Ecommerce Customer Churn Analysis and Prediction

A project report submitted to ICT Academy of Kerala

in partial fulfillment of the requirements

for the certification of

CERTIFIED SPECIALIST

IN

DATA SCIENCE & ANALYTICS

submitted by

Team 6

Members

Nivin Ashok

Riya George

Alisha Naushad

ICT ACADEMY OF KERALA

THIRUVANANTHAPURAM, KERALA, INDIA

Oct 2022

List of Figures

Fig No	Figure Name	Page No
4.1	Plot of variables: churn,city tier,hours spend on app	15
4.2	Histogram of number of device registered, complain, satisfaction score	16
4.3	Histogram of continuous numericals	18
4.4	Bar Plot of Categorical Features	20
4.5	Customer churn percentage	21
4.6	Relationship of churn with numerical variables	23
4.7	Churn V/S Discrete Numericals	24
4.8	Relationship of Churn based on Gender	25
4.9	Relationship of Churn with Categorical variables	26
4.10	Heatmap depicts Correlation	27
4.11	Numerical feature Correlation with Churn	28
6.1	Feature importance for unscaled data	39
6.2	Feature importance for Data with Boxcox scaling for Xg boost	40
6.3	Feature importance for Unscaled Data with RFC	41
6.4	Feature importance for Data with log transformation for RFC model	42

List of Abbreviations

RFC - Random Forest Classifier

XGB - Extreme Gradient Boosting

Table of Contents

	Page No
Abstract	6
Chapter 1: Problem Definition	7
1.1 Overview	7
1.2 Problem Statement	7
Chapter 2: Introduction	8
Chapter 3: Literature Survey	9
Chapter 4: Data	12
4.1 Data Exploration	13
4.1.1 Descriptive Analysis	13
4.1.2 Univariate Analysis	15
4.1.3 Bivariate Analysis	22
4.1.4 Multivariate Analysis	27
Chapter 5: Data Preprocessing	29
5.1 Data Cleaning	29
5.2 Handling Missing Values	29
5.3 Outlier Treatment	30
5.4 Data Transformation	31
5.5 Encoding	32
5.6 Handling Imbalanced Data	32

5.7 Feature Engineering 32
Chapter 6: Model 33
6.1 Building Models 33
6.1.1 Splitting Data 33
6.1.2 Machine Learning Models 33
6.1.3 Performance Metrics 35
6.2 Feature Importance 39
6.3 Model Tuning 43
6.4 Model Deployment 43
6.4.1 Home Page 44
6.4.2 Prediction Page 45
Chapter 7: Result 46
Chapter 8: Conclusion 47
References 48

Abstract

E-commerce (electronic commerce) is the buying and selling of goods and services, or the transmitting of funds or data, over an electronic network, primarily the internet. For any business , Customer churn is one of the most crucial problem. It is the rate at which clients opt out of purchasing more of a company’s product or services. Churn analysis is the evaluation of a company's customer loss rate. It is important because it helps companies to understand why customers don't return for repeat business. Higher customer churn than your company or industry average can indicate a problem with pricing, service, product quality, delivery or some other aspect of the customer experience.

We are going to analysis an E-commerce dataset. The data set belongs to ABC online E-commerce company. Company wants to know the customers who are going to churn. Here, we employ classification algorithm to predict customers that are most likely to churn, from the 5600 instances and 20 columns available in the dataset. Using the ML model, companies can know in advance the customers that are most likely to give up the company’s services and therefore, come up with customer retention strategies..

1. Problem Definition

1.1 Overview

Maintaining a churn-rate is crucial for an e-commerce busines's long-term profit. However, not all companies have a system that can detect which of their customers will churn. This situation can have bad consequences for the company, when they give a benefit/promotion to a non-churning consumer and let the other customer churn. In this case, the company will compound an expense and lose potential revenue from the churned consumer. Or, if a company wants to play safe, they can give the benefit/promotion for all of their customer base. However, it will require a quite large expenditure without any certainty that the benefit was given to the right target.

1.2 Problem Statement

The data set belongs to ABC online E-commerce company. Company wants to know the customers who are going to churn.

2. Introduction

Simply put, customer churn occurs when customers or subscribers stop doing business with a company or service. Customer churn is a costly affair for these businesses as it essentially means they are losing out on customers. So, to avoid losing customers, the company needs to be able to predict which of their customers is more likely to churn.This is where machine learning comes into play. Using machine learning techniques the business creates a model using historical data and uses this model on current customers to generate the likelihood of a churn. This model also can give the company an idea as to which factor is influencing customer churn in their business. Once the company predicts the likely churners it can offer these customers some form of incentive to stay subscribed to the company. This machine learning model can not only increase the company's growth but also help them in saving costs by letting them know in which part of the business model they need to focus their money and efforts.

3. Literature Survey

Establishment of E-Commerce Customer Churn Prediction Model

Customer churn refers to the fact that the original customers of an enterprise stop to purchase enterprise goods or accept enterprise services, and instead accept the services of competitors (Wu et al., 2017). Churn rate prediction is applied extensively in the telecommunication sector. Ecommerce customer churn is a kind of churn that customers leave the enterprise, products or services for some reasons such as low quality or delay in delivery. E-commerce customer churn is a kind of customer churn in a non-contractual relationship scenario. In a non-contractual relationship, even if the termination of this kind of business-customer relationship occurs, it is difficult for the business to detect it in advance (Shao, 2016). For e-commerce companies, it is important to be able to accurately predict the high-value customer groups that are about to churn, and at the same time to study the purchasing habits of customers who have not churned in order to retain this type of customer group.

The value of e-commerce customer churn prediction is to merge ecommerce customer data over some time and establish e-commerce customer churn prediction models by analyzing customer purchase behaviors (Zhang, 2015). Then, provide e-commerce customer churn retention measures to reduce customer churn and identify high-value non-churn e-commerce customers and do a respectable job in customer retention. According to the research of Shao (2016), the remaining customers do not need high costs as the new customers want to bring high profit in ecommerce. Comparatively, the customer purchase behavior differs for both existing and new customers; however, it is essential to identify the reasons leading to the customer's loss. In support, Lu et al., (2018) stated that in the e-commerce sector, it is extremely important to analyze the loss of customers, predict the customers who might be lost, and then take corresponding measures to retain these customers and avoid their loss. At

present, most e-commerce companies have conducted an in-depth analysis of customer basic characteristic information and transaction behavior data, and then use various methods and technologies to establish and study customer churn prediction models, and finally use this to predict customer churn (Huang, 2018). Data mining technologies have been widely used in the customer relationship management of e-commerce companies, such as customer segmentation, customer churn prediction, and fraud analysis.

XGBoost

XgBoost stands for Extreme Gradient Boosting, which was proposed by the researchers at the University of Washington. It is a library written in C++ which optimizes the training for Gradient Boosting.In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

Bentéjac, Candice & Csörgő, Anna & Martínez-Muñoz, Gonzalo. (2019). A Comparative Analysis of XGBoost. XGBoost is a scalable ensemble technique based on gradient boosting that has demonstrated to be a reliable and efficient machine learning challenge solver. This work proposes a practical analysis of how this novel technique works in terms of training speed, generalization performance and parameter setup. In addition, a comprehensive comparison between XGBoost, random forests and gradient boosting has been performed using carefully tuned models as well as using the default settings. The results of this comparison may indicate that XGBoost is not necessarily the best choice under all circumstances. Finally an extensive analysis of the XGBoost parametrization tuning process is carried out.

SMOTE

SMOTE (synthetic minority oversampling technique) is used to artificially synthesize new minority samples to reduce the imbalance of the categories which is common in churn datasets. The basic idea is to insert minority virtual samples between the minority classes that are close together. For each sample of a minority, it searches its k nearest neighbors, which randomly selects k -nearest neighbors in any of points that synthesize a new minority sample. Thus, SMOTE can enhance the performance of any machine learning algorithm if the outcome of the dataset was imbalanced.

Rout, Neelam & Mishra, Debahuti & Mallick, Manas & Mallick, Pradeep Kumar. (2022). Dealing with Imbalanced Data. 10.1007/978-981-16-9488-2_35. Class imbalance all around presents in real-world applications, which has brought more curiosity from different fields. While emphasising on accuracy for performance evaluation, studying from unbalanced data may produce unproductive outcomes. Cost-sensitive, sampling, ensemble approach and other hybrid methodologies have all been used in the past to address this imbalance problem. In machine learning, the ensemble approach is used to increase the accuracy of single base classifiers by aggregating numerous of them. To handle the issues due to imbalanced data, ensemble algorithms have to be formed specifically. Several performance assessing functions showed that the ensemble method outperformed the other techniques. In this article, different methods are described to handle imbalanced datasets with the special description of SMOTE with the ensemble method. The complexity of the ensemble model is defined. The clustering methods are also used to manage the issues due to imbalanced datasets.

4.DATA

The data used in the models have been obtained from Kaggle. The data contains 5630 observations and 20 variables including the target variable.

Variables	Description
CustomerID	Unique customer ID
Churn	Churn Flag
Tenure	Tenure of customer in organization
PreferredLoginDevice	Preferred login device of customer
CityTier	City tier
WarehouseToHome	Distance in between warehouse to home of customer
PreferredPaymentMode	Preferred payment method of customer
Gender	Gender of customer
HourSpendOnApp	Number of hours spend on mobile application or website
NumberOfDeviceRegistered	Total number of deceives is registered on particular customer
PreferedOrderCat	Preferred order category of customer in last month
SatisfactionScore	Satisfactory score of customer on service
MaritalStatus	Marital status of customer
NumberOfAddress	Total number of added added on particular customer
Complain	Any complaint has been raised in last month
OrderAmountHikeFromlastYear	Percentage increases in order from last year
CouponUsed	Total number of coupon has been used in last month
OrderCount	Total number of orders has been places in last month
DaySinceLastOrder	Day Since last order by customer
CashbackAmount	Average cashback in last month

4.1 Data Exploration

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

4.1.1 Descriptive Analysis

The following were found out from the initial descriptive analysis:

- 5630 rows and 20 columns

  There are 5630 rows and 20 columns including the target feature churn.

