Retailkart.com Customer Segmentation and Churn Case Study

DSC 55 Batch

Submission on 23-03-2024

Schema

Business Understanding and Objective

Customer Segmentation

- Approach
- Customer Segments
- Supportings
- Recommendation

Customer Churn

- Approach
- Logistic Regression Model
- Random Forest Model
- Supportings
- Model Comparison and Selection
- Recommendations

Business Understanding and Objective

Business Understanding :

- Retailkart.com is a small and medium-scale organization that deals in wine, fruit and meat products, leading the offline domain with 35% market share. Due to increased competition, they decided to move online but is facing many challenges to stay competitive in the market.

Business Challenge :

- It has constraints on marketing spend but has to increase revenue
- It has to retain the market share. But churn rate is at 17%, churned customer is as good as new customer requiring more marketing spend .

Business Objective :

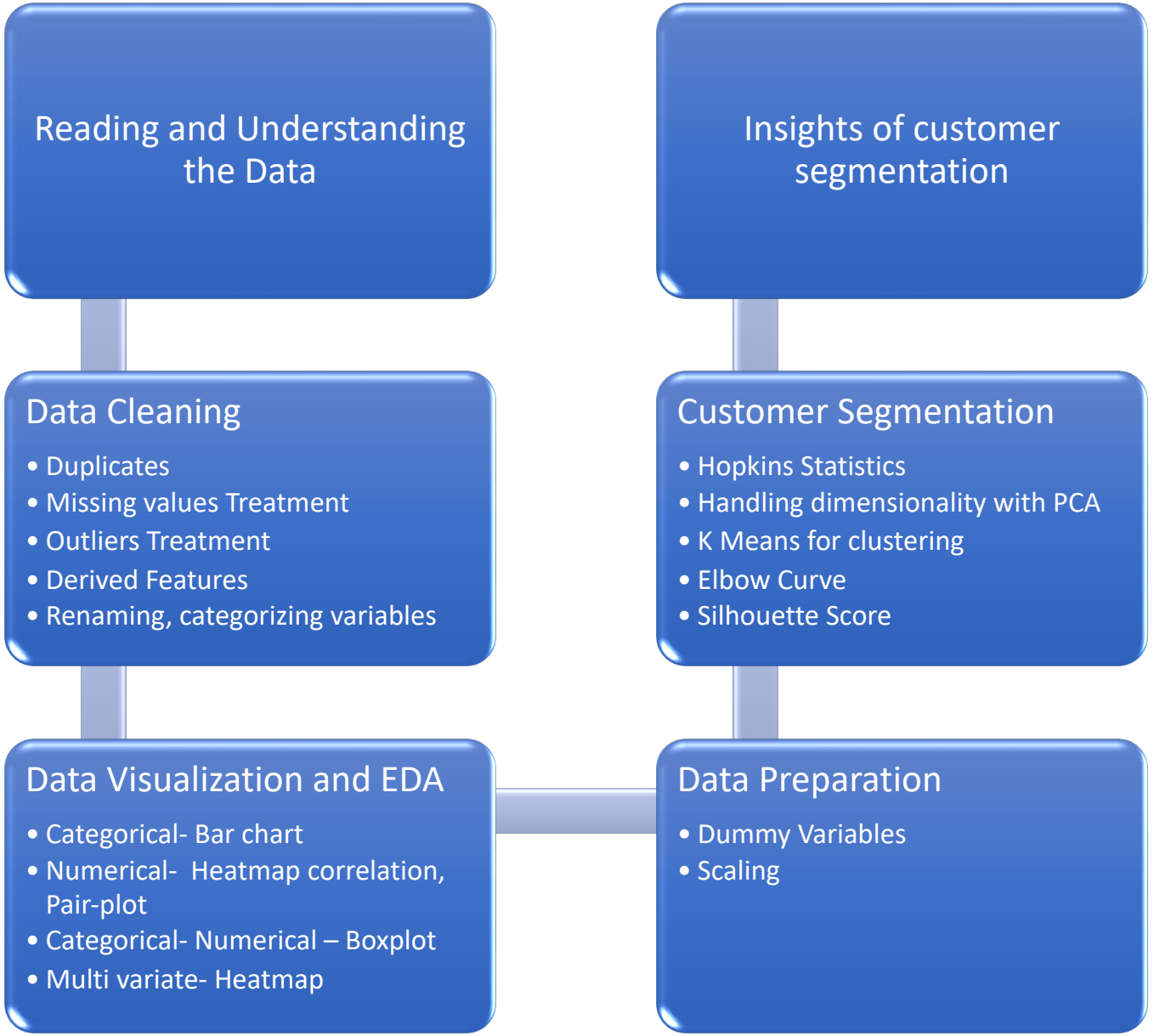
- Improve conversions :When offered personalized experiences based on their needs, customers tend to have a 1% better conversion as compared to offering a regular experience.
- Reduce Churn :When offered a token of appreciation (TOA) of ₹200, customers are likely to have a better repeat rate, lower churn by around 5%.

Data Analysis Objective :

- Customer Segmentation
- Model to predict Customer Churn

Customer Segmentation

Approach


Reading and Understanding the Data
Data Cleaning
• Duplicates
• Missing values Treatment
• Outliers Treatment
• Derived Features
• Renaming, categorizing variables
Data Visualization and EDA
• Categorical- Bar chart
• Numerical- Heatmap correlation, Pair-plot
• Categorical- Numerical – Boxplot
• Multi variate- Heatmap
Data Preparation
• Dummy Variables
• Scaling
Customer Segmentation
• Hopkins Statistics
• Handling dimensionality with PCA
• K Means for clustering
• Elbow Curve
• Silhouette Score
Insights of customer segmentation

Customer Segments

Cluster 0 : Average Value Customers

- Customers with moderate income
- Moderate spend on Wines, Meat, Fish, Fruits, Snacks and Sweets
- Purchase more when there are deals
- 50% customers have made 5-10 store purchases
- Prefer Web Purchases
- Median Age of this cluster is 60

Cluster 1 : High Value Customers

- Customers with high income
- High spend on Wines, Meat, Fish, Fruits, Snacks and Sweets
- 50% customers have made 6-11 store purchases
- Prefer Catalog Purchases
- Median Age of this cluster is 55
- Respond well to campaigns

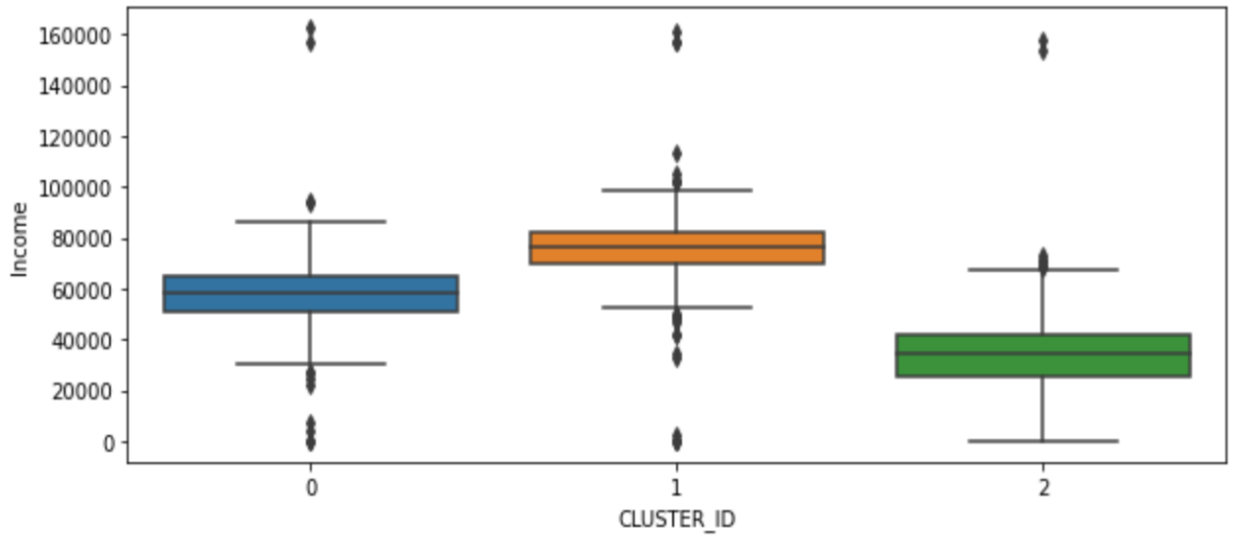
Cluster 2 : Budget Customers

- Customers with low income
- Low spend on Wines, Meat, Fish, Fruits, Snacks and Sweets
- Opt for deals on Purchases
- 50% customers have made 3-4 store purchases
- Frequent Web Visits
- Median Age of this cluster is 50
- Have 1 kid at home
- Includes customers with Basic Education

