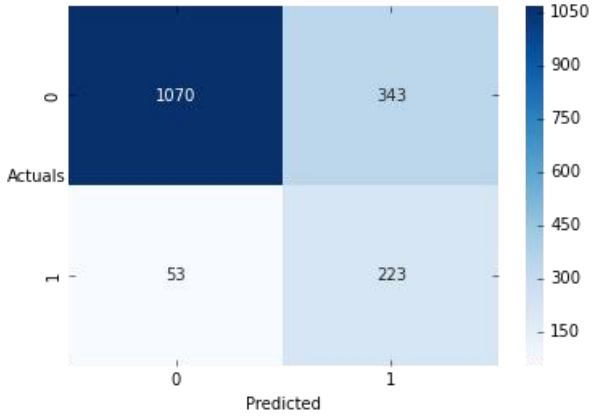
2. **Accuracy of test data:**

   0.7655

3. **Confusion matrix on train data:**



4. **Confusion matrix on test data:**



5. **Classification report on train data:**

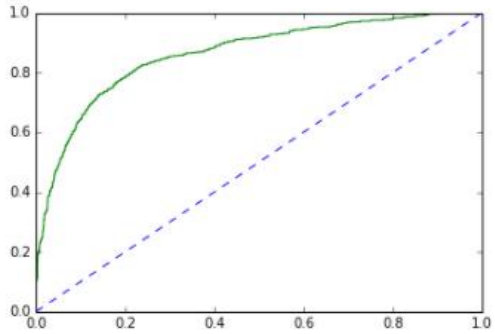
Dependent variable	Precision	Recall	F1-score	Support
0	0.95	0.77	0.85	3269
1	0.42	0.82	0.55	672

6. **Classification report on test data:**

Dependent variable	Precision	Recall	F1-score	Support
0	0.95	0.76	0.84	1413
1	0.39	0.81	0.53	276

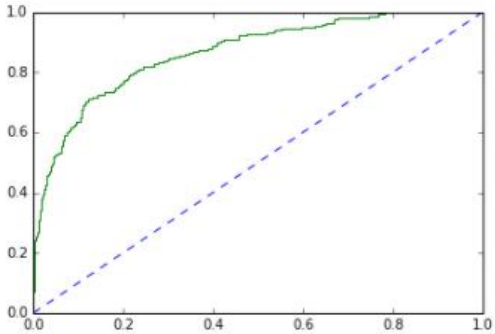
7. **AUC on train data:**

```
AUC: 0.862
[<matplotlib.lines.Line2D at 0x27055ab8d0>]
```



8. **AUC on test data:**

```
AUC: 0.868
[<matplotlib.lines.Line2D at 0x270566b470>]
```



5.2.2 Linear Discriminant Analysis:

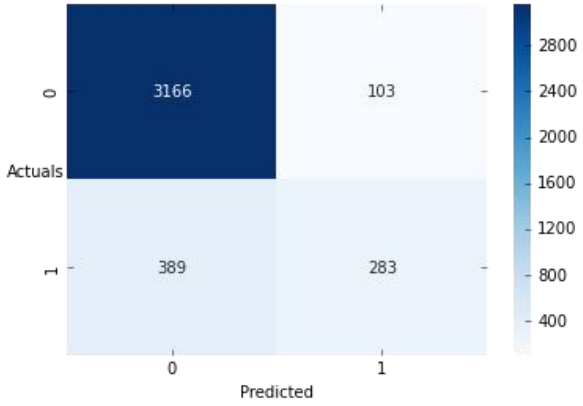
1. **Accuracy of train data:**

   0.8751

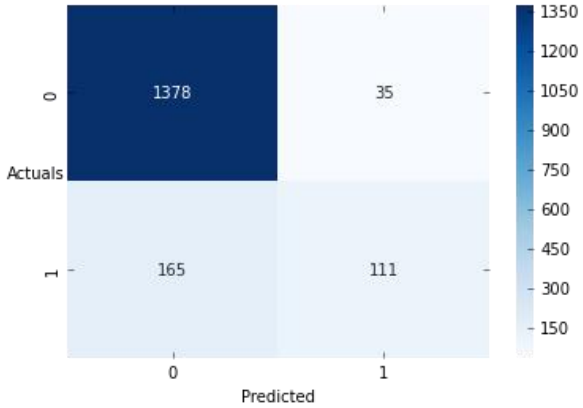
2. **Accuracy of test data:**

   0.8815

3. **Confusion matrix on train data:**



4. Confusion matrix on test data:



5. Classification report on train data:

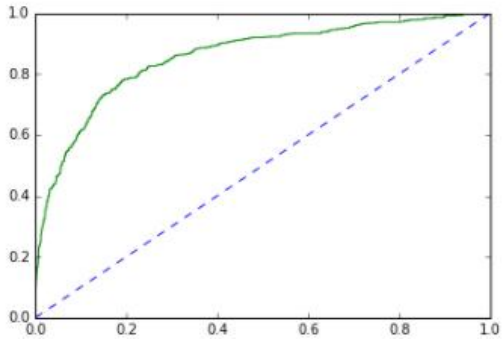
Dependent variable	Precision	Recall	F1-score	Support
0	0.89	0.97	0.93	3269
1	0.73	0.42	0.53	672

6. Classification report on test data:

Dependent variable	Precision	Recall	F1-score	Support
0	0.89	0.98	0.93	1413
1	0.76	0.40	0.53	276