- 15 numerical and 5 categorical variables

  15 numerical features which includes discrete and continuous numerical feature and 5 categorical features are in this dataset

- There are no duplicate entry

  No duplicate entries were found

- There are null values in some columns

Variable	No. of Null Values
DaySinceLastOrder	307
OrderAmountHikeFromlastYear	265

Tenure	264
OrderCount	258
CouponUsed	256
HourSpendOnApp	255
WarehouseToHome	251

- There's a quite a lot of features with outliers
- Little amount of skewness are present in all continuous variables

Variable name	Skewness
Churn	1.77
Tenure	0.74
CityTier	0.74
WarehouseToHome	1.62
HourSpendOnApp	-0.03
NumberOfDeviceRegistered	-0.40
SatisfactionScore	-0.14
NumberOfAddress	1.09
Complain	0.95
OrderAmountHikeFromlastYear	0.79
CouponUsed	2.55

OrderCount	2.20
DaySinceLastOrder	1.19
CashbackAmount	1.15

4.1.2 Univariate Analysis

4.1.2.1 Discrete Numericals

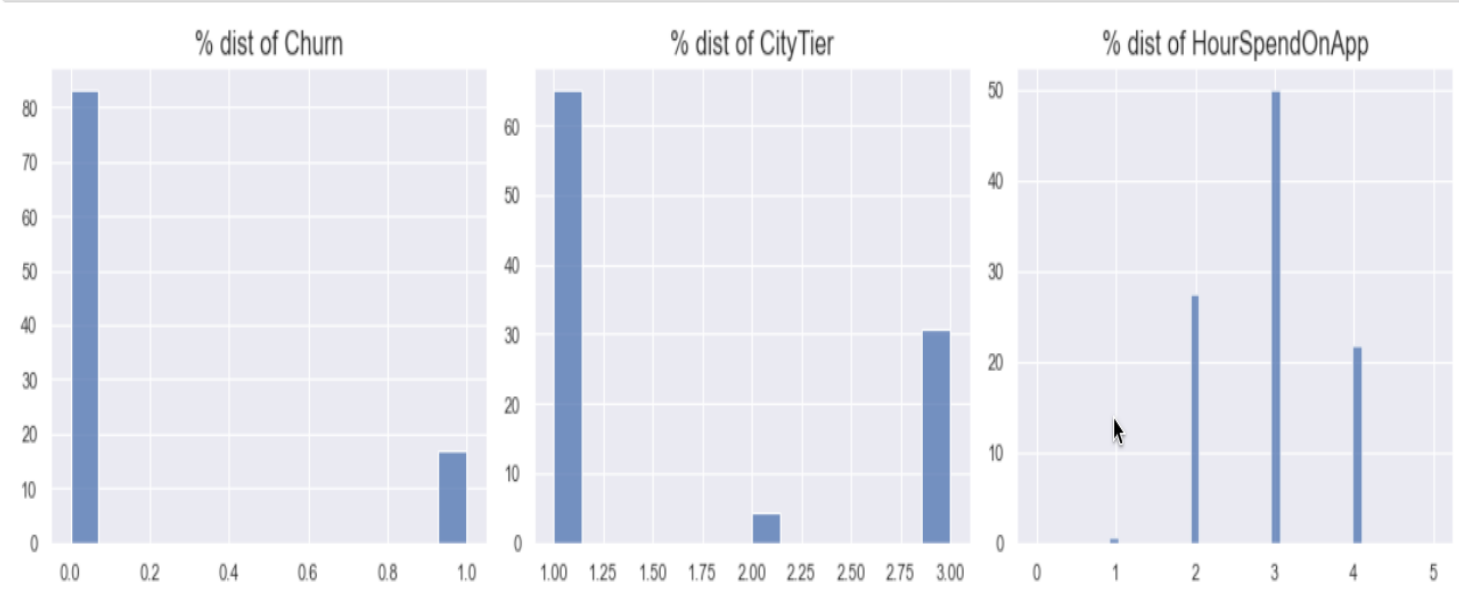


Fig 4.1 plot of variables Churn,City Tier,Hourspend on App

Insights

- Target feature Churn is imbalanced.
- Most customers are in tier 1 city.
- Columns HourSpendOnAPP have mode in 3 hours

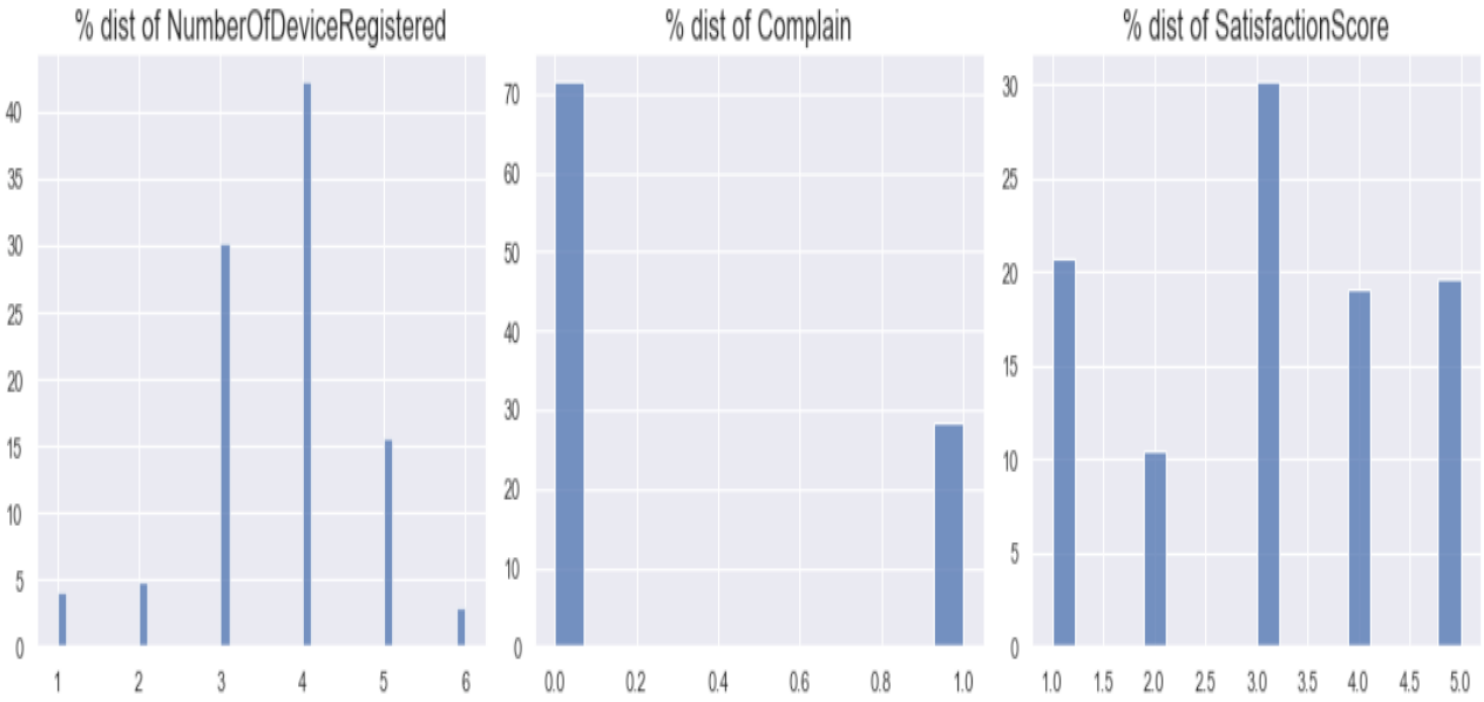
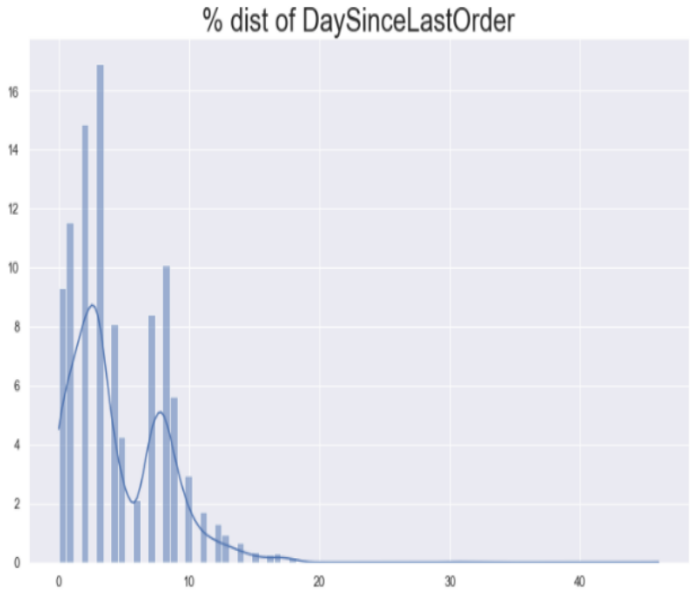


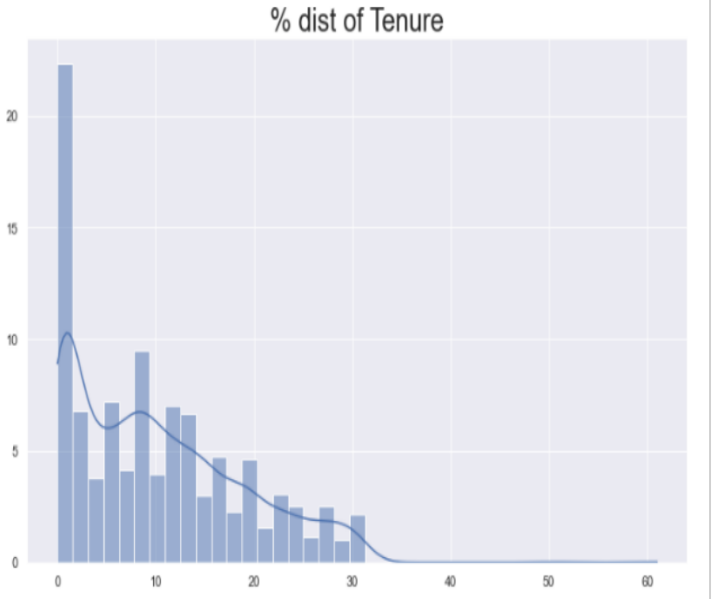
Fig 4.2 Histogram of Number of device registered, Complain, Satisfaction score