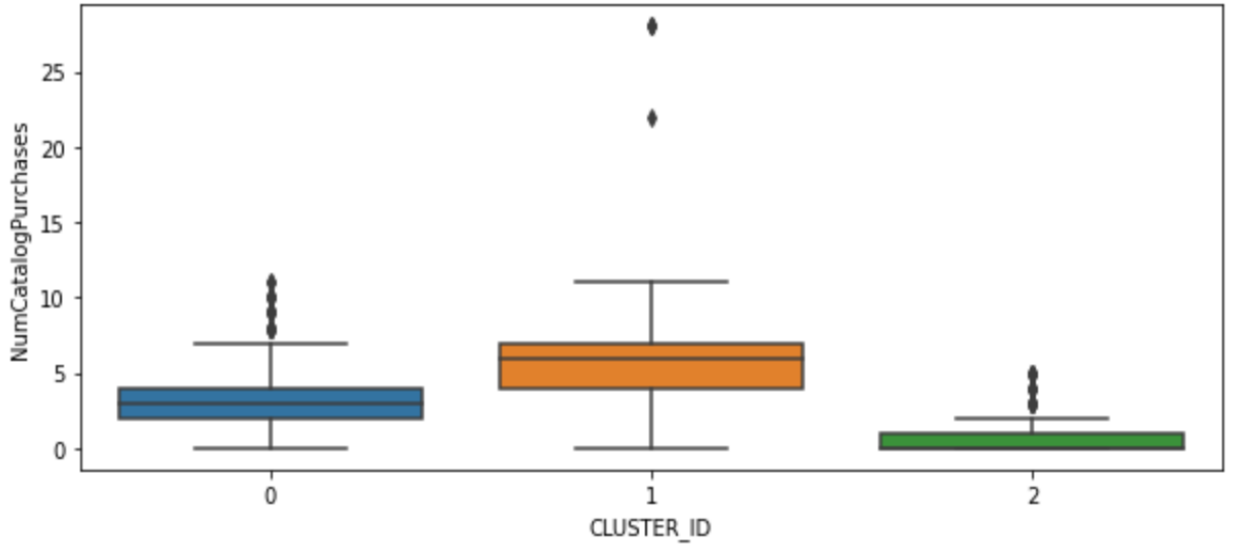
- Silhouette Score for K Means cluster 0.49
- Number of clusters were decided using Elbow Curve and Silhouette Score
- Dimensionality handled through PCA

Insights Customer Segments

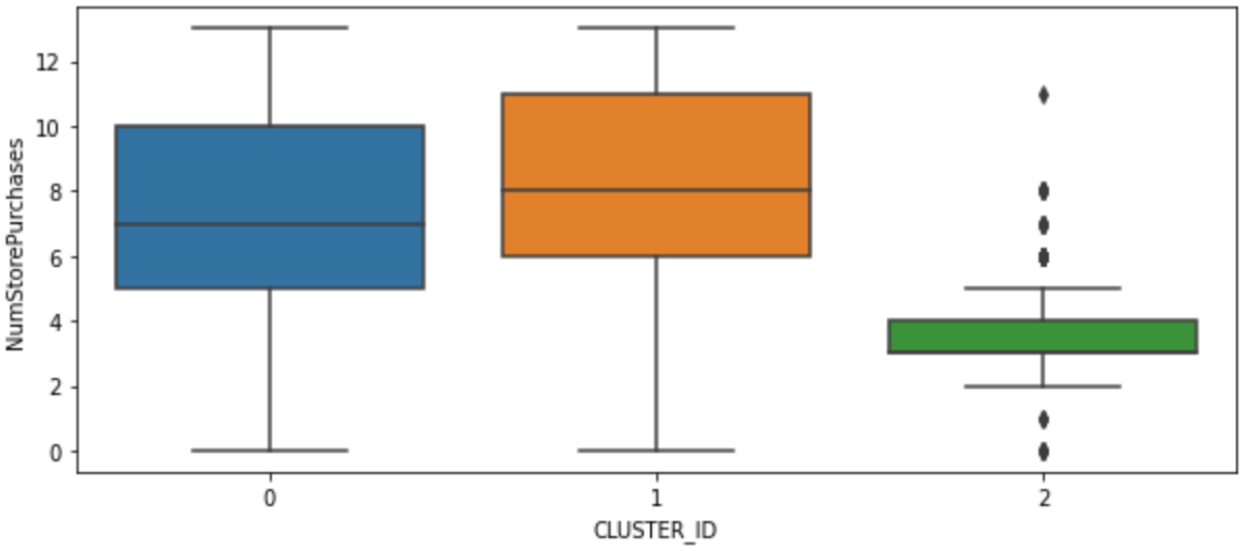
Boxplot Income Vs Cluster ID



Boxplot NumCatalogPurchases Vs Cluster ID



Boxplot NumStorePurchases Vs Cluster ID



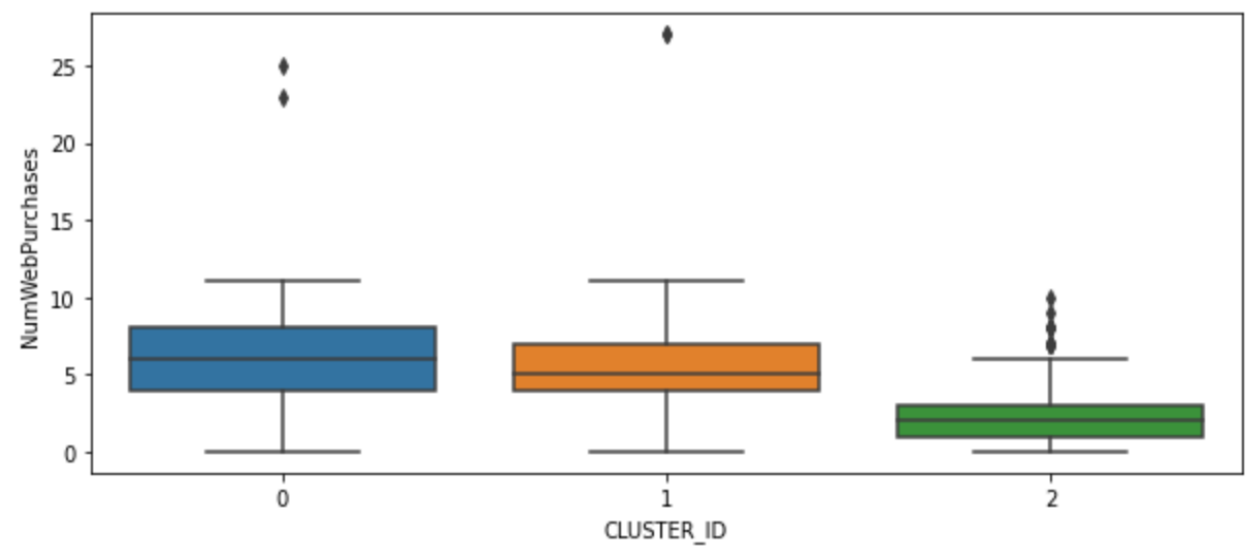
Cluster 0 : Income in the range of 30-85K,preference for Store Purchase