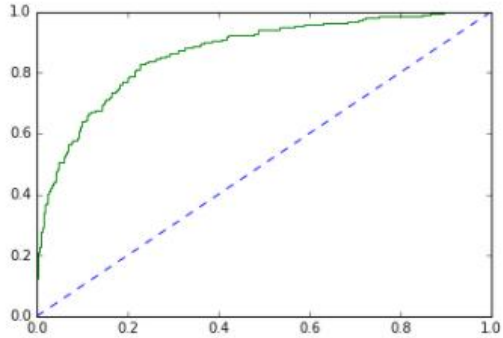
7. AUC on train data:

```
AUC: 0.857

[<matplotlib.lines.Line2D at 0x2708293e80>]
```



8. **AUC on test data:**

```
AUC: 0.869

[<matplotlib.lines.Line2D at 0x27082be128>]
```



5.2.3 K-Nearest Neighbours:

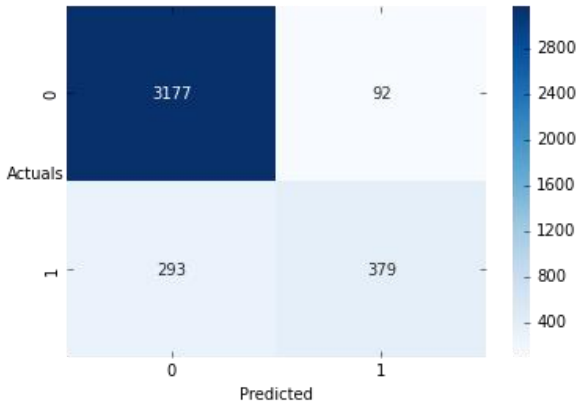
1. **Accuracy of train data:**

   0.9023

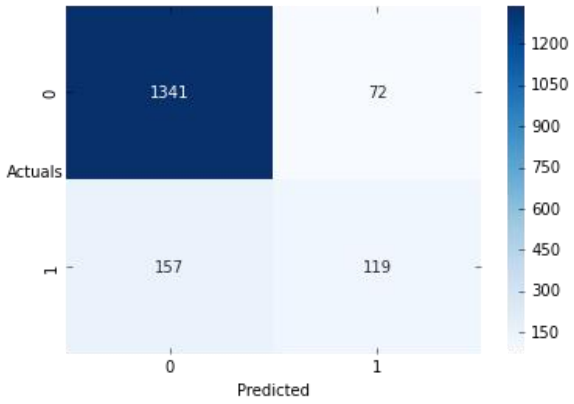
2. **Accuracy of test data:**

   0.8644

3. **Confusion matrix on train data:**



4. **Confusion matrix on test data:**



5. **Classification report on train data:**

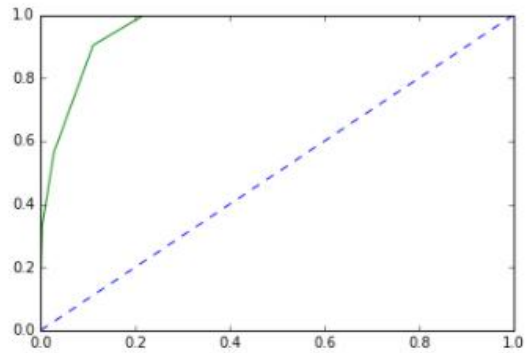
Dependent variable	Precision	Recall	F1-score	Support
0	0.92	0.97	0.94	3269
1	0.80	0.56	0.66	672

6. **Classification report on test data:**

Dependent variable	Precision	Recall	F1-score	Support
0	0.90	0.95	0.92	1413
1	0.62	0.43	0.51	276

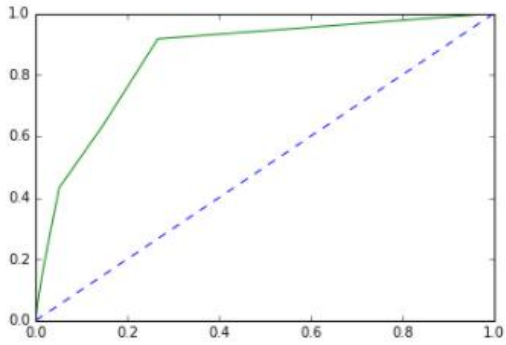
7. **AUC on train data:**

```
AUC: 0.956

[<matplotlib.lines.Line2D at 0x2708407710>]
```



8. **AUC on test data:**

```
AUC: 0.859

[<matplotlib.lines.Line2D at 0x2708498470>]
```