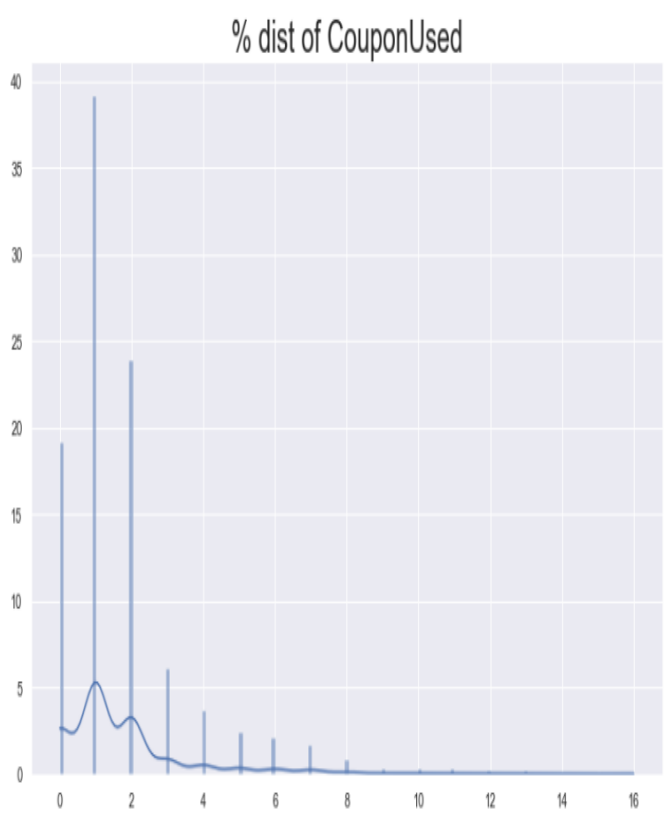
Insights

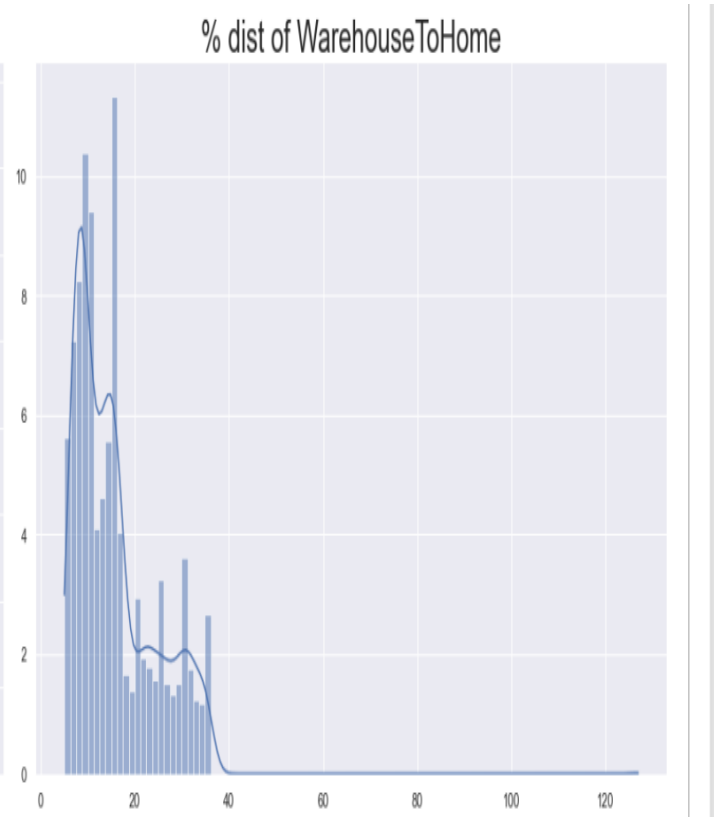
- Most customers registered 4 devices.
- Complaints were barely reported by customers
- Majority Satisfaction score is 3

4.1.2.2 Continuous numerical









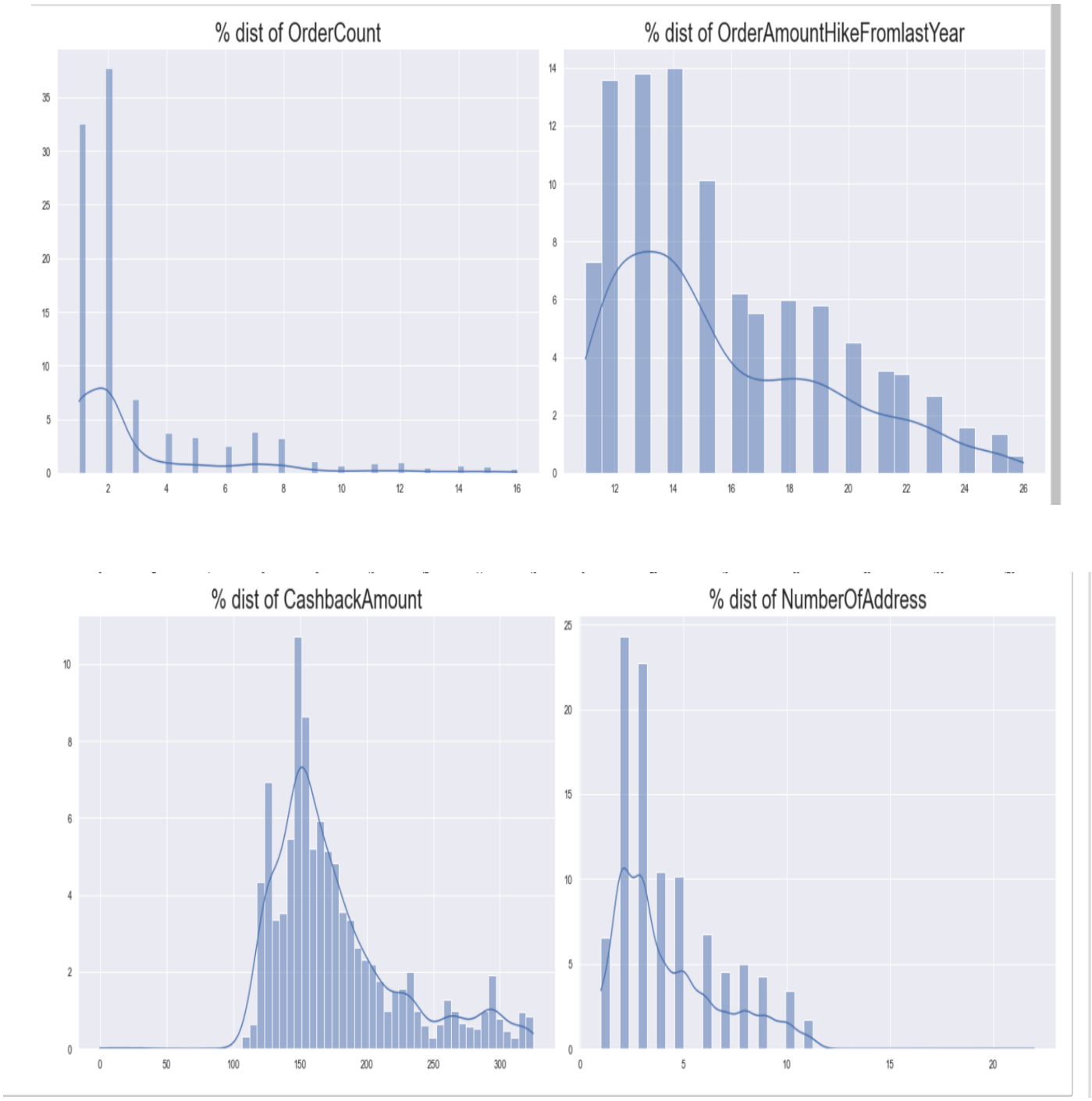
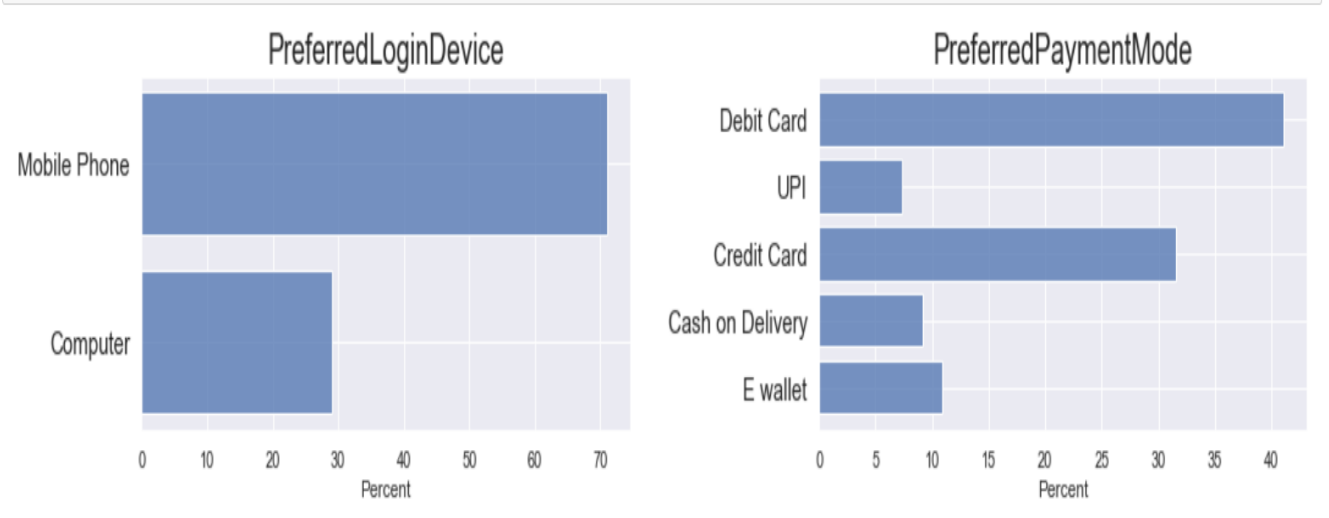


Fig 4.3 Histogram of Continuous Numericals

Insights

- The data are widely distributed in Satisfaction score, Order amount hike from last year, Days since last order and Cashback amount.
- Most of the customers are new as tenure is around 1 month which means recently joined.
- Most of the continuous numerical columns are skewed