Cluster 1 : Income more than 50K,preference for Store and Catalog Purchase

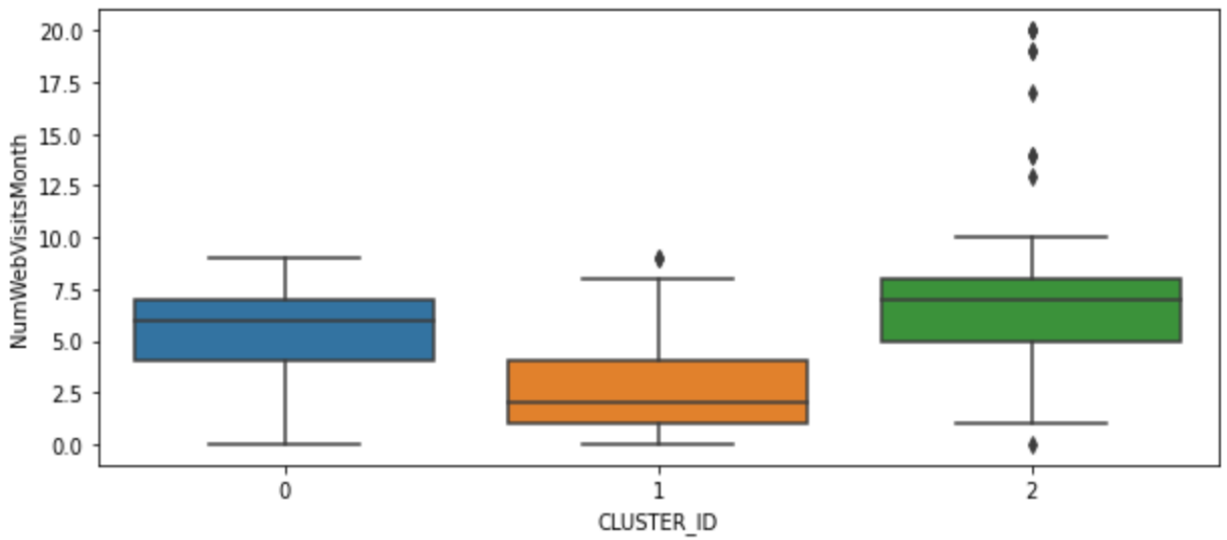
Cluster 2: Maximum Income of customers in this cluster is 60K