5.2.4 Naive Bayes:

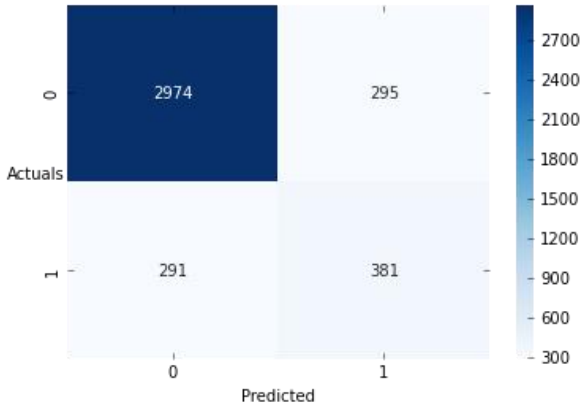
1. **Accuracy of train data:**

   0.8513

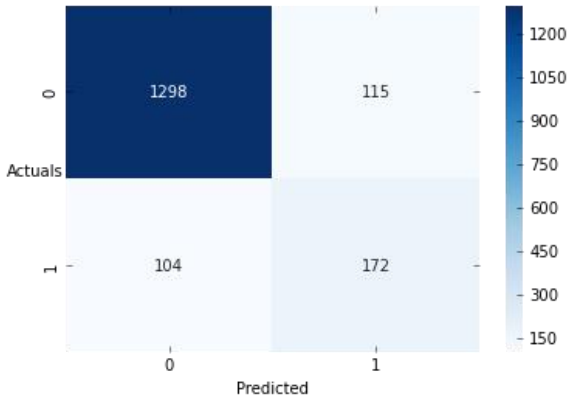
2. **Accuracy of test data:**

   0.8703

3. **Confusion matrix on train data:**



4. **Confusion matrix on test data:**



5. Classification report on train data:

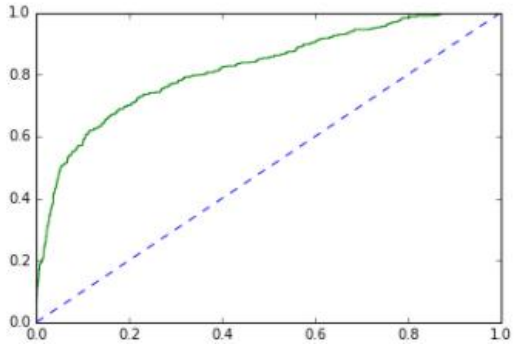
Dependent variable	Precision	Recall	F1-score	Support
0	0.91	0.91	0.91	3269
1	0.56	0.57	0.57	672

6. Classification report on test data:

Dependent variable	Precision	Recall	F1-score	Support
0	0.93	0.92	0.92	1413
1	0.60	0.62	0.61	276

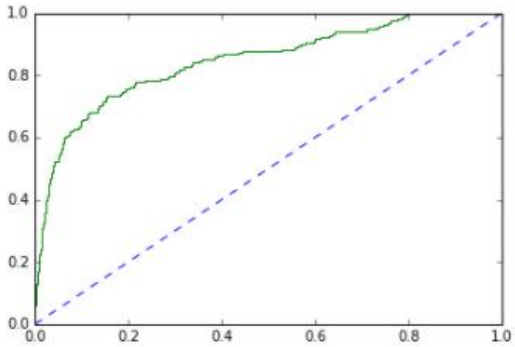
7. AUC on train data:

```
AUC: 0.821

[<matplotlib.lines.Line2D at 0x270857ba58>]
```



8. AUC on test data:

```
AUC: 0.847

[<matplotlib.lines.Line2D at 0x270857b278>]
```



5.2.5 Support Vector Machine:

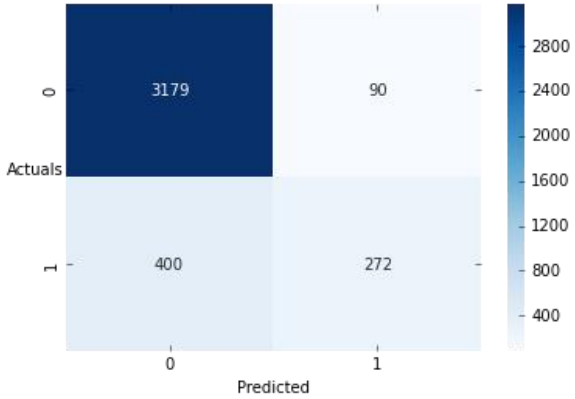
1. **Accuracy of train data:**

   0.8756

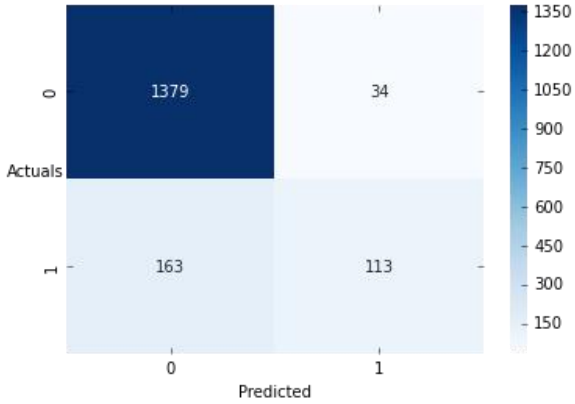
2. **Accuracy of test data:**

   0.8833

3. **Confusion matrix on train data:**



4. **Confusion matrix on test data:**



5. **Classification report on train data:**

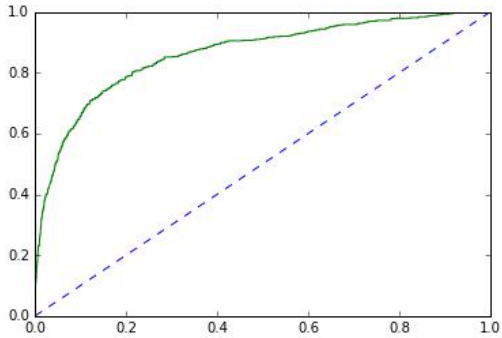
Dependent variable	Precision	Recall	F1-score	Support
0	0.89	0.97	0.93	3269
1	0.75	0.40	0.53	672

6. **Classification report on test data:**

Dependent variable	Precision	Recall	F1-score	Support
0	0.89	0.98	0.93	1413
1	0.77	0.41	0.53	276

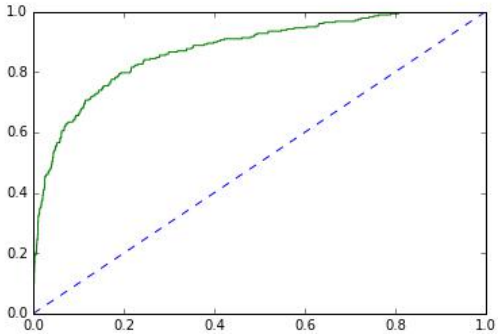
7. **AUC on train data:**

```
AUC: 0.863

[<matplotlib.lines.Line2D at 0x4256de2f60>]
```



8. **AUC on test data:**

```
AUC: 0.874

[<matplotlib.lines.Line2D at 0x4256de2f98>]
```



