Title: Ecommerce Customers Churn Analysis

Group 1: Club Ecommerce

- Abhishek Kukreja
- Ravina Ingole
- Akash Dhage

Contents:

1.Executive Summary:

This project involves deploying the supervised machine learning algorithm. Supervised machine learning algorithm can be classification or regression. This project is based on building a classifier. As a pre-elementary step, data is preprocessed by treating the null and outlier values. Then this processed data is split into training and test data (80%-20% split). Model is made to learn the true characteristic of the dataset using training data and evaluated using test data to build a generalized model. This project involves using decision tree and ensemble machine learning approach - random forest, since they are robust machine learning algorithms and are known to have better prediction power. We have incorporated several evaluation methodologies - Precision, Recall, Accuracy and F1 score to access the classifier performance in a better way and gain more insights about the classifier's prediction ability. We have also used feature importance to identify the most important predictor variable impacting the target variable.

2.Problem Statement & Research questions:

Success of any companies hugely depends on how well they can analyze the data on their clients' behavior, but Costs of gaining new customers are usually 5-6 times higher than the costs of retaining an existing customer so it's good idea to switch focus from acquiring new customers to retaining existing ones i.e., reducing customer churn. Churn prediction is one of the most critical indicators of a healthy growing business, irrespective of the size or channel of sales.

Research questions:

1. In this project, we will investigate the churn of each customer affected by other input independent variables.
2. We will analyze the variables that affects most to the churn of each customer that will help to predict the churn of each customer.
3. we will use machine learning approach – Decision Tree and Random Forest.

3.Data Preprocessing and Visualization:

Dataset for the project was derived from Kaggle. Dataset originally had 5,630 records and 21 features (20 independent variables + 1 target variable). As a pre-elementary step, dataset was probed to highlight any missing values and outlier values.

3.1. Missing values treatment:

Dataset had 1,856 records with null values (33% of total data). Eliminating all these records will result into huge data loss and less data availability for creating a machine learning model. Histograms were plotted for each feature to identify the data distribution.

Features with skewed distribution (right-tail/left-tail) null values were replaced by median. On the other hand, one with roughly skewed distribution were replaced by mean.

3.2. Outliers' values treatment:

Dataset was also probed to identify the outlier values. As outliers, greatly impact the machine learning prediction since algorithms learn to model the noise by accounting the outliers. Boxplots were plotted for features demonstrating outliers (by comparing the max value and 75 percentile). All the datapoints beyond 99 percentiles were regarded as outliers and removed - summing up to 352 records (making 6.25% of total data).

3.3. Scaling:

Scaling hinders the machine learning algorithm's ability to learn the true characteristics of the data points. Hence, all the datapoints were scaled between 0 and 1 to improve the prediction accuracy.

3.4 Exploratory Data Analysis:

Plotted Heat Map, Bar Plot and Scatter plot to understand relationship and observed that there is strong relation between:

- Tenure vs Churn (More association with organization lesser is the churn).
- Tenure vs Cashback amount
- Coupon used vs Order count
- Order count vs Day since last order.
- Customer ID vs Hour spends on app (Irrelevant)

Observations:

1. Dataset contains 17% of churn users (37% females and 63% males).
2. Churn is similar at gender level (15% vs 18% for females and males respectively).
3. 65% of users are from Tier 1 followed by Tier 2 and Tier 3. Churn rate is similar across all Tiers (~15%).
4. Most of the users spend ~3hrs on app/website.
5. Debit card and credit card is the most often payment mode of delivery followed by UPI, wallet,

COD etc.
6. Churn rate amongst users who have complained is 46% while Churn rate amongst users who have not complained is observed at 12%.
7. Tenure and Churn has negative correlation.
8. Tenure and Cashback amount has strong positive correlation.