Insights Customer Segments

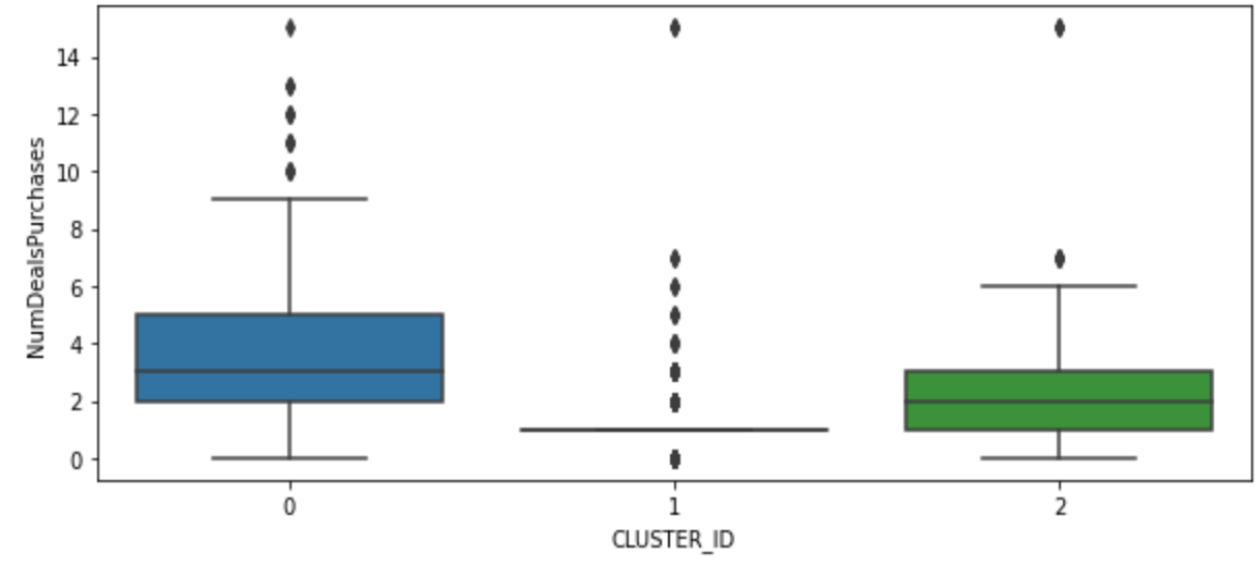
Boxplot NumWebPurchases Vs Cluster ID



Boxplot NumWebVisitsMonth Vs Cluster ID



Boxplot NumDealsPurchases Vs Cluster ID

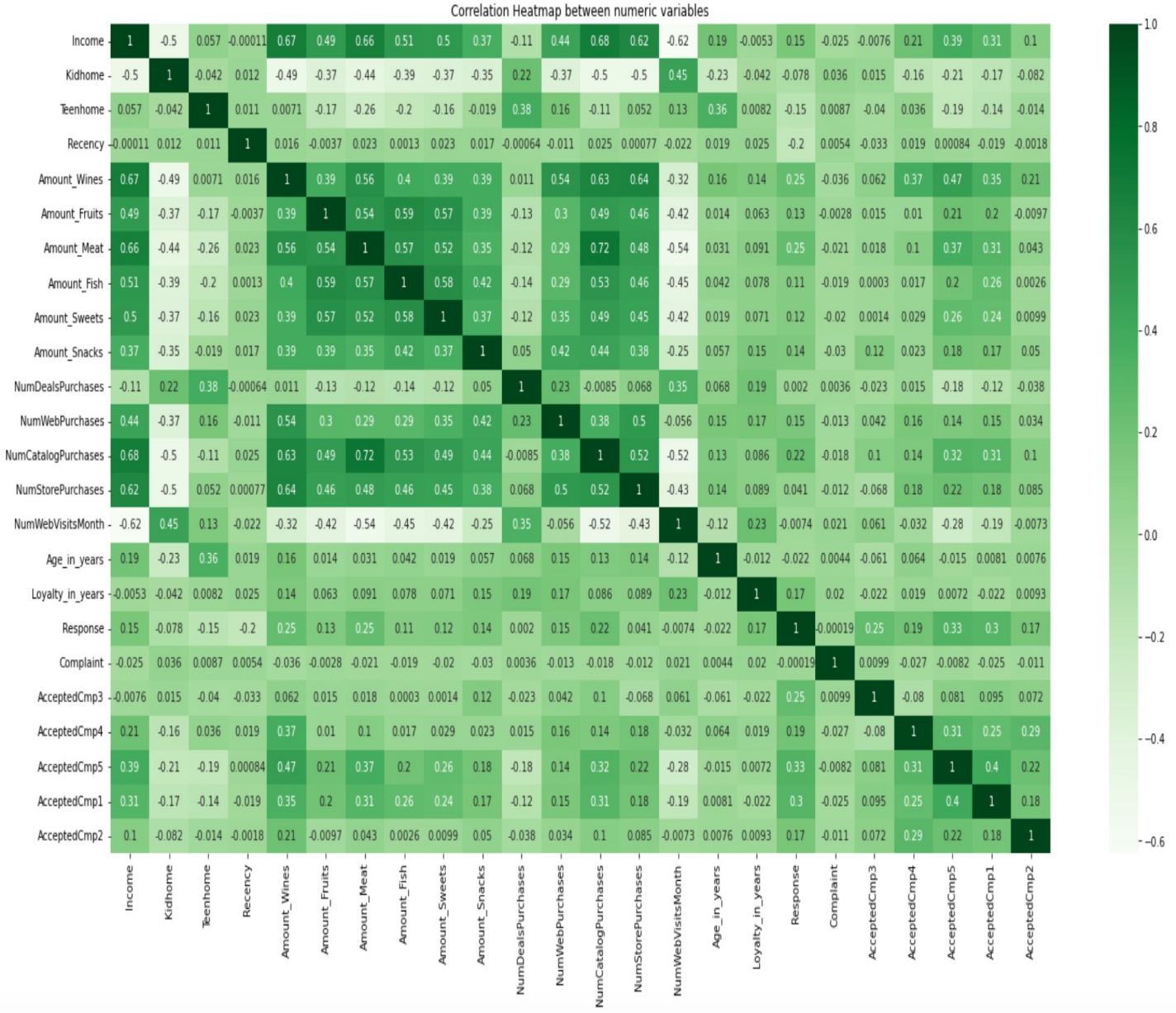


Cluster 0 : More Purchases when there are deals, prefer Web Purchases

Cluster 1 : Do not wait for deals , fewer Web Visits

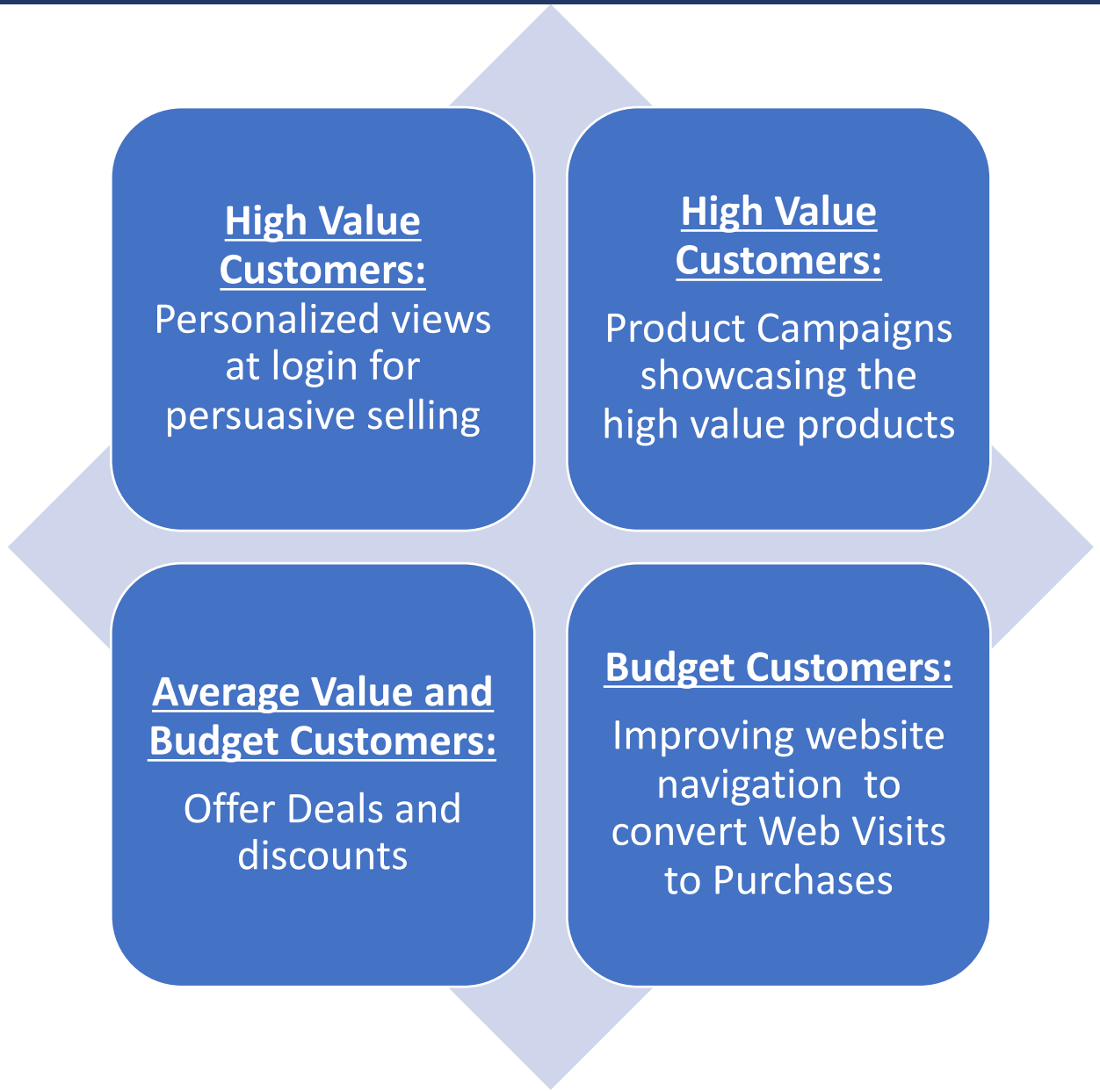
Cluster 2: Customers with at-least 1 Web visit and opt for deals on purchases

Insights Customer Segments



- Income is positively correlated to spend on Wines, Meat, Fish, Fruits, Snacks and Sweets, higher the income, higher the spend.
- Customers with kids at home make more web visits
- Customers prefer Store for purchase of wines and catalog purchases for meat.
- Customers spending on Wines also spend more on meat.
- Customers who spend more on Fruits also spend more on Fish and Sweets

Recommendations



Customer Churn

Approach

Reading and Understanding the Data

Data Cleaning
- Duplicates
- Missing values Treatment
- Outliers Treatment
- Renaming, categorizing variables

Data Visualization and EDA
- Categorical- Bar chart, Pie Chart
- Numerical- Heatmap correlation, Pair-plot
- Categorical- Numerical – Boxplot

Data preparation
- Conversion of categorical to dummy variable
- Dividing Data into X and y
- Train Test Split
- Scaling the data
- Handling Data Imbalance using SMOTE after evaluating ADASYN, Tomek, SMOTE+TOMEK

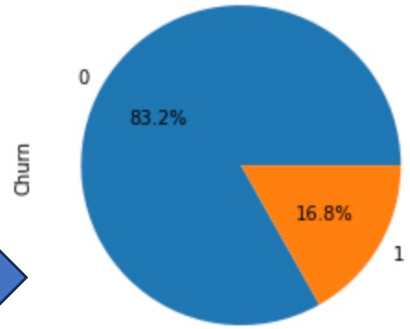
Logistic Regression Model
- Feature Selection RFECV
- Manual tuning- p value<0.05 VIF<5
- Prediction & Evaluation on training data- accuracy, precision, recall AUC, cross val score

Random Forest Model
- Hyper Parameter Tuning with GridSearchCV
- Prediction & Evaluation metrics of best model- accuracy, AUC, precision, recall, cross val score, oob

Prediction and Evaluation on test set
- Model Comparison logistic regression vs random forest
- Selecting the best model based on performance on test set based on F1 Score