5.2.6 Artificial Neural Network:

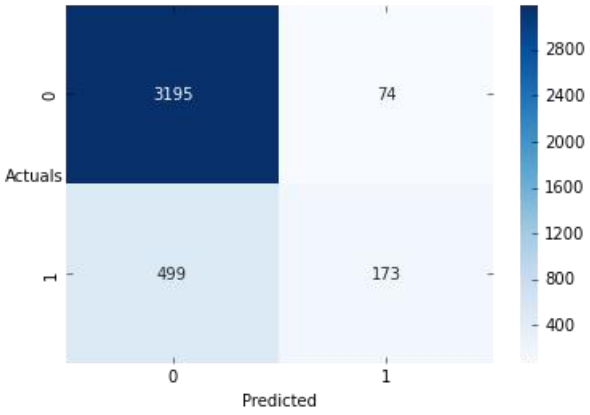
1. **Accuracy of train data:**

   0.9373

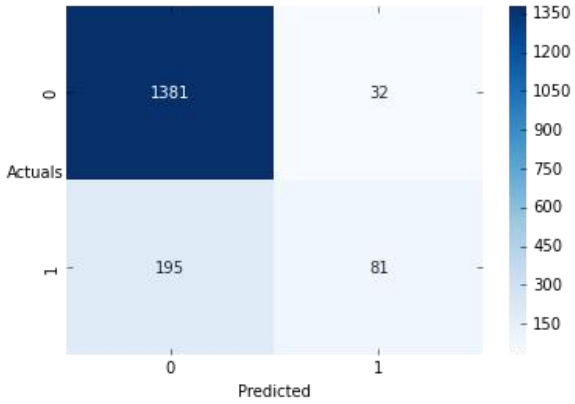
2. **Accuracy of test data:**

   0.9153

3. **Confusion matrix on train data:**



4. **Confusion matrix on test data:**



5. **Classification report on train data:**

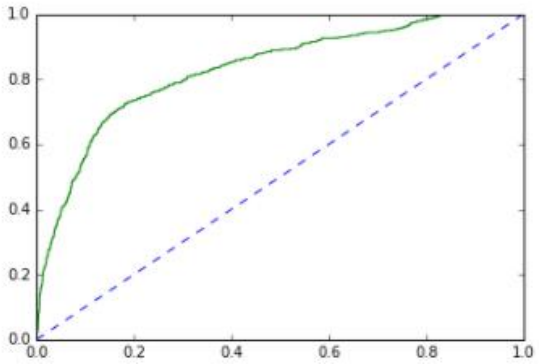
Dependent variable	Precision	Recall	F1-score	Support
0	0.86	0.98	0.92	3269
1	0.70	0.26	0.38	672

6. **Classification report on test data:**

Dependent variable	Precision	Recall	F1-score	Support
0	0.88	0.98	0.92	1413
1	0.72	0.29	0.42	276

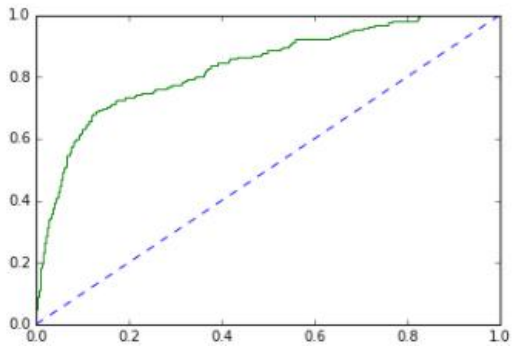
7. **AUC on train data:**

```
AUC: 0.830

[<matplotlib.lines.Line2D at 0x969483d630>]
```



8. **AUC on test data:**

```
AUC: 0.834

[<matplotlib.lines.Line2D at 0x969483d6d8>]
```



5.3 Interpretation of Models:

Once the models are built using training data, the models performance are determined by using testing data to predict the target variable.

1. **Logistic Regression:**

   This turns up with accuracy of 0.7749 on train data and 0.7655 on test data which is a decent performance as there are no underfit or overfit of data. Also the model has good Recall value but the Precision rates and F1 score are very low on Churned customers of both train and test data.

2. **Linear Discriminant Analysis:**

   This model has accuracy of 0.8751 on train data and 0.8815 on test data which is a good performance model. Here, the model has good Precision value but the Recall rates and F1 score are low on Churned customers of both train and test data.

3. **K-Nearest Neighbors:**

   The model's accuracy is 0.9023 on train data and 0.8644 on test data which is overall a good performance model, but the Precision value, Recall value and F1 score are comparatively low on Churned customers of both train and test data.

4. **Naive Bayes:**

   Accuracy is 0.8513 on train data and 0.8703 on test data of the model which is overall a good performance model, but the Precision value, Recall value and F1 score are very low on Churned customers of both train and test data.

5. **Support Vector Machine:**

   This model performs poorly even though the accuracy is 0.8294 on train data and 0.8365 on test data. The Precision value, Recall value and F1 score are zero on Churned customers of both train and test data.

6. **Artificial Neural Network:**

   This model has accuracy of 0.9373 on train data and 0.9153 on test data which is a good performance model. Here, the model has decent Precision value but the Recall rates and F1 score are poor on Churned customers of both train and test data.

6 Model Tuning

However the models which are built can be fine tuned to improve the models performance and can validate the models using some techniques.

6.1 Ensemble Modeling

Ensemble technique builds many models and combines in order to produce one good model. The following ensemble techniques are used.

6.1.1 Grid Search CV and Random Forest

Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model. This is significant as the performance of the entire model is based on the hyper parameter values specified.