4. Methods (Models, Experiments, Analysis) and Results:

4.1 Methods:

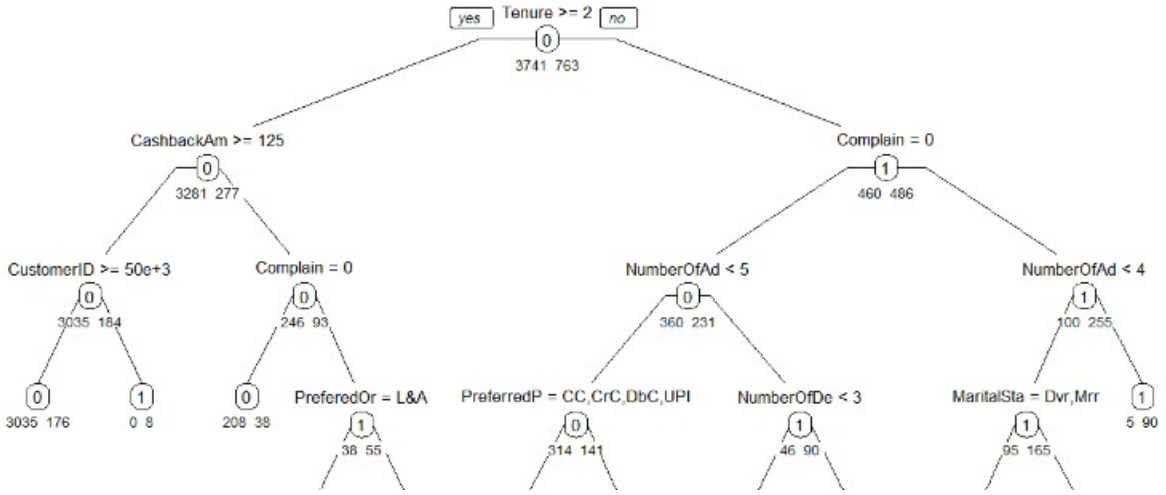
Post pre-processing, dataset was split into training and test data (80%-20% split). Idea was to build model using training data and evaluate the performance on unseen data i.e. test data to build a generalized model. Two machine learning approaches were deployed in this project viz; Decision trees and Random Forest.

4.2 Decision trees:

Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The idea is to divide the entire predictor variables dataset into rectangles such that each rectangle is as homogeneous or “pure” as possible. By pure, we mean belonging to one class/label.

The algorithm examined each predictor variable and all possible split values for each variable to find the best split. Best split is decided basis the reduction in impurity before and after the split (Gini impurity or entropy).

For our project, tenure is the root node and the resulting child nodes are cashback amount and complain.



4.3 Pruned Decision Tree:

Decision Tree has disadvantage of overfitting on the t

raining data and models the noise in the training data. This is addressed by controlling the depth of the tree using complexity parameter. By plotting the complexity parameter, we can find that it has minimum accuracy at 20th row i.e. 0.5409 and the corresponding terminal nodes are 110.

4.4 Random Forest:

Random forest is an ensemble machine learning approach which combines multiple machine learning models for prediction. It encompasses two approaches - bagging (improves predictive power by combining multiple classifiers or predictive algorithms) and bootstrapping (draw random samples with replacement from the data). Apart from instilling randomness at drawing samples from dataset, randomness is brought into effect while selecting a subset of predictors at each stage. Label is identified by voting for classification and averaging for prediction.

For our project, we have used 500 decision trees to build random forest.

4.5 Results:

Accuracy was used to access the three-model performance at a rudimentary level. However, accuracy doesn't consider true positives, true negative, false positivity rate etc. Hence, metrics such as precision, recall and F1 score were also evaluated to better comprehend the performance. Result summary for the three models is shown below:

ML models	Accuracy	Precision	Recall	F1 Score
Decision tree	89%	78%	53%	63%
Pruned decision tree	95%	90%	72%	80%
Random Forest	96%	100%	71%	83%