5.2.3 Categorical Variables



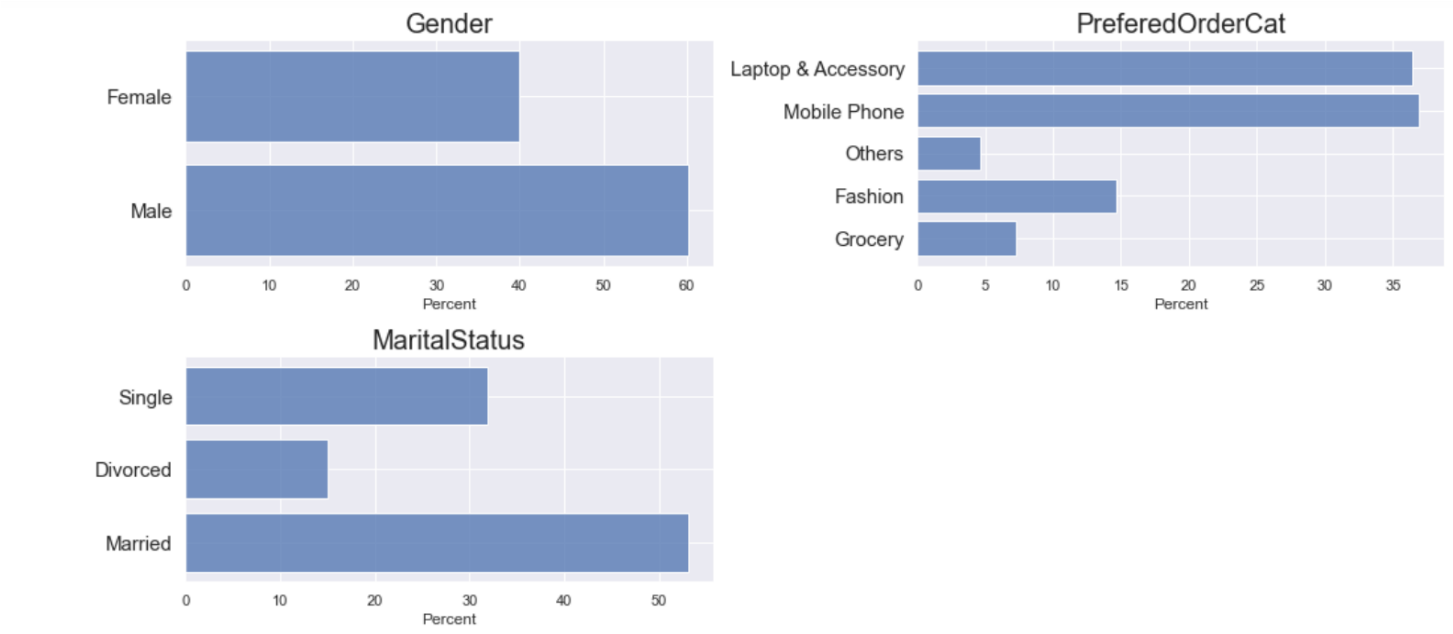


Fig 4.4 Barplot of Categorical Features

Insights

- Mobiles Phones were preferred by most customers
- Most customers chose Debit Card for payment
- Majority of the customers were male
- Laptop and Accessories were the most ordered category
- Most of the customers are married

5.3 Target Variable

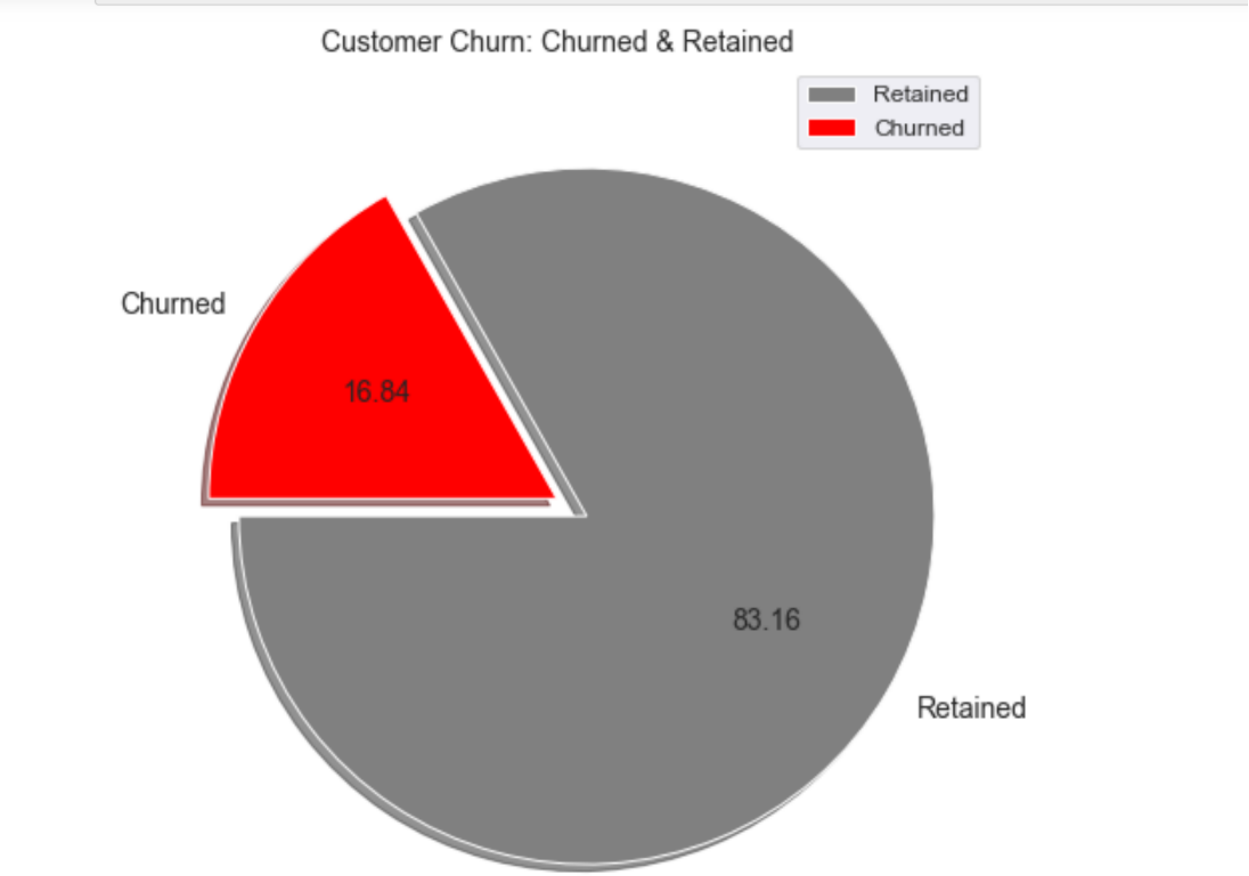


Fig 4.5: Customer Churn Percentage

•16.8% of customers have churned from total customers

•83.16 % customers have retained out of total customers

•The pie chart shows us that in Churn the data isn't uniformly distributed in other words, it is *imbalanced*.

•Label definition:

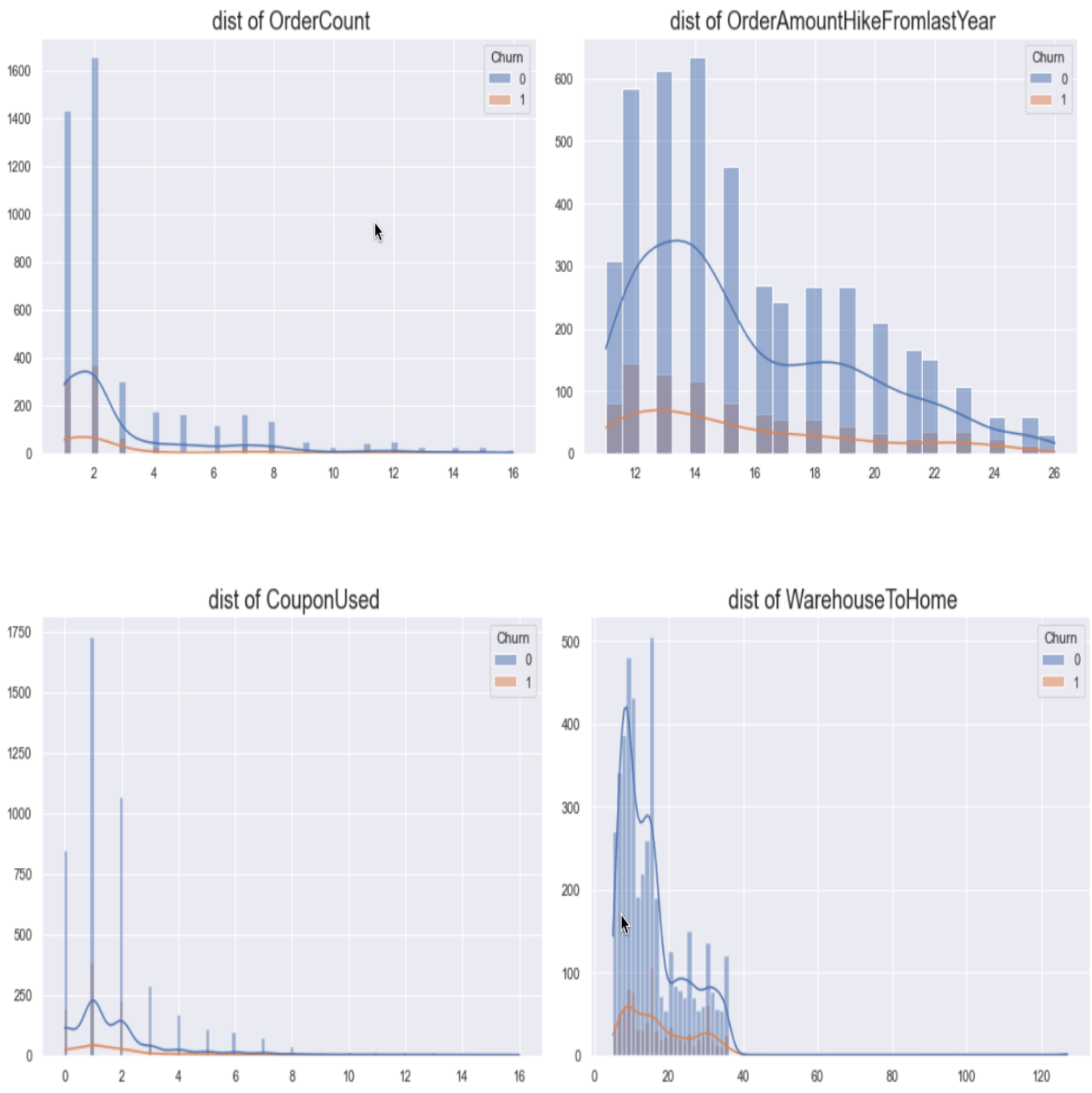
0 = Customer not churn (Retained customer)

1 = Customer churn (stop using the service)