The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

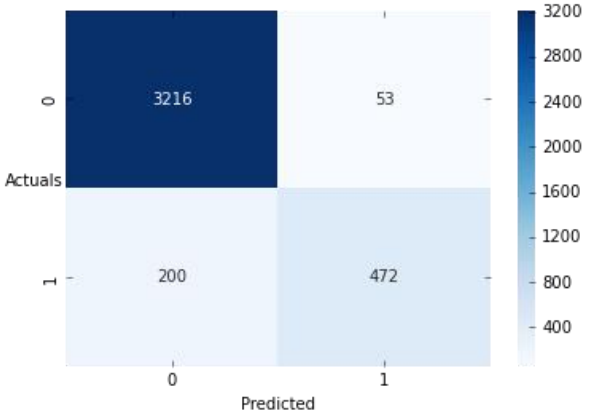
1. **Accuracy of train data:**

   0.9358

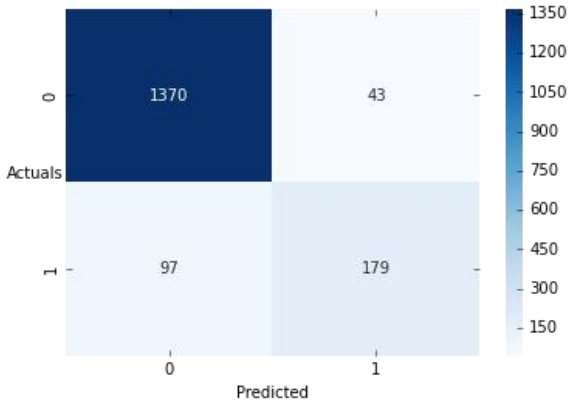
2. **Accuracy of test data:**

   0.9171

3. **Confusion matrix on train data:**



4. **Confusion matrix on test data:**



5. **Classification report on train data:**

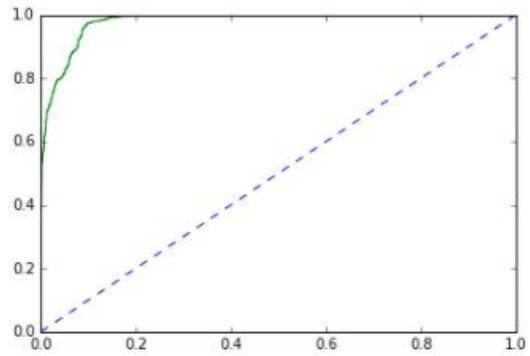
Dependent variable	Precision	Recall	F1-score	Support
0	0.94	0.98	0.96	3269
1	0.90	0.70	0.79	672

6. **Classification report on test data:**

Dependent variable	Precision	Recall	F1-score	Support
0	0.93	0.97	0.95	1413
1	0.81	0.65	0.72	276

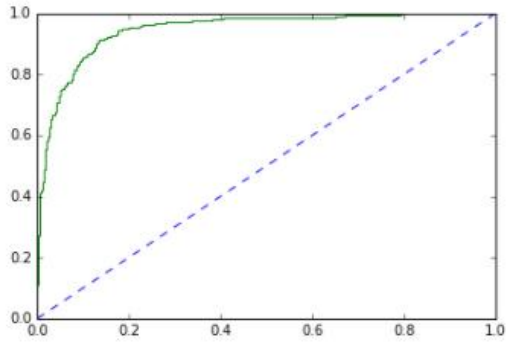
7. **AUC on train data:**

```
AUC: 0.980

[<matplotlib.lines.Line2D at 0xcfeb5f1748>]
```



8. **AUC on test data:**

```
AUC: 0.947

[<matplotlib.lines.Line2D at 0xcfebd97240>]
```



6.1.2 Ada Boost

It aims to convert a set of weak classifiers into a strong one.

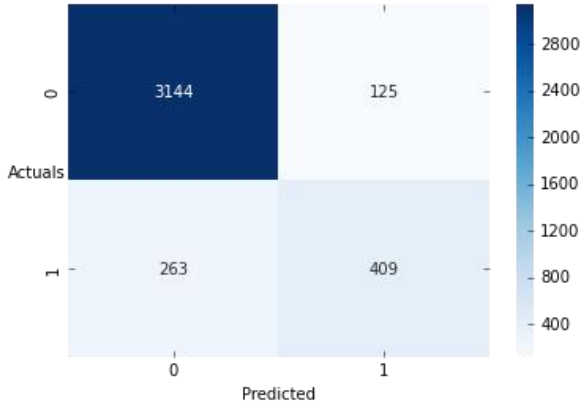
1. **Accuracy of train data:**

   0.9015

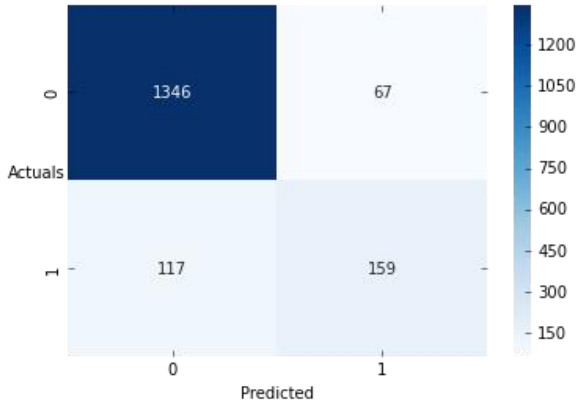
2. **Accuracy of test data:**

   0.8910

3. **Confusion matrix on train data:**



4. **Confusion matrix on test data:**



5. Classification report on train data:

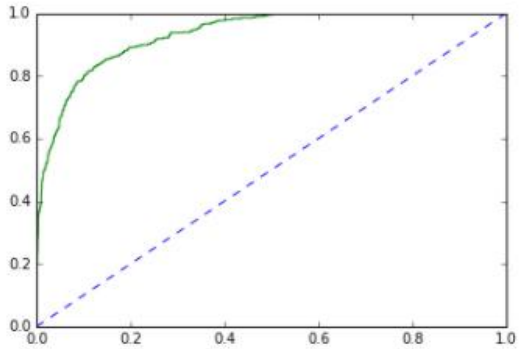
Dependent variable	Precision	Recall	F1-score	Support
0	0.92	0.96	0.94	3269
1	0.77	0.61	0.68	672