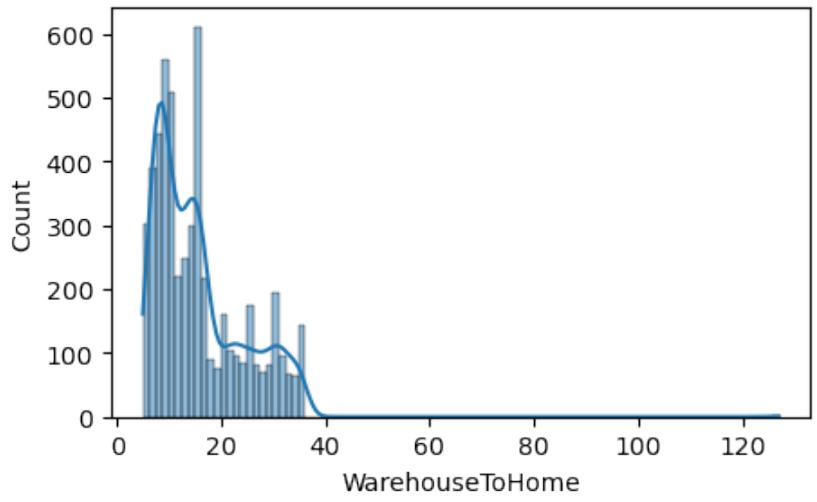
Table 4.1.1

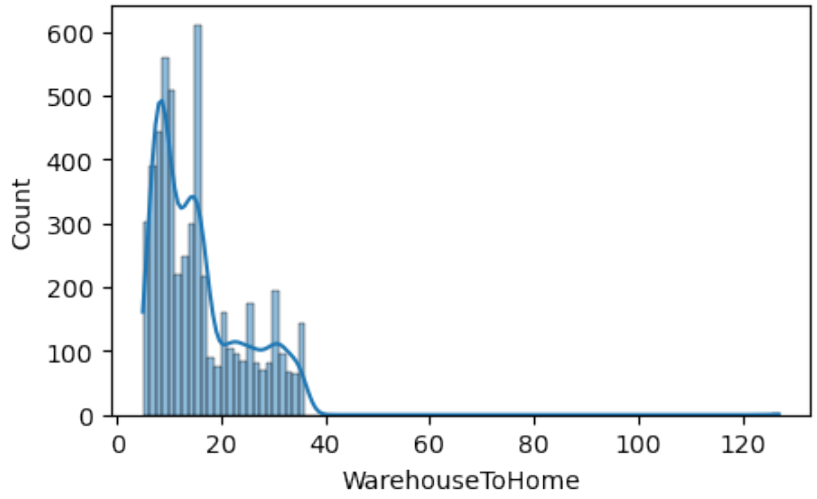
5.Conclusion:

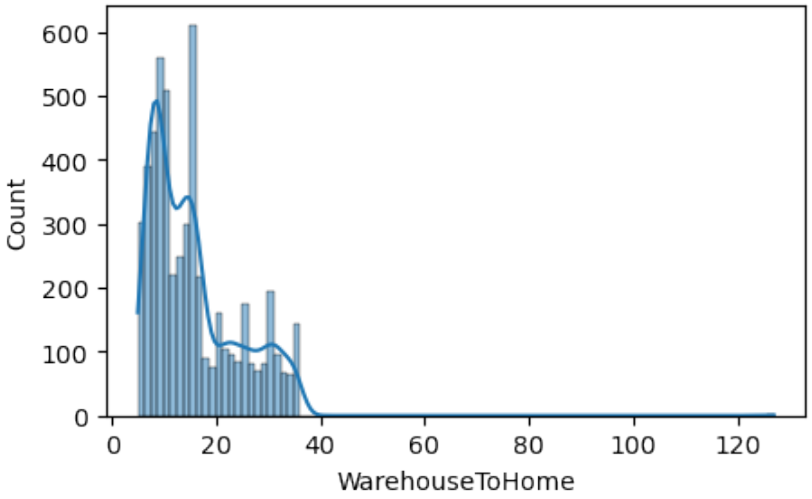
- Model was built using Decision Tree and Random Forest and accuracy for the same is 89% and 96%. Therefore, as expected, random forest performed better than decision tree model.
- Confusion matrix was created across each model and f1 score was evaluated to compare the prediction power of each model.
- F1 score for Decision Tree and Random Forest is 63% and 83% result.