Recommendations

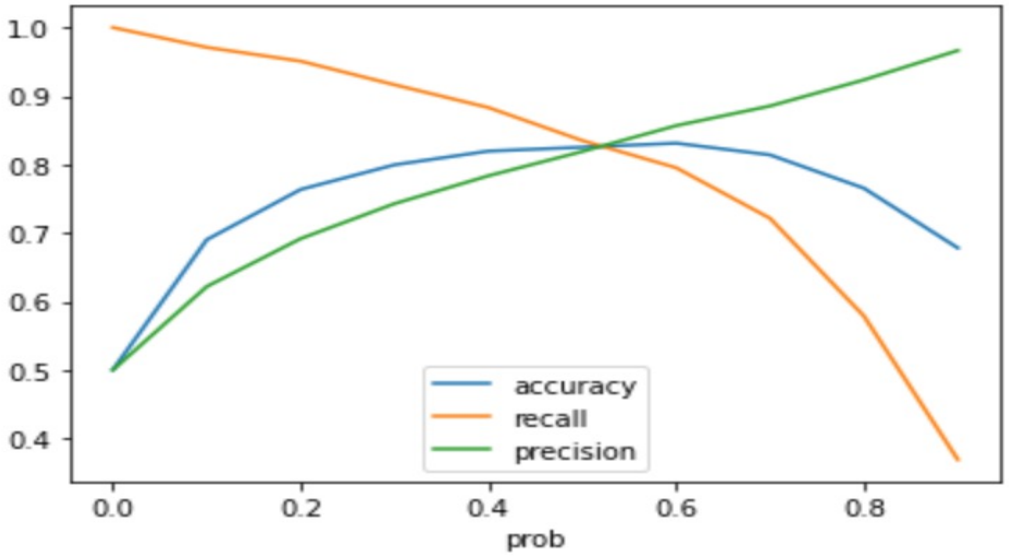
Pie Chart showing the split of the target variable-Churn



Logistic Regression Model

Variables considered in logistic regression model:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3108	0.050	-26.410	0.000	-1.408	-1.213
Tenure	-1.7175	0.057	-29.986	0.000	-1.830	-1.605
WarehouseToHome	0.3917	0.038	10.180	0.000	0.316	0.467
NumberOfDeviceRegistered	0.3389	0.039	8.587	0.000	0.262	0.416
SatisfactionScore	0.4098	0.038	10.837	0.000	0.336	0.484
NumberOfAddress	0.5994	0.040	15.126	0.000	0.522	0.677
Complaint	0.7166	0.035	20.546	0.000	0.648	0.785
OrderAmountHikeFromlastYear	-0.1515	0.037	-4.066	0.000	-0.224	-0.078
OrderCount	0.4732	0.047	10.039	0.000	0.381	0.566
DaySinceLastOrder	-0.4890	0.048	-10.112	0.000	-0.584	-0.394
CashbackAmount	-0.4216	0.067	-6.260	0.000	-0.554	-0.290
PreferredLoginDevice_Mobile Phone	-0.1883	0.044	-4.301	0.000	-0.274	-0.103
PreferredLoginDevice_Phone	-0.2315	0.042	-5.498	0.000	-0.314	-0.149
PreferredPaymentMode_Credit Card	-0.2886	0.065	-4.421	0.000	-0.417	-0.161
PreferredPaymentMode_Debit Card	-0.2170	0.068	-3.205	0.001	-0.350	-0.084
PreferredPaymentMode_E wallet	-0.1362	0.060	-2.287	0.022	-0.253	-0.019
Gender_Male	0.2003	0.037	5.356	0.000	0.127	0.274
PreferedOrderCat_Laptop & Accessory	-0.7139	0.042	-16.992	0.000	-0.796	-0.632
PreferedOrderCat_Others	0.3338	0.059	5.632	0.000	0.218	0.450
MaritalStatus_Single	0.4057	0.035	11.491	0.000	0.336	0.475
CityTier_2	0.1044	0.036	2.931	0.003	0.035	0.174
CityTier_3	0.4757	0.042	11.406	0.000	0.394	0.557



- Logistic Regression Model model6 with 21 features has cross val score of 83% on accuracy and 84% on recall.
- Optimal cut-off point 0.54
- Probability threshold considered as 0.4 as recall (0.88) is high at that level

Random Forest Model

Feature wise importance in random forest:

VarName	Imp
Tenure	0.326835
Complaint	0.134160
DaySinceLastOrder	0.070426
PreferedOrderCat_Mobile Phone	0.063436
NumberOfDeviceRegistered	0.047406
CashbackAmount	0.042640
MaritalStatus_Single	0.039566
SatisfactionScore	0.036830
NumberOfAddress	0.031921
CityTier_3	0.030532
MaritalStatus_Married	0.026613
WarehouseToHome	0.024676
HourSpendOnApp	0.021608
CouponUsed	0.019377
OrderCount	0.019133
PreferredLoginDevice_Mobile Phone	0.016606
PreferedOrderCat_Laptop & Accessory	0.015378
OrderAmountHikeFromlastYear	0.011759
Gender_Male	0.005716
PreferredPaymentMode_Credit Card	0.005308
PreferredPaymentMode_Debit Card	0.003543
PreferredLoginDevice_Phone	0.002775
PreferredPaymentMode_E wallet	0.002423
PreferedOrderCat_Grocery	0.000928
PreferedOrderCat_Others	0.000404

Parameters given for hyper parameter tuning :

```
Fitting 5 folds for each of 256 candidates, totalling 1280 fits

GridSearchCV(cv=5,
             estimator=RandomForestClassifier(n_jobs=-1, oob_score=True,
                                              random_state=42),
             n_jobs=-1,
             param_grid={'max_depth': [5, 10, 15, 20],
                         'max_features': [5, 10, 15, 20],
                         'min_samples_leaf': [50, 100, 250, 350],
                         'n_estimators': [25, 50, 80, 100]},
             verbose=1)
```

Parameters of the best random forest:

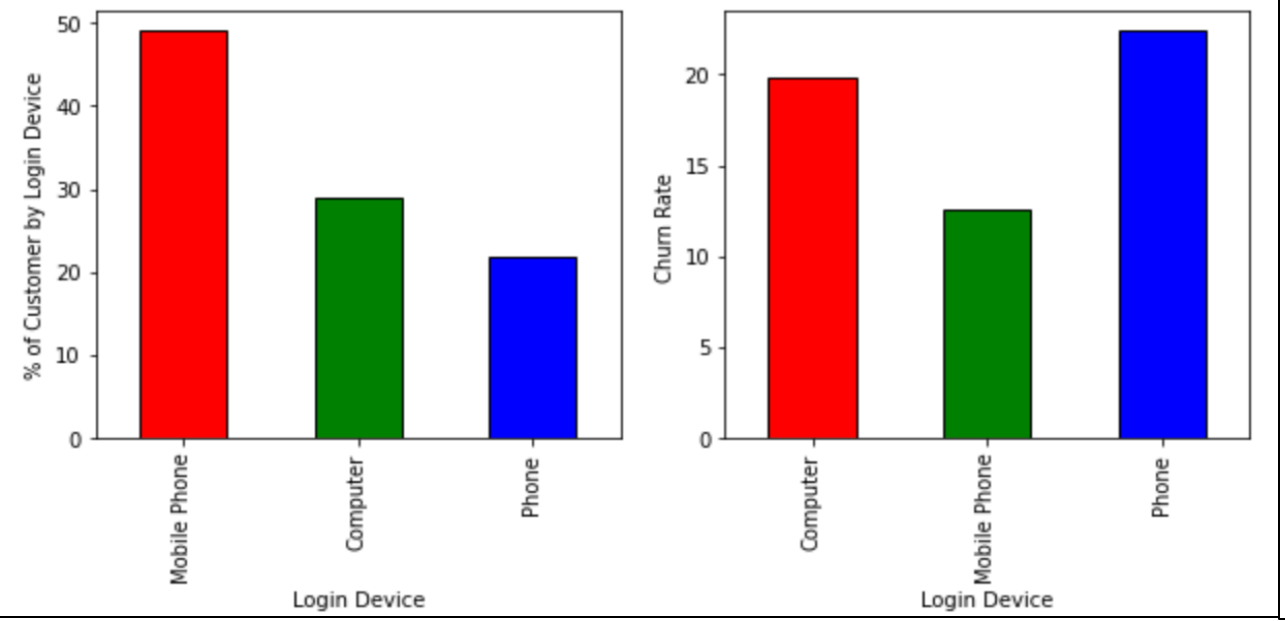
```
# finding the best random forest
rf_best=rf_cv.best_estimator_
rf_best
```

```
RandomForestClassifier(max_depth=10, max_features=5, min_samples_leaf=50,
                       n_jobs=-1, oob_score=True, random_state=42)
```