6. Classification report on test data:

Dependent variable	Precision	Recall	F1-score	Support
0	0.92	0.95	0.94	1413
1	0.70	0.58	0.63	276

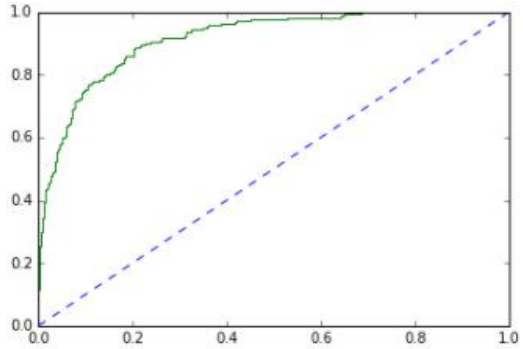
7. AUC on train data:

```
AUC: 0.933

[<matplotlib.lines.Line2D at 0xcfebfac860>]
```



8. AUC on test data:

```
AUC: 0.916

[<matplotlib.lines.Line2D at 0xcfebfc8a90>]
```



6.1.3 XG Boost

XG Boost is an implementation of gradient boosted decision trees designed for speed and performance.

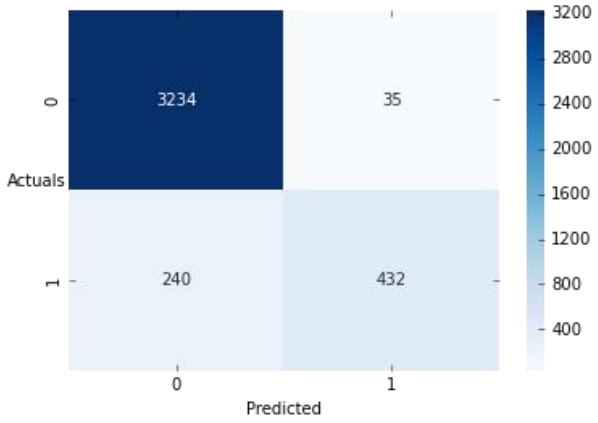
1. **Accuracy of train data:**

   0.9302

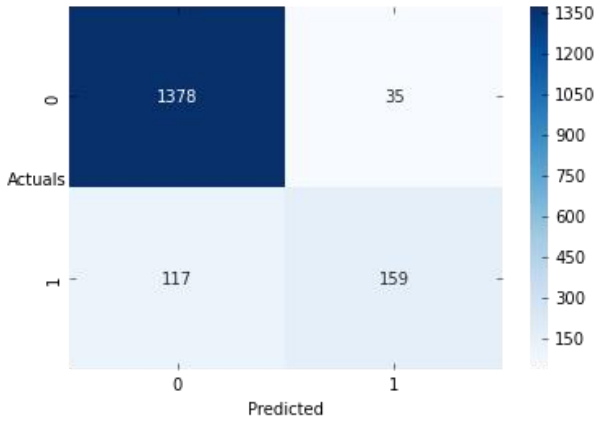
2. **Accuracy of test data:**

   0.9100

3. **Confusion matrix on train data:**



4. **Confusion matrix on test data:**



5. **Classification report on train data:**

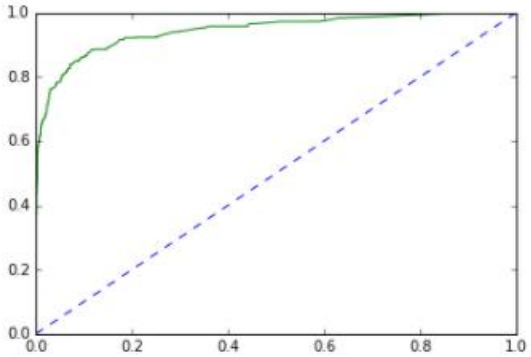
Dependent variable	Precision	Recall	F1-score	Support
0	0.93	0.99	0.96	3269
1	0.93	0.64	0.76	672

6. Classification report on test data:

Dependent variable	Precision	Recall	F1-score	Support
0	0.92	0.98	0.95	1413
1	0.82	0.58	0.68	276

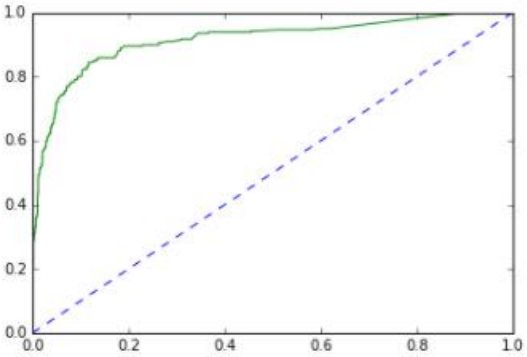
7. AUC on train data:

```
AUC: 0.944

[<matplotlib.lines.Line2D at 0xcfec1a1cf8>]
```



8. AUC on test data:

```
AUC: 0.915

[<matplotlib.lines.Line2D at 0xcfec1a1a58>]
```



6.2 Model Tuning Measures

Cross validation is an important technique to verify the models built by splitting the data into 10 folds.