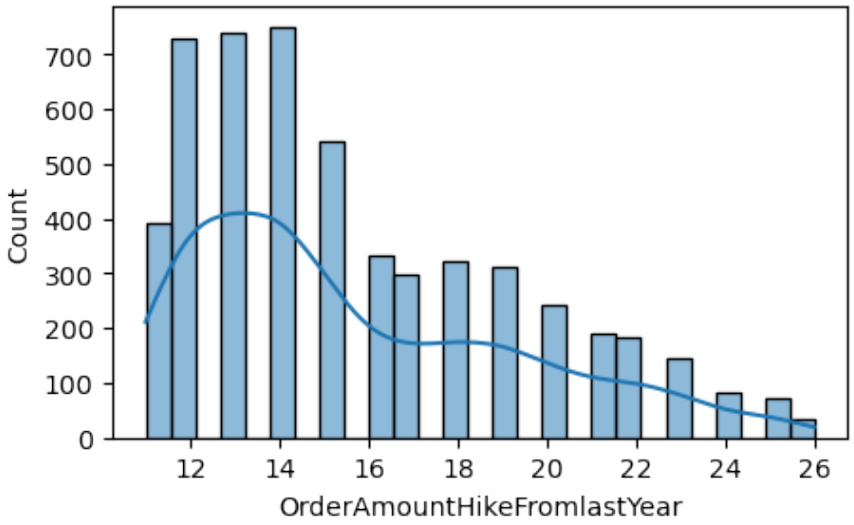
6.Appendix:

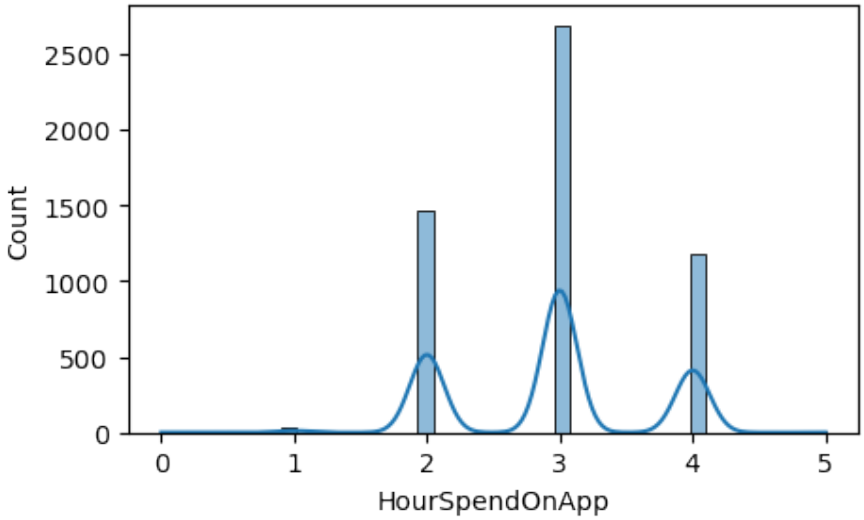
6.1 Histogram:

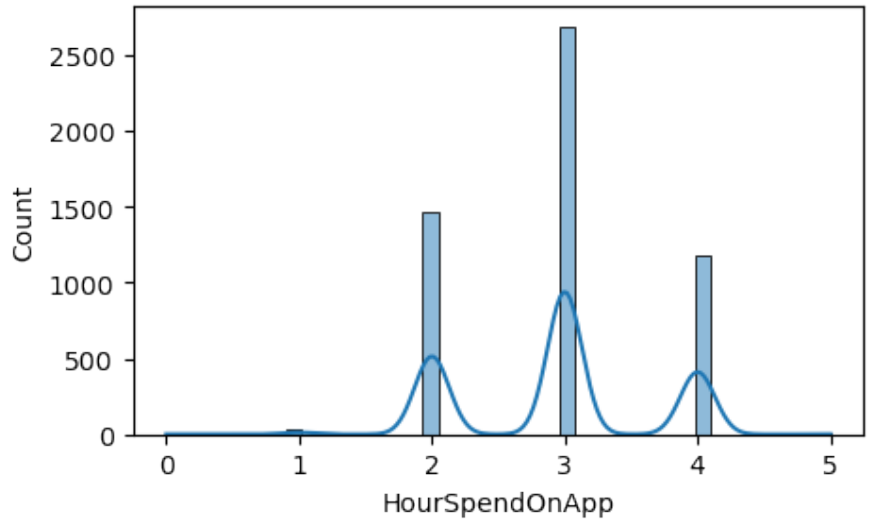





Count
WarehouseToHome
0
100
200
300
400
500
600
0
20
40
60
80
100
120

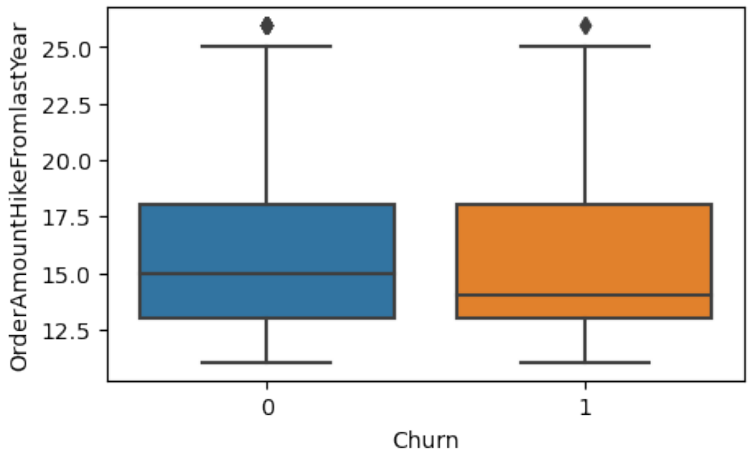

Count
OrderAmountHikeFromlastYear
0
100
200
300
400
500
600
700
12
14
16
18
20
22
24
26