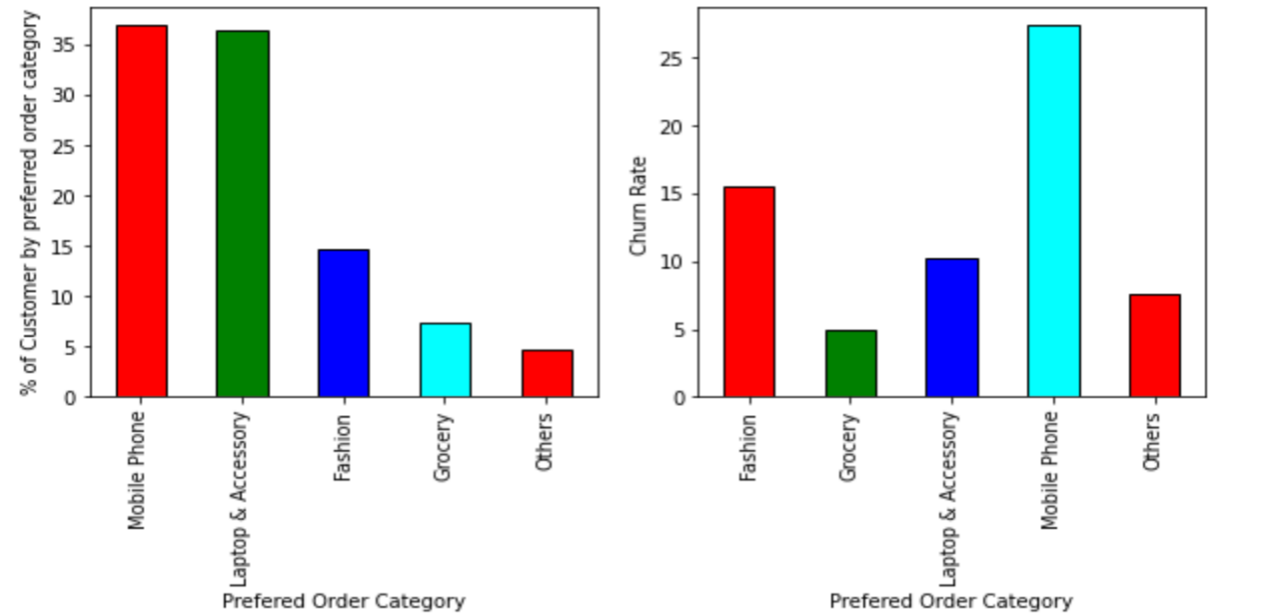
- OOB Score of the Random forest is 88% and cross val score 87.61% . There is no overfitting in this model.

Insights supporting the models

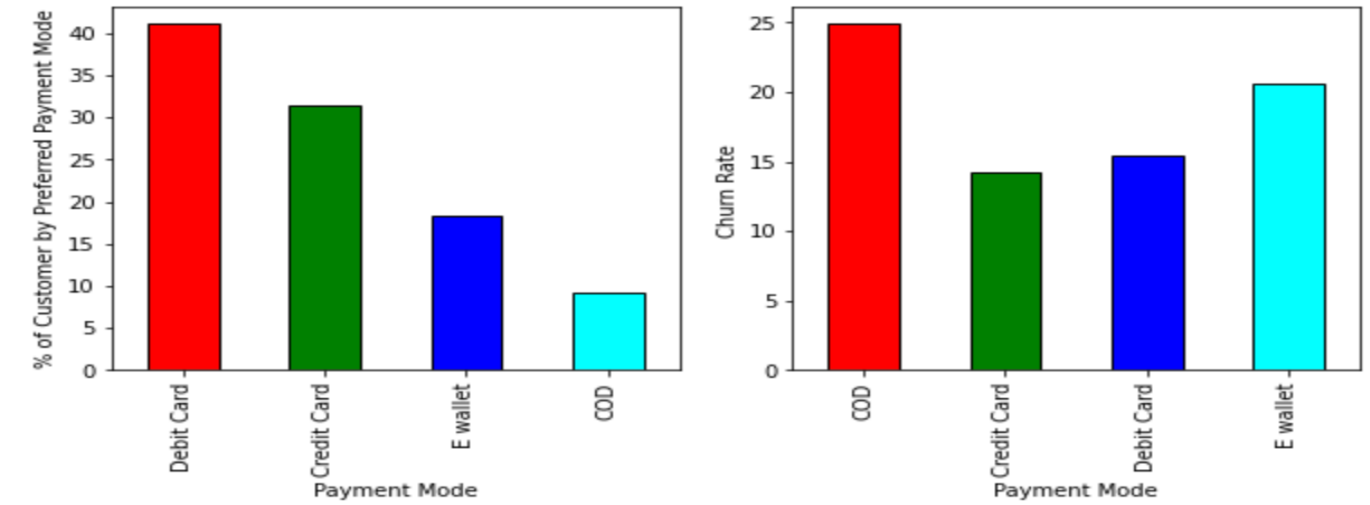
Bar Chart showing the Customers by Preferred Login Device and Churn Rate



Bar Chart showing the Customers by preferred order category and Churn Rate



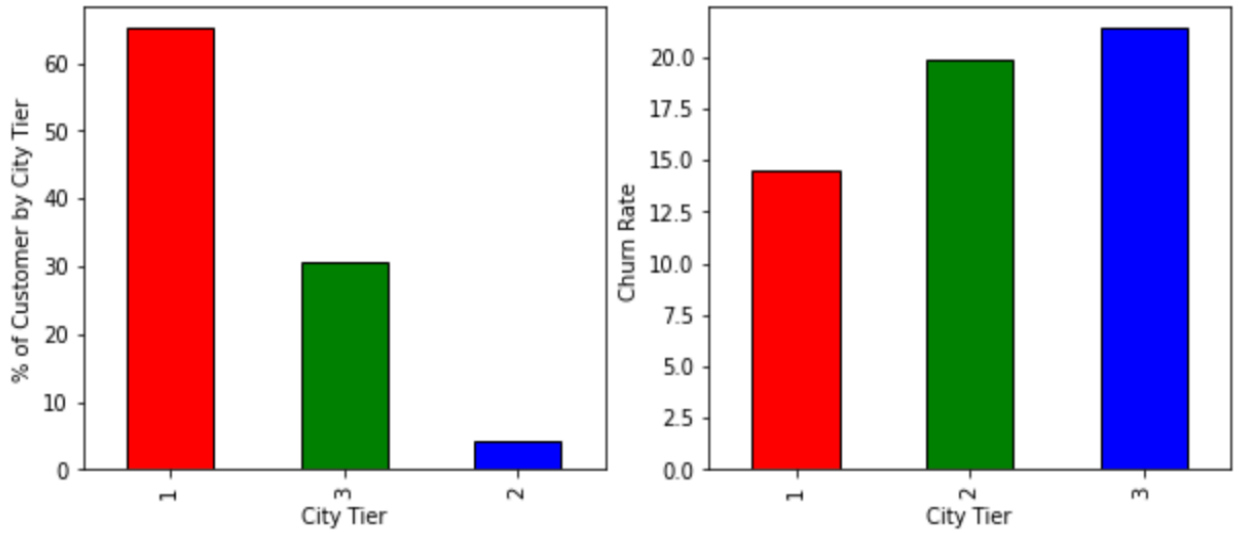
Bar Chart showing the Customers by Preferred Payment Mode and Churn Rate