6.2.1 Logistic Regression

1. Train data:

0.7645	0.7538	0.7741	0.8071	0.7639
0.7157	0.7741	0.7487	0.8147	0.7944

2. **Test data:**

0.8047	0.7514	0.7692	0.7869	0.8284
0.8106	0.7633	0.7455	0.8106	0.7619

6.2.2 Linear Discriminant Analysis

1. **Train data:**

0.8506	0.8883	0.8578	0.8781	0.8502
0.8578	0.8680	0.8857	0.8857	0.8832

2. **Test data:**

0.8698	0.8698	0.8816	0.8934	0.8757
0.8934	0.8698	0.8698	0.8639	0.8809

6.2.3 K-Nearest Neighbour

1. **Train data:**

0.8683	0.8629	0.8426	0.8604	0.8502
0.8527	0.8629	0.8730	0.8629	0.8705

2. **Test data:**

0.8875	0.8461	0.8639	0.8224	0.8639
0.8875	0.8402	0.7988	0.8284	0.8392

6.2.4 Naive Bayes

1. **Train data:**

0.8075	0.8502	0.8502	0.8730	0.8477
0.8147	0.8553	0.8477	0.8883	0.8654

2. **Test data:**

0.8461	0.8461	0.8639	0.8698	0.8934
0.8520	0.8343	0.8402	0.8698	0.8392

6.2.5 Support Machine Vector

1. **Train data:**

0.8278	0.8299	0.8299	0.8299	0.8299
0.8299	0.8299	0.8299	0.8299	0.8274

2. **Test data:**

0.8402	0.8402	0.8402	0.8343	0.8343
0.8343	0.8343	0.8343	0.8343	0.8392

6.2.6 Artificial Neural Network

1. **Train data:**

0.8531	0.8527	0.8477	0.8553	0.8401
0.8274	0.8934	0.8680	0.8502	0.8527

2. **Test data:**

0.8757	0.8698	0.8520	0.8461	0.8402
0.8343	0.8343	0.8402	0.8402	0.8511

6.2.7 Random Forest

1. **Train data:**

0.8962	0.9111	0.9111	0.9162	0.8908
0.8883	0.9213	0.9137	0.9162	0.9035

2. **Test data:**

0.9112	0.8875	0.8875	0.8579	0.9289
0.9171	0.8934	0.8816	0.9053	0.9226

6.2.8 Ada Boost

1. **Train data:**

0.8860	0.8984	0.8934	0.9010	0.8832

0.8857	0.9035	0.9137	0.9035	0.8908

2. Test data:

0.8934	0.9053	0.8875	0.8875	0.8934
0.8994	0.8875	0.8875	0.9171	0.8869

6.2.9 XG Boost

1. Train data:

0.9012	0.9035	0.9060	0.9238	0.8807
0.8705	0.9213	0.9086	0.8934	0.8959

2. Test data:

0.9112	0.9230	0.8816	0.8639	0.9171
0.9171	0.8994	0.8816	0.9112	0.8988

6.3 Interpretation of Ensemble Model

Once the models are built using training data, the models performance are determined by using testing data to predict the target variable.

1. **Grid Search CV and Random Forest:**

   This turns up with accuracy of 0.9358 on train data and 0.9171 on test data which is a good performance as there are no underfit or overfit of data. Also the model has good Precision rates but the Recall value and F1 score are comparatively low on Churned customers of both train and test data.

2. **ADA Boost:**

   This model has accuracy of 0.9015 on train data and 0.8910 on test data which is a good performance model. Here, the model has comparatively low Precision value and the Recall rates and F1 score are low on Churned customers of both train and test data.

3. **XG Boost:**

   The model's accuracy is 0.9302 on train data and 0.9100 on test data which is overall a good performance model, also the model has good Precision rates but the Recall value and F1 score are low on Churned customers of both train and test data.

6.4 Interpretation of Optimum Model

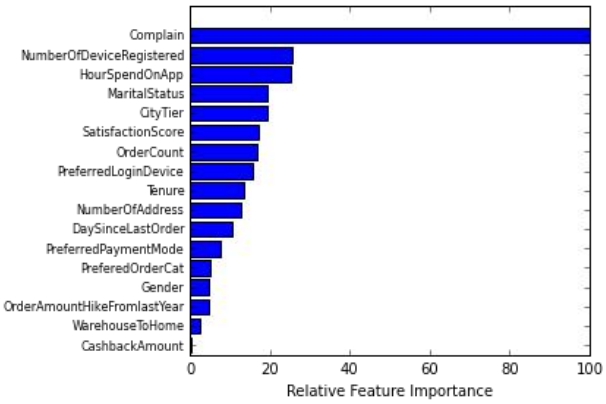
Model	Accuracy		Precision of not Churned		Precision of Churned		Recall of not Churned		Recall of Churned		F1-score of not Churned		F1-score of Churned		AUC Value	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.77	0.76	0.95	0.95	0.42	0.39	0.77	0.76	0.82	0.81	0.85	0.84	0.55	0.53	0.86	0.86
LDA	0.87	0.88	0.89	0.89	0.73	0.76	0.97	0.98	0.42	0.40	0.93	0.93	0.53	0.53	0.85	0.86
KNN	0.90	0.86	0.92	0.90	0.80	0.62	0.97	0.95	0.56	0.43	0.94	0.92	0.66	0.51	0.95	0.85
Naive Bayes	0.85	0.87	0.91	0.93	0.56	0.60	0.91	0.92	0.57	0.62	0.91	0.92	0.57	0.61	0.82	0.84
SVM	0.87	0.88	0.89	0.89	0.75	0.77	0.97	0.98	0.40	0.41	0.93	0.93	0.53	0.53	0.86	0.87
ANN	0.93	0.91	0.86	0.88	0.70	0.72	0.98	0.98	0.26	0.29	0.92	0.92	0.38	0.42	0.83	0.83
Random Forest	0.93	0.91	0.94	0.93	0.90	0.81	0.98	0.97	0.70	0.65	0.96	0.95	0.79	0.72	0.98	0.94
Ada Boost	0.90	0.89	0.92	0.92	0.77	0.70	0.96	0.95	0.61	0.58	0.94	0.94	0.68	0.63	0.93	0.91
XG Boost	0.93	0.91	0.93	0.92	0.93	0.82	0.99	0.98	0.64	0.58	0.96	0.95	0.76	0.68	0.94	0.91