4.1.3 Bivariate Analysis:

This method of analysis describes the relationship between the variables

4.1.3.1 Numerical Variables



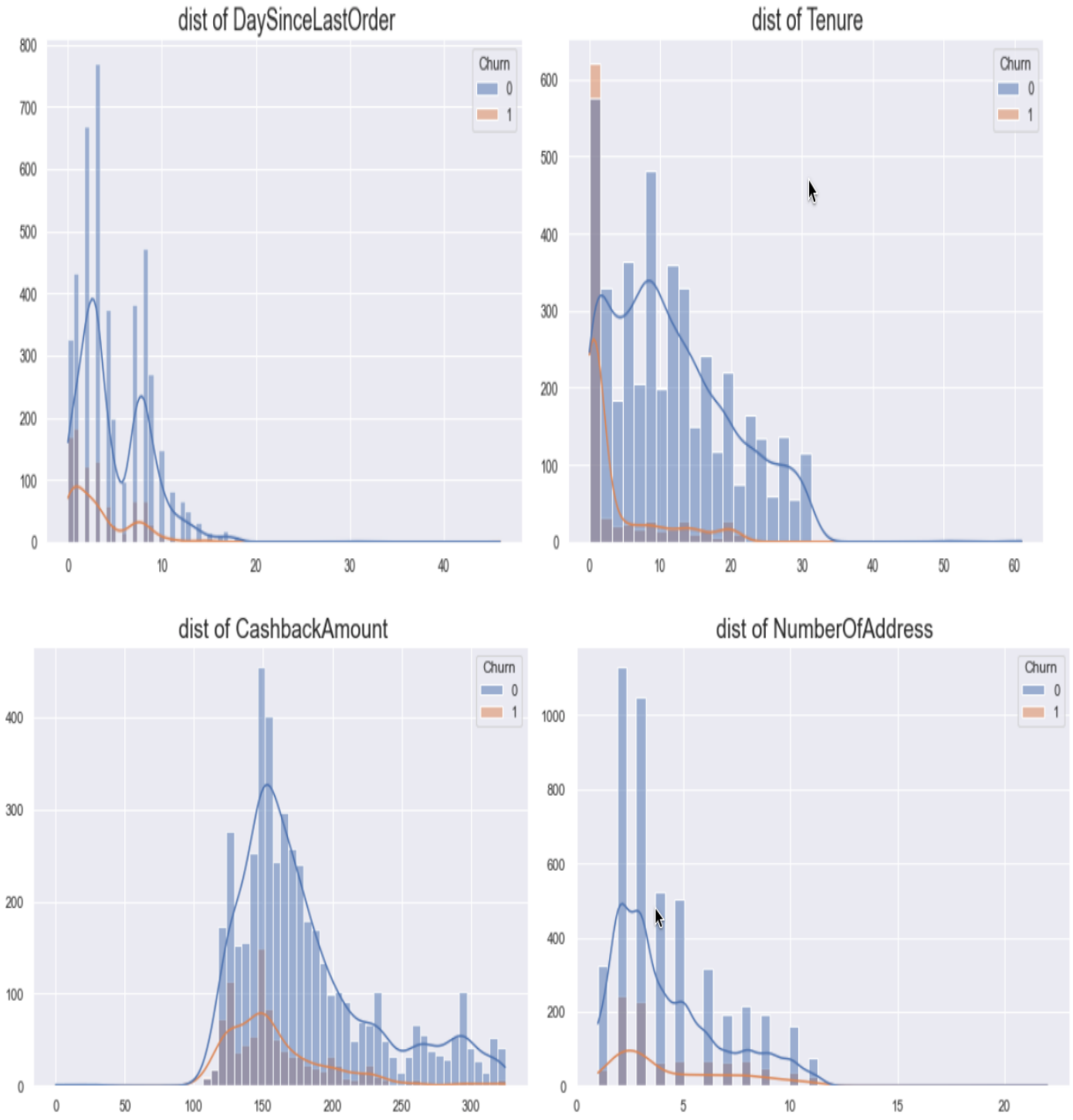


Fig 4.6 Relationship of Churn with numerical variables

Insights

- The distribution pattern of continuous variables is almost same for Churn as well as no churn customers , with slight variations .
- Churn is decreasing for Cashback amounts greater than 150.
- Churn is slowly increasing after around 5 days since the last order.
- The customers with comparatively less Cashback amount, more distance from Warehouse to home are likely to churn
- Customers with more Satisfaction score and less number of Days since last order are tending to Churn.

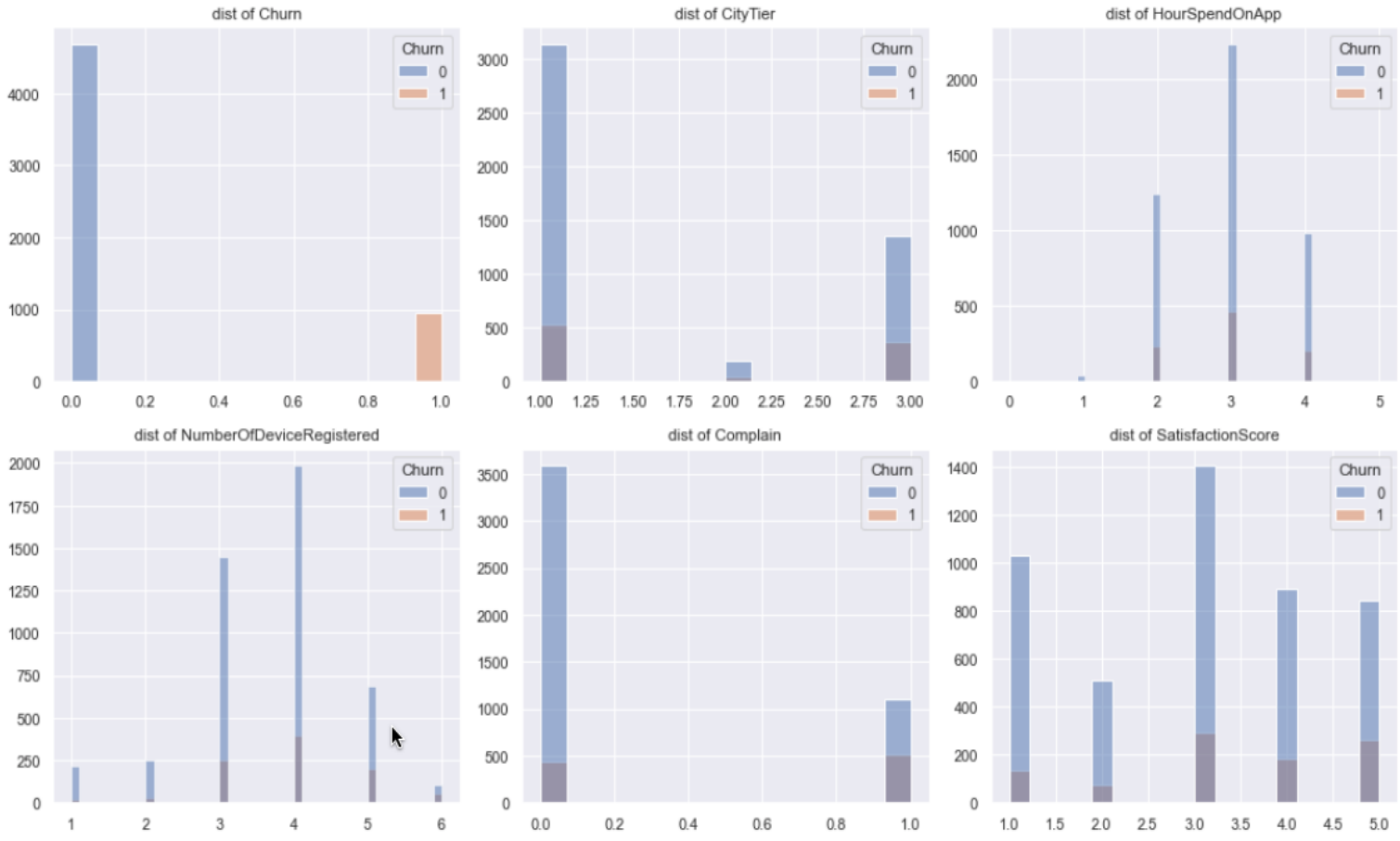


Fig 4.7 Churn V/s Discrete Numericals

Insights

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

4.1.3.2 Categorical

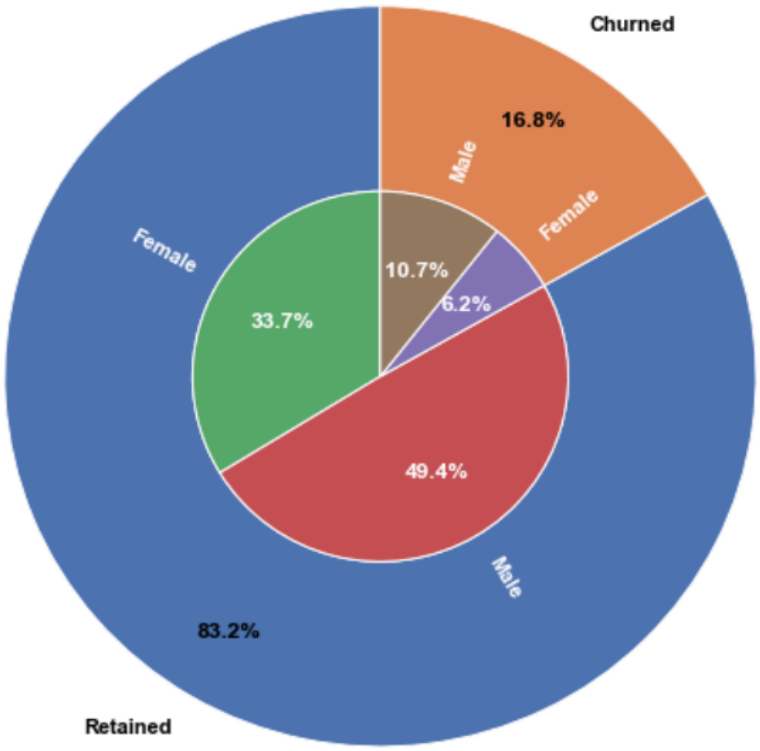


Fig 4.8 : Churn based on gender

Insights:

- Male customers churn more than female

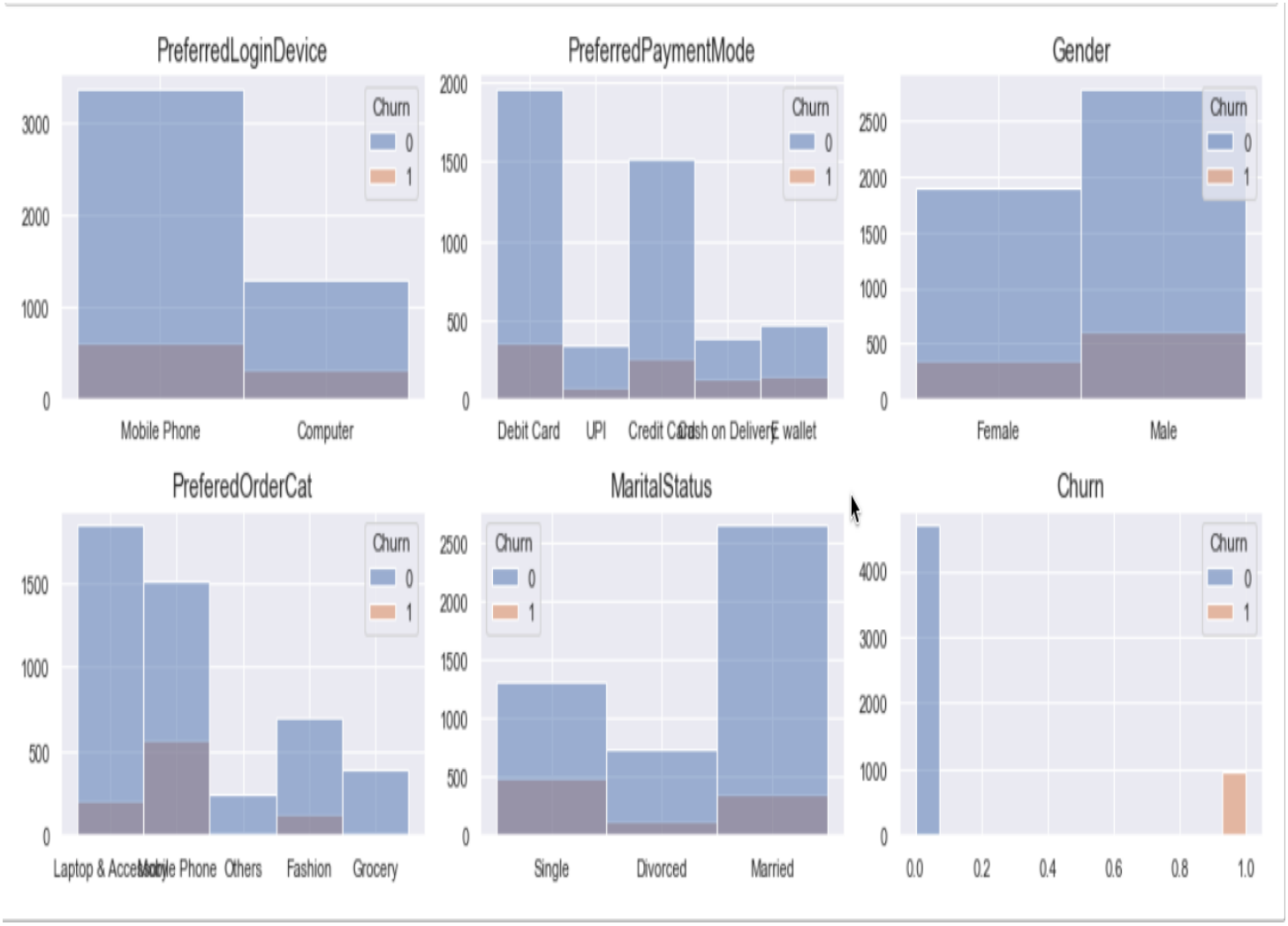


Fig 4.9: Relationship of Churn with Categorical Features

Insights

- Mobiles Phones were preferred by both customers who churned and retained
- Most customers chose Debit Card for payment
- Churned customers mostly paid through Debit card
- Most Married customers retained while most of the single customers churned

- The most ordered category was Laptop and Accessories for customers who didn't churn while mobile phones were for customers who churned

4.1.4 Multivariate Analysis:

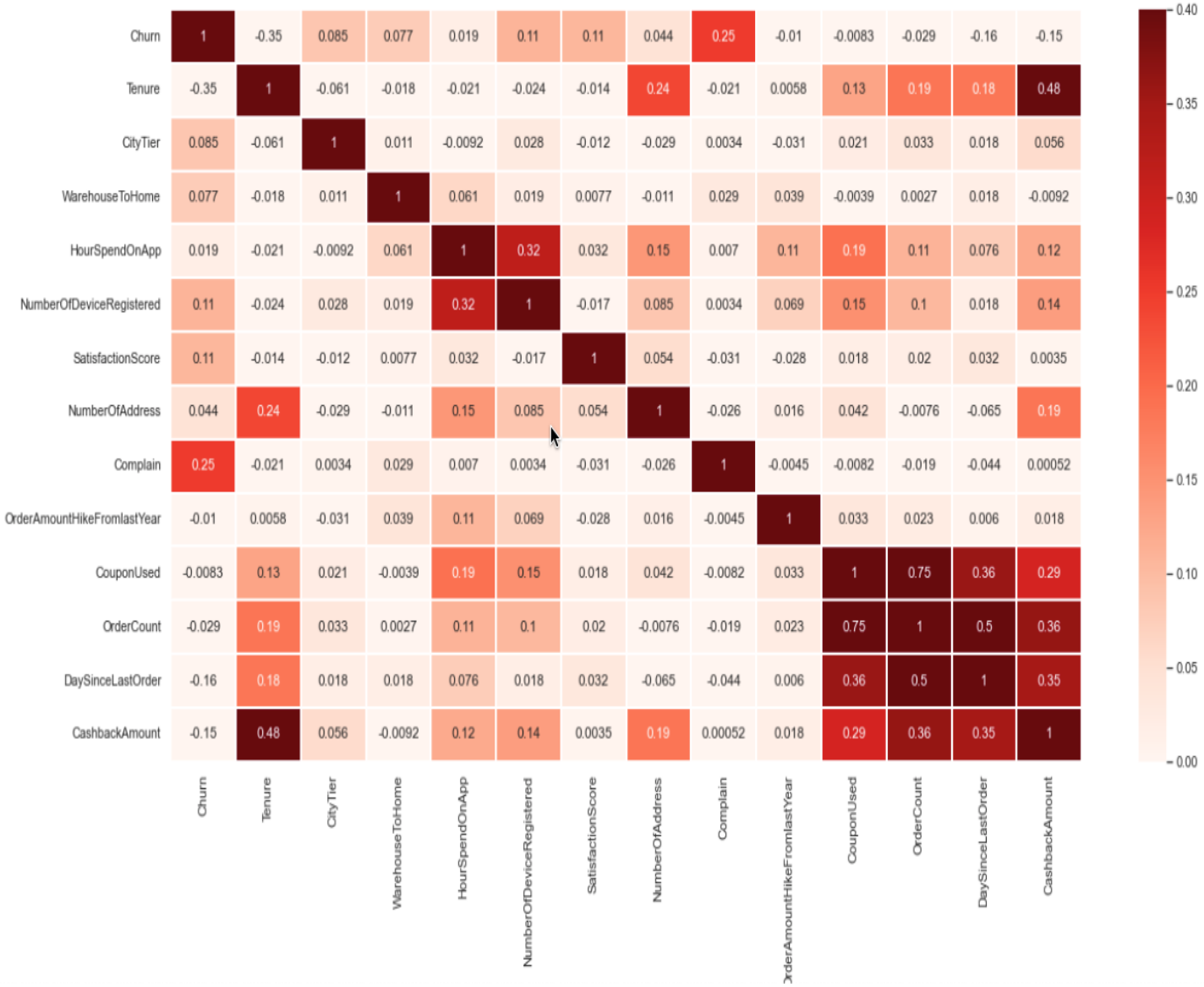


Fig 4.10 Heatmap depicting Correlation

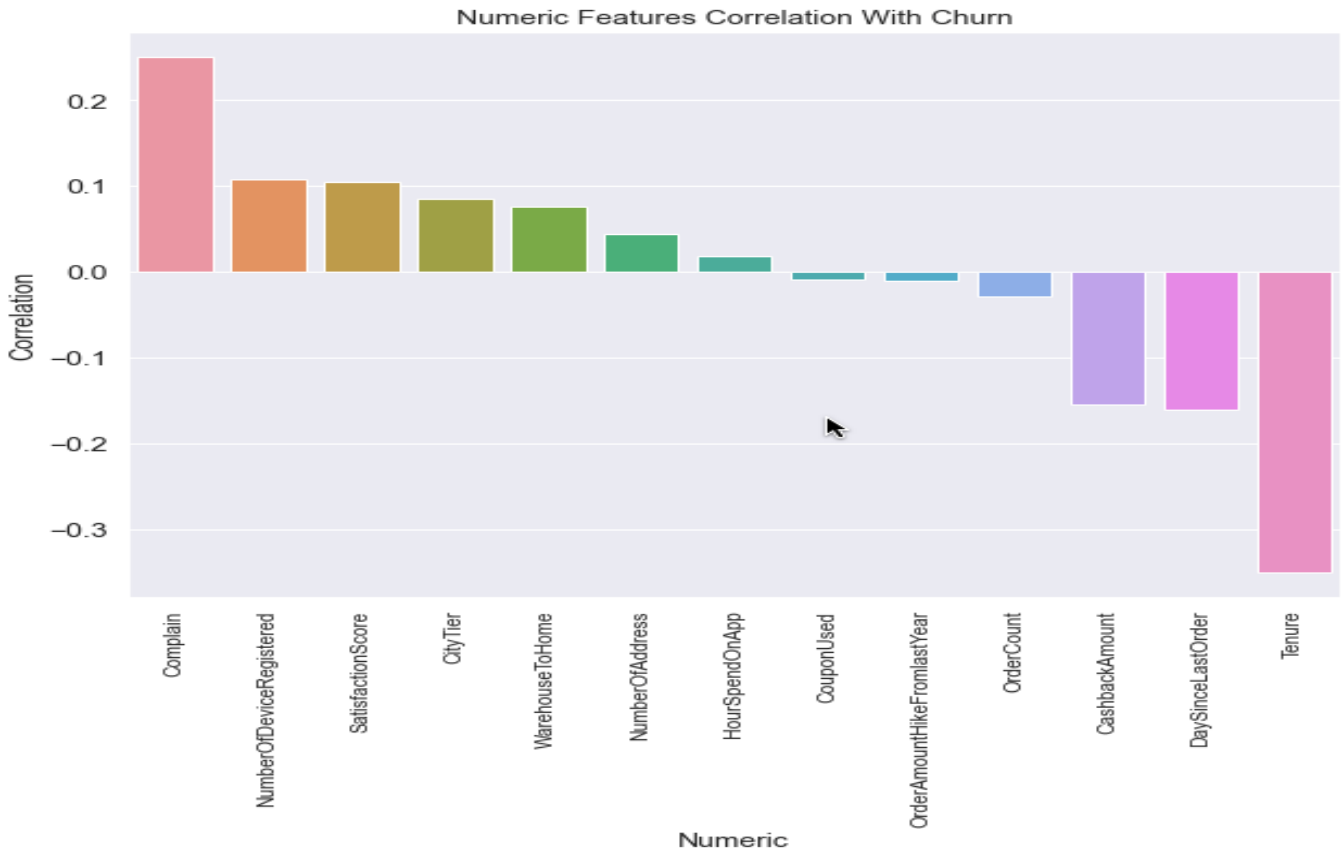


Fig 4.11 Numerical Feature correlation with Churn

Insights:

- Target,Churn has a Positive correlation with CityTier, WarehouseToHome, NumberOfDeviceRegistered, SatisfactionScore, Complain
- Churn has a Negative correlation with Tenure, DaySinceLastOrder, and CashbackAmount.
- Churn with HourSpendOnApp, NumberOfAddress, OrderAmountHikeFromlastYear, CouponUsed, OrderCount has correlation very weak ~0, this indicates that the feature may not have potential.

5.DATA PREPROCESSING

When creating a machine learning project, it is not always a case that we come across clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put it in a formatted way. So for this, we use data preprocessing.