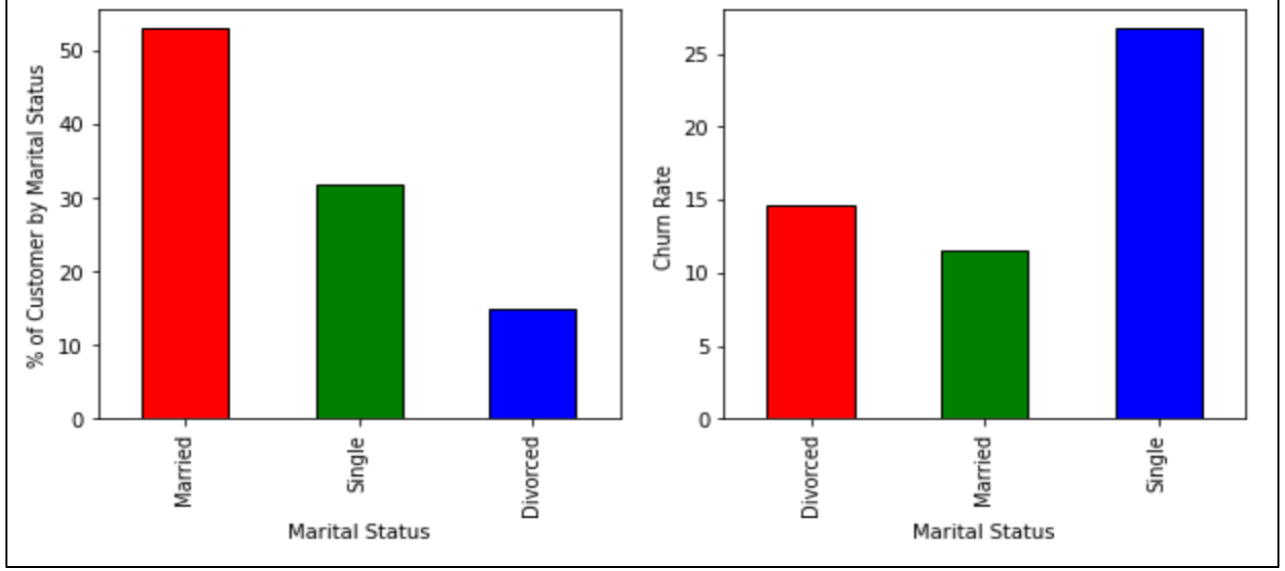
- Almost 50% customers prefer mobile phone to login but their churn rate is low.
- More than 35% customers prefer order category Laptop & Accessory and this category has a low churn rate of around 10%
- 91% customers prefer payment mode other than COD and also have lower churn rates.

Insights supporting the models

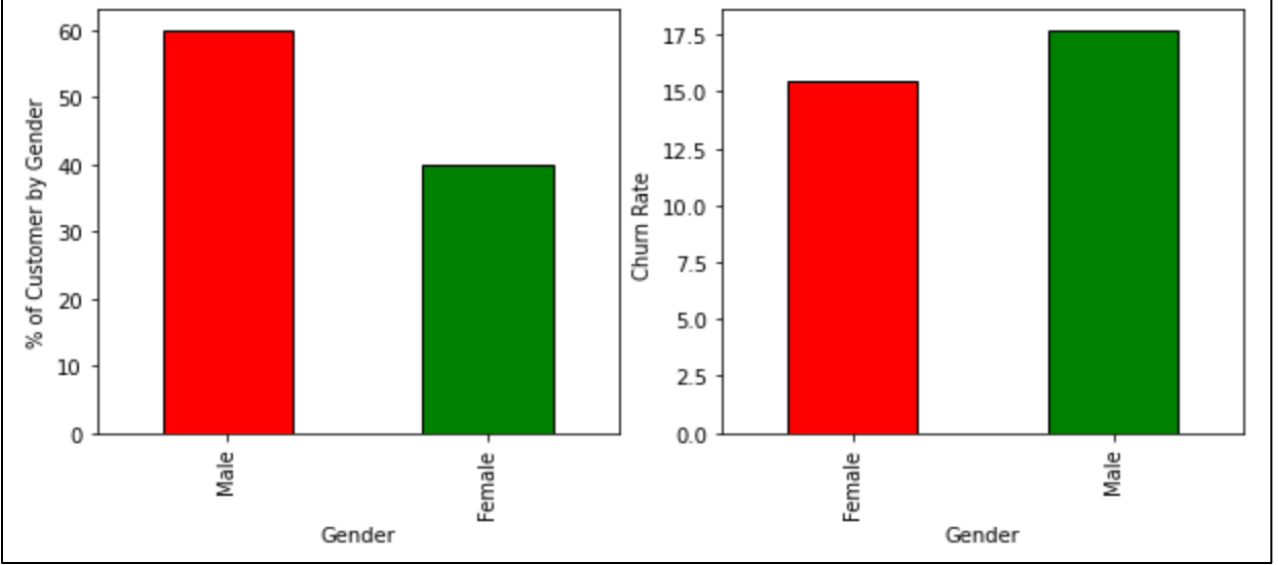
Bar Chart showing the Customers by City Tier and Churn Rate