Note: The Out of Bag score of Random forest is 0.9063

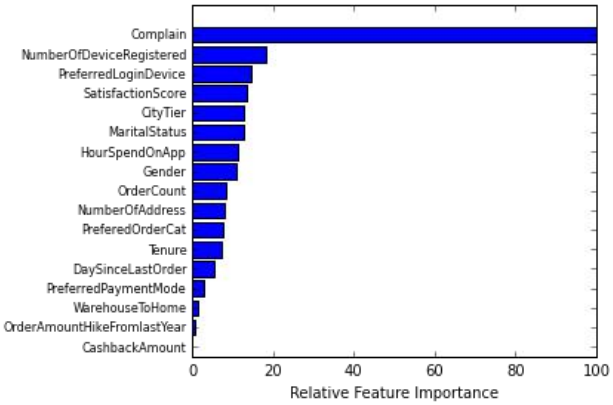
From the above table the models are compared and interpreted as follows:

- Random Forest using Grid Search CV turned to be a best model as they have best accuracy, precision, recall and f1- score on both train and test data.
- Generally Ensemble techniques performs better than other classification techniques.
- Apart from Ensemble techniques, SVM and LDA performs good but with low recall.
- After Random Forest, XG Boost model performs better.

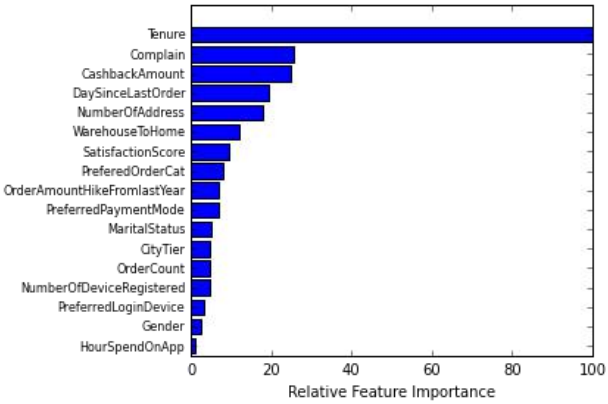
6.5 Implication on the Business



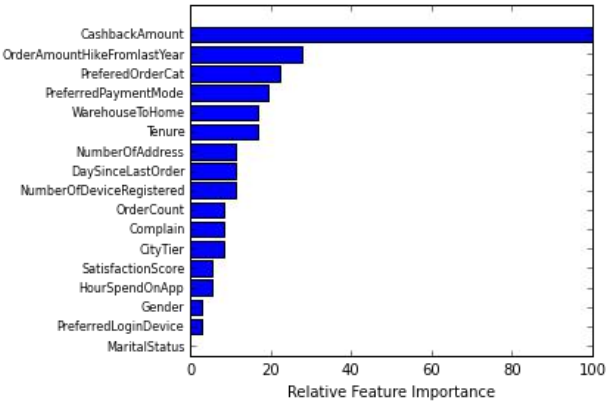
Logistic Regression



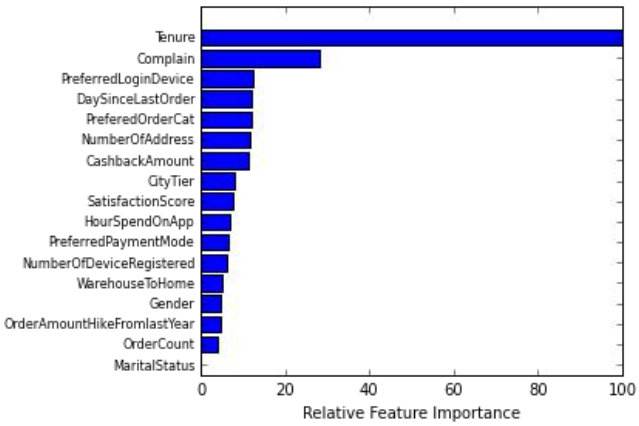
LDA



Random Forest



Ada Boost



XG Boost

- Complain plays the most important role, followed by Number of Device Registered in Logistic Regression and LDA model and Cashback has the least importance.
- Tenure plays the most important role, followed by Complain in Random Forest and XG Boost.

6.5.1 Overall Observations

- Tenure and Complain are important factors for Churn where Order Amount Hike from Last year and Gender seems to be least important factor which plays in model building.
- Other variables like City Tier, Days since Last Order, Satisfaction Score are also some of the important factors which plays in model building.
- Random Forest using Grid Search CV model performs best among all other models so it is good to implement but also it takes much time to build, so if time is a concern then the next best is XG Boost model to implement.
- Company should be conscious on above important factors to reduce Churn rate and the model built predicts the Churn rate of customers with respect to given features.

7 Appendix

- Tools used: Python, Tableau, Knime
- Python code file, Tableau public link and Knime file are attached here for reference

Final Project_Karthihesv

https://public.tableau.com/profile/karthiheswar#!/vizhome/Projectnote-1/ChurnvsCitytier

Project Note-1.knwf