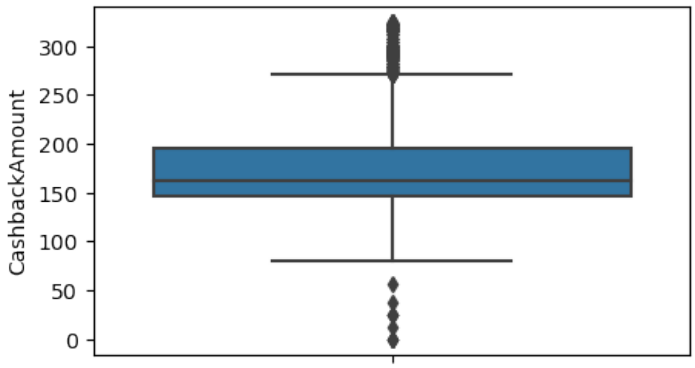

Count
HourSpendOnApp
0
500
1000
1500
2000
2500
0
1
2
3
4
5

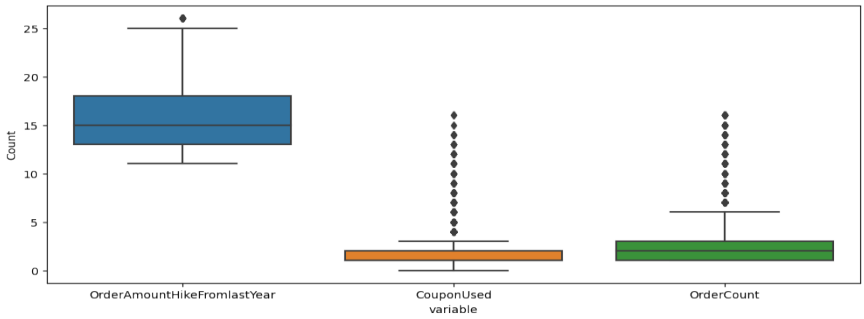


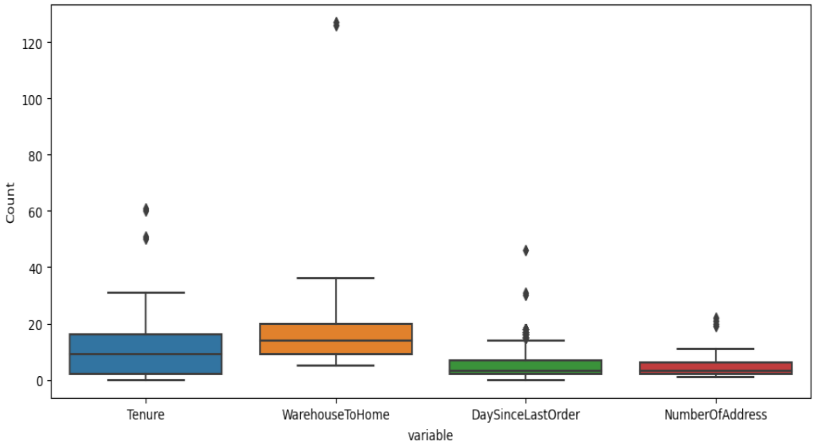
6.2 Boxplots:

Outliers:

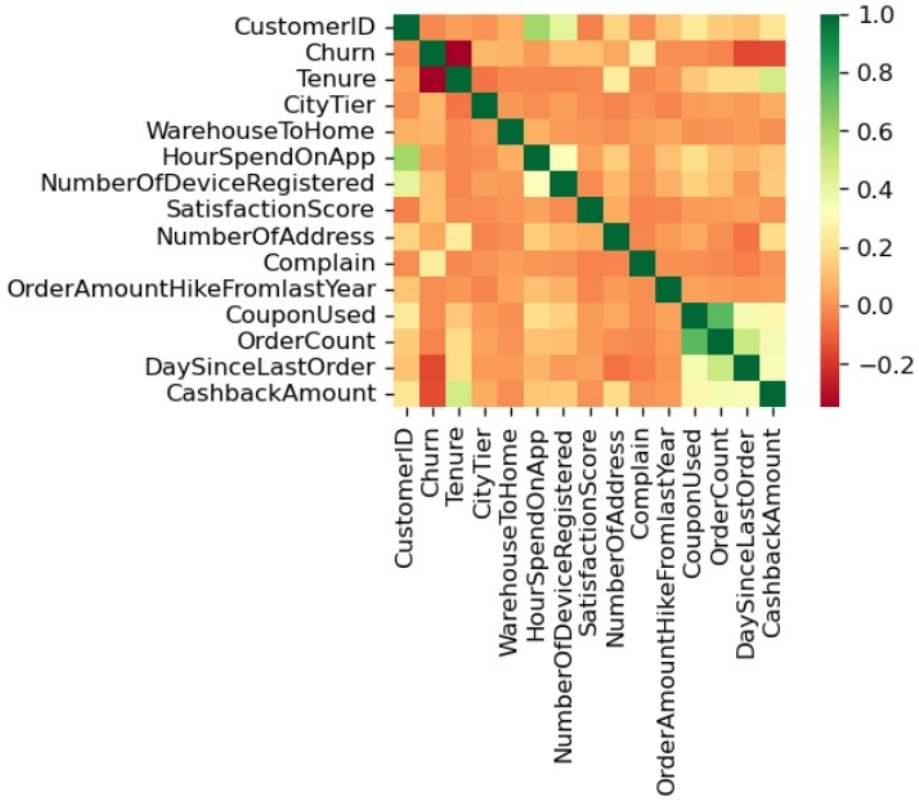




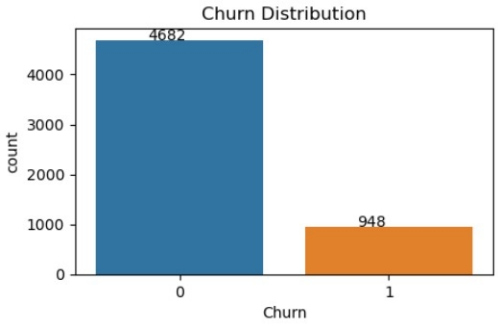

25
20
15
10
5
0
Count
OrderAmountHikeFromlastYear
CouponUsed
OrderCount
variable

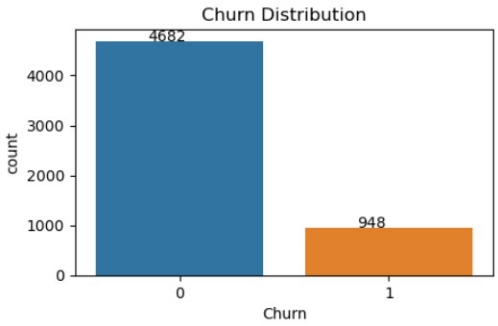

120
100
80
60
40
20
0
Count
Tenure
WarehouseToHome
DaySinceLastOrder
NumberOfAddress
variable