Bar Chart showing the Customers by Marital Status and Churn Rate



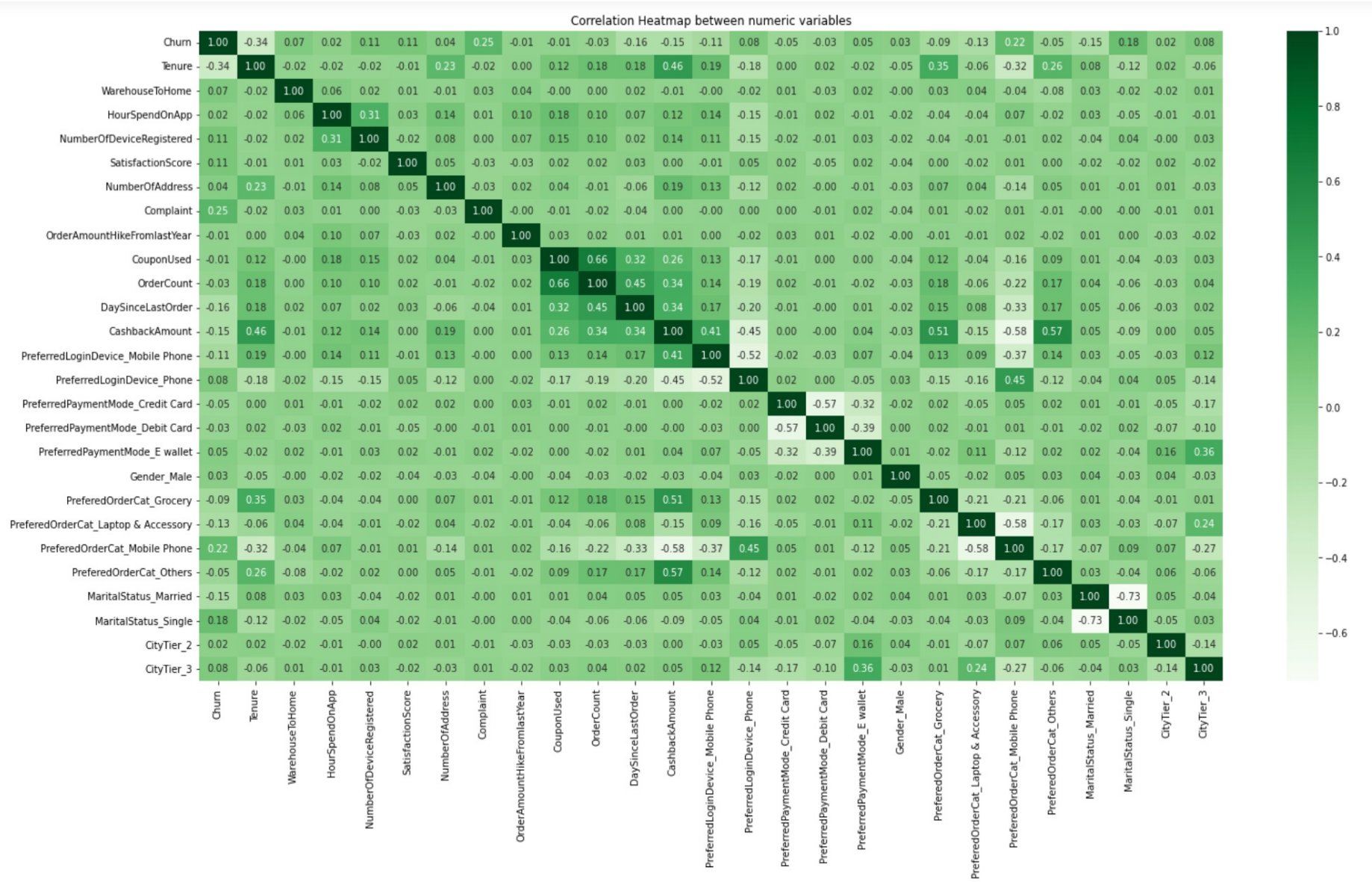
Bar Chart showing the Customers Gender wise and Churn Rate by Gender



Churn Rate is high in

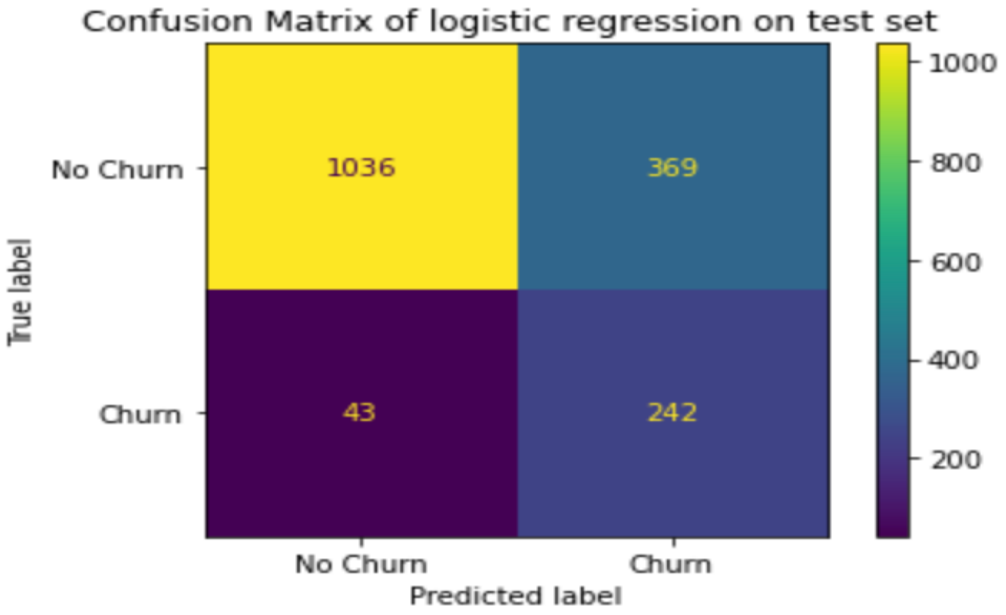
- Tier 3 City
- Customers with marital status-single
- Male customers

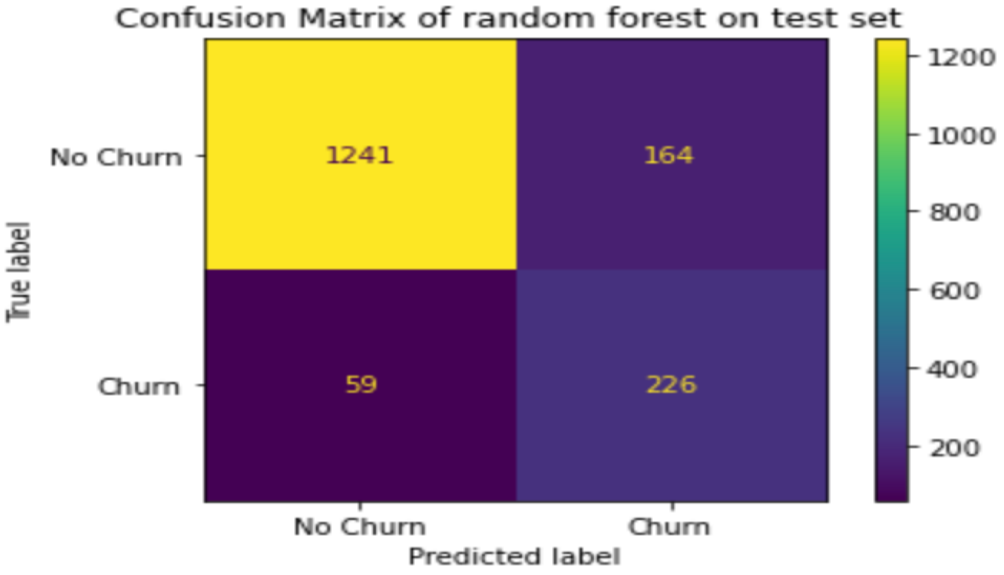
Insights supporting the models



- As the customer tenure increases the Churn Rate decreases
- Increase in Number of Address and registered Device, WarehousetoHome Distance also increases chances of churn

Random Forest vs Logistic Regression

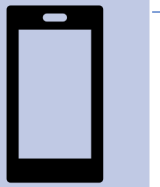




Random Forest is the model to be chosen for deployment

- **F1 Score** is used for comparison as the objective of the model is to reduce False Negative and False Positive
- F1 Score of **Random Forest(0.67)** is better than Logistic Regression(0.54)
- Recall Random Forest - 0.79 ,Logistic Regression - 0.85
- Precision Random Forest- 0.58 Logistic Regression0.4
- Both the models above i.e Random Forest and Logistic Regression have been built ***handling data imbalance, feature selection and hyper parameter tuning.***
- F1 Score on Logistic Regression models handling data imbalance **but** without feature selection are as follows:
  - SMOTE -0.58
  - ADAYSN- 0.56
  - SMOTE+TOMEK – 0.58
  - Tomek -0.62

Recommendations

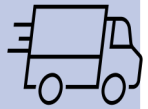


Improve Product Offerings and offer discounts in the Product Category –Mobile Phones.

Loyalty programs to be rolled out for customers with long tenure

Set up team to address Customers Grievances and Complaints

Discourage Cash on Delivery payment mode by customers.



Reduce delivery charges where Warehouse is far from Customers home

City Specific Campaigns in Tier3 City

Thank You