5.1 Data Cleaning

Initially, in PreferedOrderCat column, there were two variables indicating the same meaning “Mobile” and “Mobile Phone” which were grouped in one level labeled as “Mobile”. Then, In PreferredPaymentMode column “Cash on Delivery” and “COD” levels that have been grouped in one level labeled as “Cash on Delivery” in addition to “Credit Card” and “CC” levels have been grouped in in group called “Credit Card”

5.2 Handling Missing Values

- There was 7 features with missing values:

Features	Missing Values
Tenure	264
WarehouseToHome	251
HourSpendOnApp	255
CouponUsed	256
OrderCount	258
DaySinceLastOrder	307

OrderAmountHikeFromlastYear	265

- We used grouby to group the missing values by churn 0 and 1 and filled it with corresponding medians

5.3 Outlier Treatment:

The dataset has many features with outliers,but most of them have genuine values,which are retained and the rest is capped.

CashbackAmount :

For Cashback amount ,the support was from 0 to 325. From the outlier plot, values from 0-50 and 275-325 were outliers. Details will be retained as such as they are genuine values and not any extreme values. Will choose an ML model robust with outliers.

DaySinceLastOrder :

The support for this feature is from 0 to 46. There were 3 entries with value of the feature greater than 20.Out of them, the extreme value 46 had churn 1 and other 2 values had churn 0.We will be retaining all the 3 entries as all 3 entries have varied info which will be helpful in modelling.

OrderAmountHikeFromlastYear:

The support for this feature is 0 to 26. Even though many data points fall in the outlier category, since the support is small range and has varied information, the datapoints considered outlier will be retained.

Tenure :

It has a support from 0 to 61.99th th percentile value for tenure is 30. ie 99% of values fall within 30. Remaining values which are above 30 are all Churn =0. So we capped all the values of tenure above 99 percentile to 30.

WarehouseToHome :

This feature has a support from 5 to 127.There were few extreme points as per the outlier plot. There were only 2 datapoints which were greater than 36 (>99 percentile) and both of the points had churn =0. So we capped the extreme points to 36.

5.4 Data Transformation

- The transformation methods we used here are Log transform, BoxCox Method and Robust Scaling method.
- We used these different methods in different copies of the same dataset.
- We had 3 discrete numerical features which had many unique values. In order to bring these features under a manageable range and to match with the range of scaled continuous variables , we did binning.It is done for the following features. Bins are selected based on the distribution of data.
    - NumberOfAddress : bins : [0, 3, 7, 25], labels:['low','med','high']
    - CouponUsed : [-1, 0, 1, 2, 20], labels: ['zero','low','med','high']
    - OrderCount : [0, 1, 2,3, 20], labels: ['zero','low','med','high']

5.5 Encoding

We used One Hot Encoding to encode the categorical variables.

For encoding binned discrete numerical variables , LabelEncoder is used.

5.6 Handling Imbalanced Data

Our target variable was imbalanced. Hence, we used SMOTE to upsample the target variable.

5.7 Feature Engineering

We created a new column Tenure in years from variable Tenure by converting it into years.

6.MODEL

The next step is to build the models for the Churn prediction from the given dataset, model results are interpreted to find the best suited model.

6.1 Building Models:

Various models are built using various machine learning algorithms.

6.1.1 Splitting the dataset

Initially the processed dataset is splitted into 2 following subsets by dropping CustomerID and Churn 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.

6.1.2 Machine Learning Models

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

- **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

- **Random Forest:**Random Forest is a robust machine learning algorithm that can be used for a variety of tasks including regression and classification.

- **Decision Tree:**Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

- **Extra Trees Classifier:** It is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result.

- **Bagging Classifier**:Bagging is a technique for improving the accuracy of predictions made by a supervised learning algorithm. The basic idea is to train a number of different models on different randomly-selected subsets of the training data, and then to combine the predictions of these models using some sort of voting scheme.

- **Gaussian Naive:**Gaussian Naive Bayes is a probabilistic classification algorithm based on applying Bayes' theorem with strong independence assumptions.

- **SVM:**The principle of SVM is to find an hyperplane which, can classify the training data points into labeled categories. The input of SVM is the training data and use this training sample point to predict class of test point

- **Gradient Boosting Classifier:**Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model,when combined with previous models, minimizes the overall prediction error.

- **ADA Boost Algorithm:**.AdaBoost also called Adaptive Boosting is a type of ensemble learning technique. The most common algorithm

used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split.

- **XGBoost**: It stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

6.1.3 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. Here, we used F1-Score, Precision and Accuracy.

Model 1:Logistic Regression

Data transformation	F1 Score	Accuracy	Precision
Unscaled	0.744	0.832	0.745
with log transformation	0.780	0.859	0.800
Boxcox Scaling	0.781	0.859	0.797
Robust Scaling	0.752	0.839	0.765

Model 2:Random Forest

Data transformation	F1 Score	Accuracy	Precision
Unscaled	0.960	0.974	0.971
with log transformation	0.949	0.967	0.962

Boxcox Scaling	0.967	0.978	0.959
Robust Scaling	0.970	0.980	0.969

Model 3:Decision Tree

Data transformation	F1 Score	Accuracy	Precision
Unscaled	0.913	0.943	0.909
with log transformation	0.933	0.956	0.936
Boxcox Scaling	0.925	0.951	0.933
Robust Scaling	0.902	0.936	0.913

Model 4:Extra Trees Classifier

Data transformation	F1 Score	Accuracy	Precision
Unscaled	0.969	0.980	0.975
with log transformation	0.974	0.983	0.980
Boxcox Scaling	0.973	0.982	0.969
Robust Scaling	0.976	0.984	0.974

Model 5:Bagging Classifier

Data transformation	F1 Score	Accuracy	Precision
Unscaled	0.938	0.960	0.959
with log transformation	0.953	0.969	0.954

Boxcox Scaling	0.946	0.965	0.951
Robust Scaling	0.951	0.968	0.975

Model 6:Gaussian Naive

Data transformation	F1 Score	Accuracy	Precision
Unscaled	0.631	0.682	0.510
with log transformation	0.679	0.748	0.585
Boxcox Scaling	0.672	0.733	0.564
Robust Scaling	0.643	0.699	0.527

Model 7:SVM

Data transformation	F1 Score	Accuracy	Precision
Unscaled	0.736	0.810	0.679
with log transformation	0.783	0.862	0.814
Boxcox Scaling	0.795	0.868	0.812
Robust Scaling	0.798	0.869	0.807

Model 8:Gradient Boosting Classifier

Data transformation	F1 Score	Accuracy	Precision
Unscaled	0.866	0.926	0.901
with log transformation	0.869	0.916	0.898
Boxcox Scaling	0.863	0.912	0.884

Robust Scaling	0.878	0.920	0.889

Model 9:ADA Boost Algorithm

Data transformation	F1 Score	Accuracy	Precision
Unscaled	0.855	0.905	0.858
with log transformation	0.816	0.881	0.828
Boxcox Scaling	0.831	0.890	0.839
Robust Scaling	0.802	0.874	0.827

Model 10:XGBoost

Data transformation	F1 Score	Accuracy	Precision
Unscaled	0.976	0.984	0.982
with log transformation	0.976	0.984	0.980
Boxcox Scaling	0.972	0.982	0.971
Robust Scaling	0.976	0.984	0.980