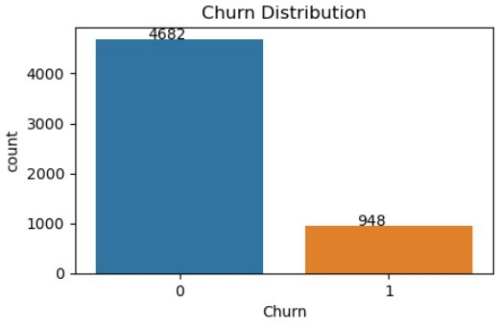
6.3. Heat Map

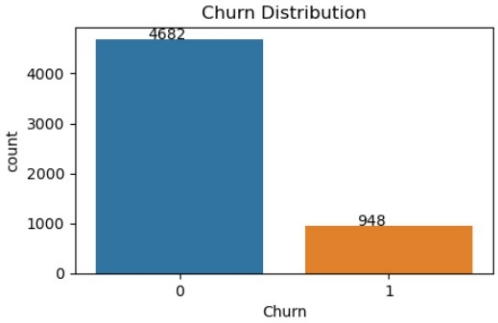

CustomerID
Churn
Tenure
CityTier
WarehouseToHome
HourSpendOnApp
NumberOfDeviceRegistered
SatisfactionScore
NumberOfAddress
Complain
OrderAmountHikeFromlastYear
CouponUsed
OrderCount
DaySinceLastOrder
CashbackAmount
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

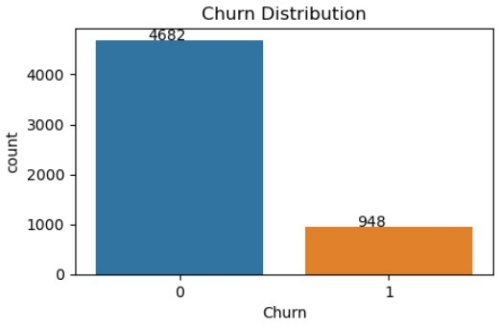
6.4 Bar Plot

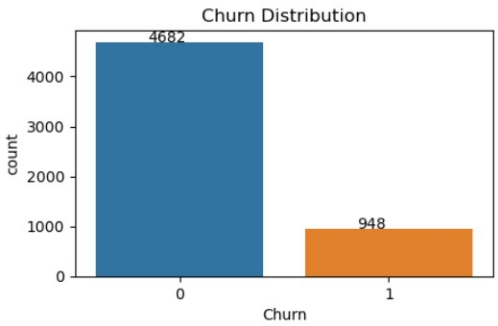

Churn Distribution
4682
948
count
4000
3000
2000
1000
0
0
1
Churn

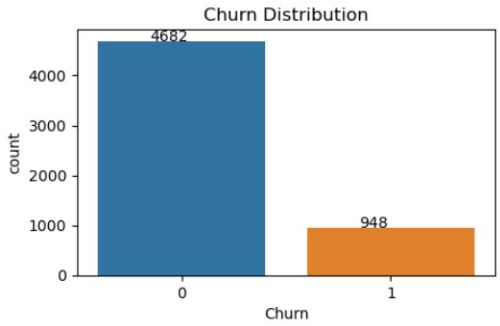

Churn Distribution
4682
948
count
4000
3000
2000
1000
0
0
1
Churn


Churn Distribution
4682
948
count
0
1000
2000
3000
4000
0
1
Churn


Churn Distribution
4682
948
count
0
1000
2000
3000
4000
0
1
Churn


Churn Distribution
4682
948
count
0
1000
2000
3000
4000
0
1
Churn


Churn Distribution
4682
948
count
0
1000
2000
3000
4000
0
1
Churn



6.5 Scatter Box