6.2 Feature Importance

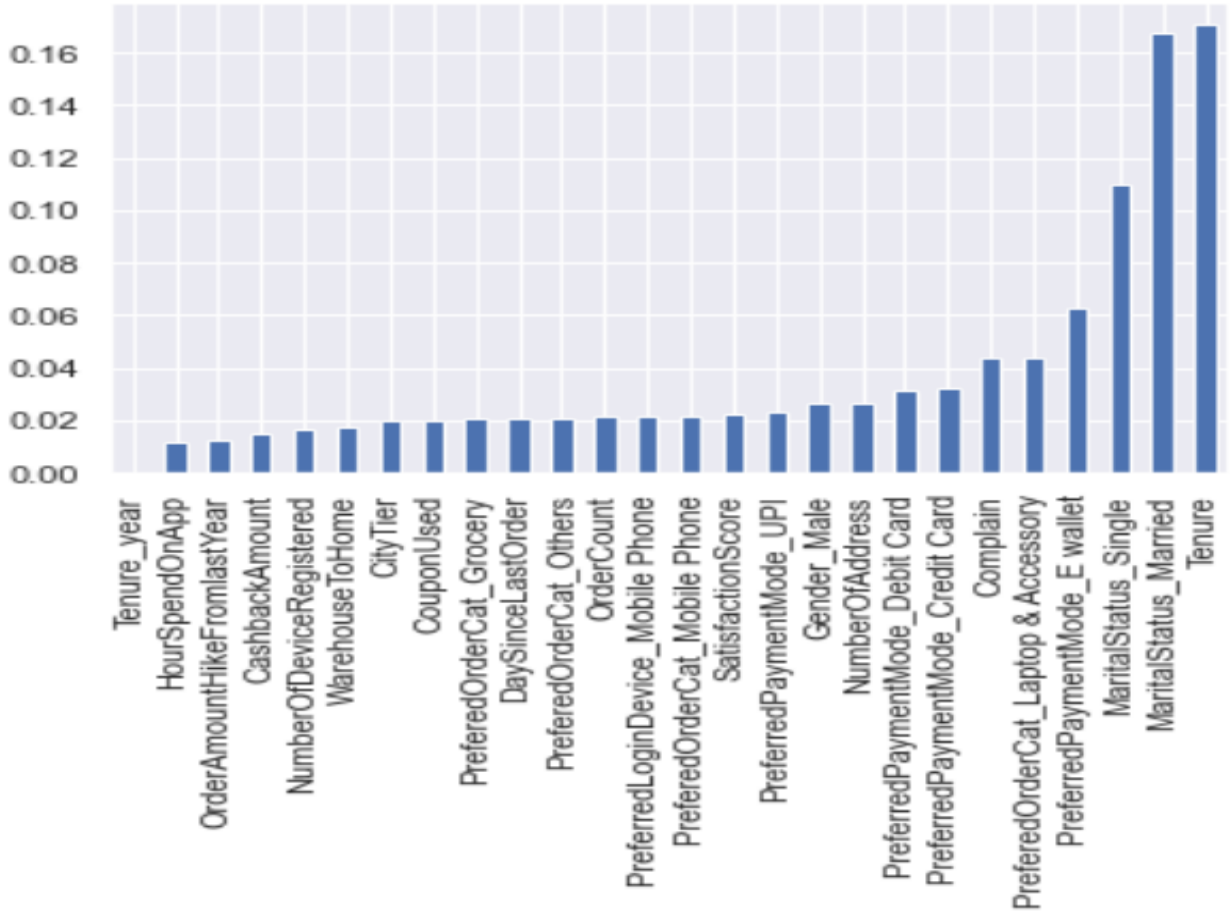


Fig 6.1 Feature Importance for Unscaled Data for Xg boost

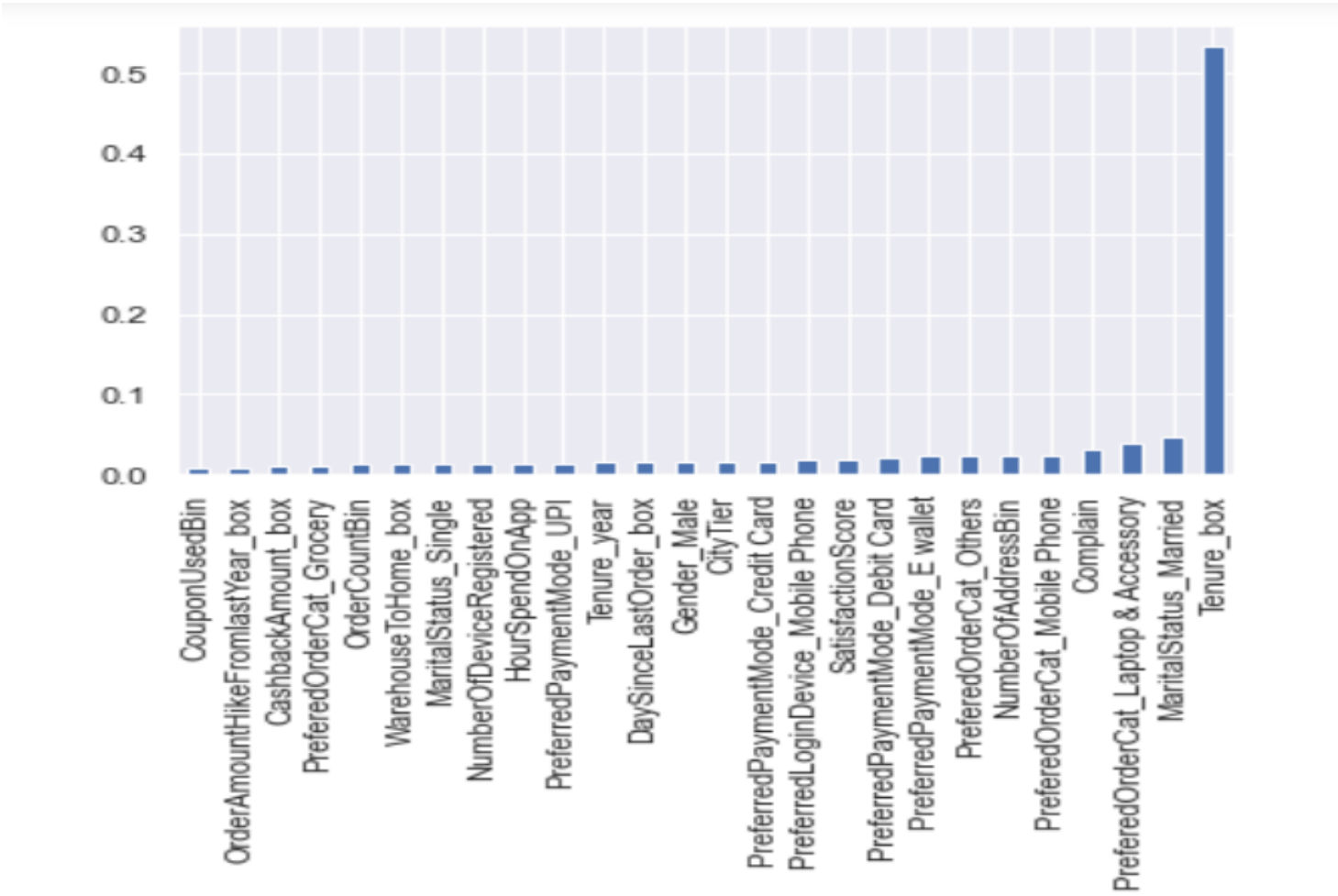


Fig 6.2 Feature Importance of Data with boxcox Scaling for Xg boost model

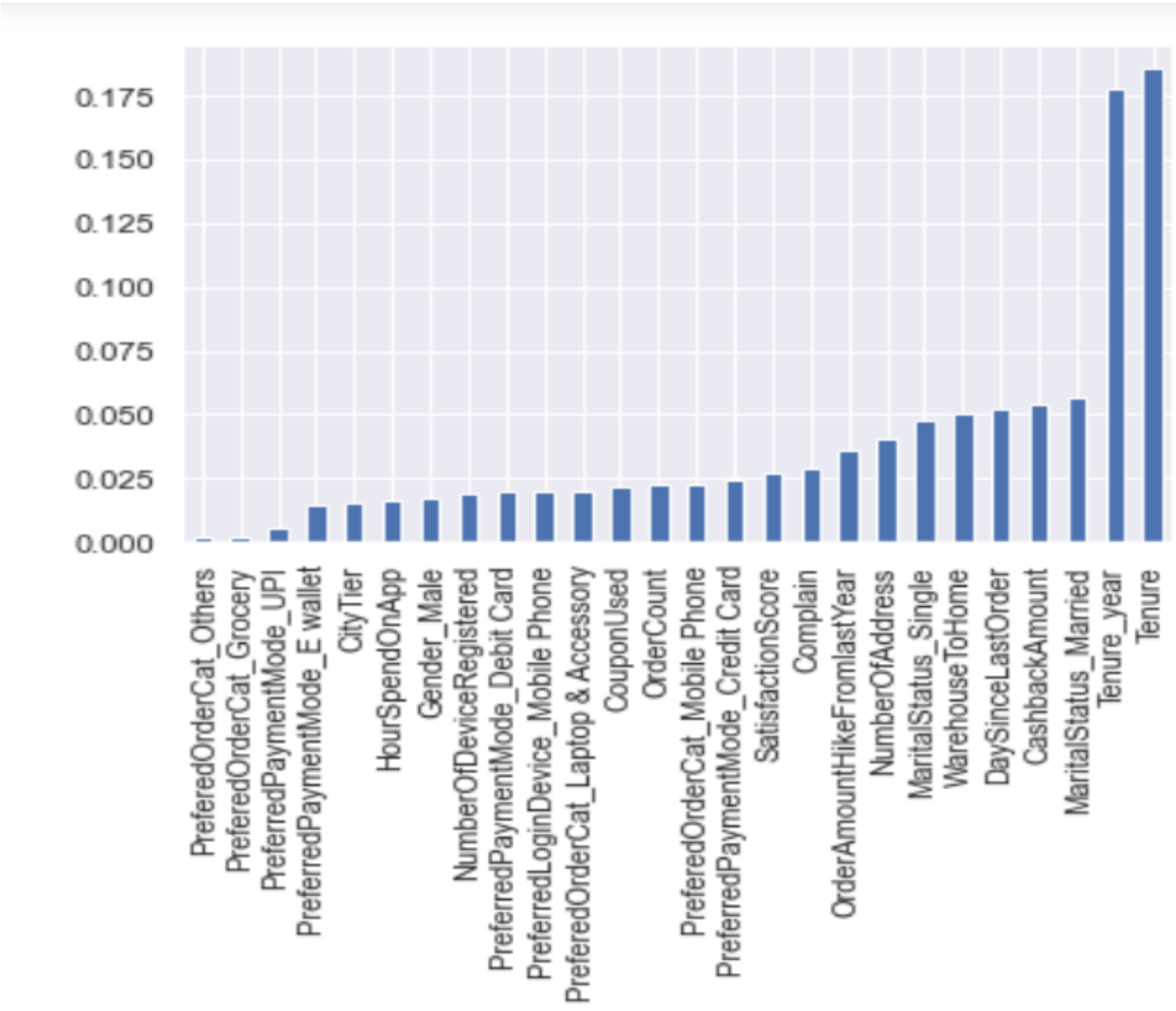


Fig 6.3 Feature Importance for Unscaled Data with RFC

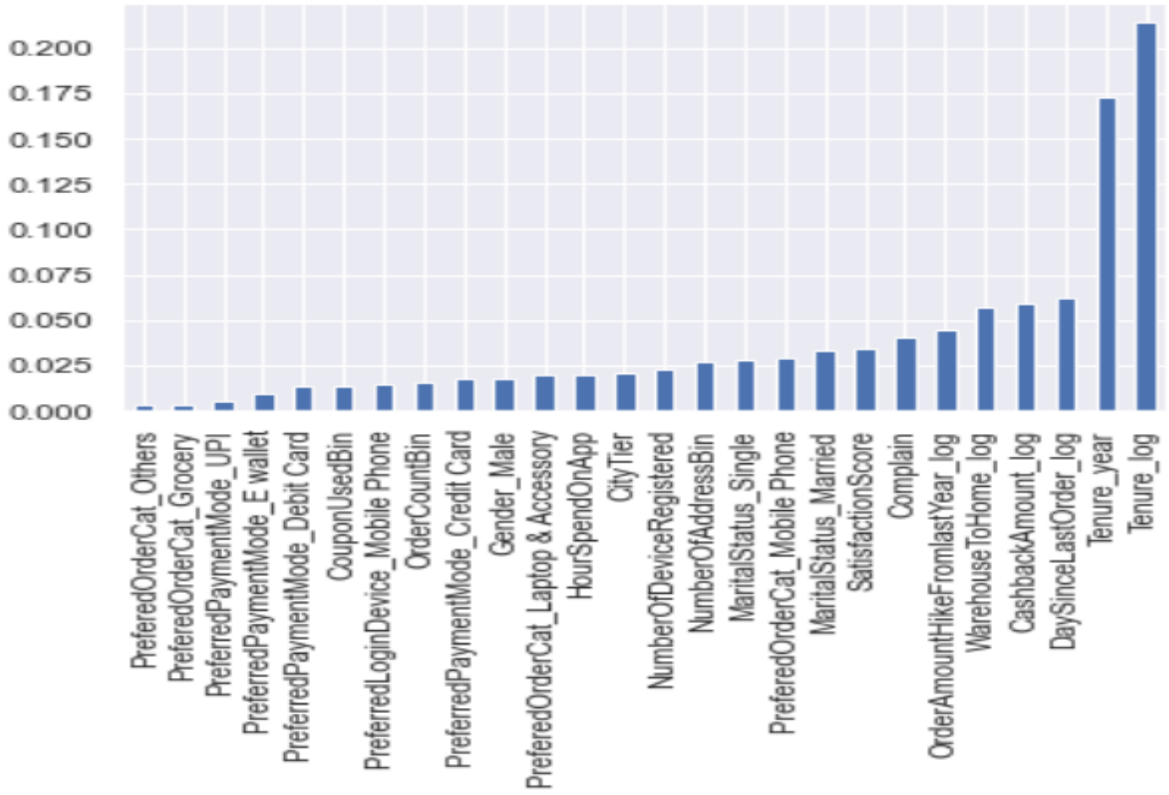


Fig 6.4 Feature Importance of Data with log transformation for RFC model

XGB with unscaled data can be used as the final model. Scaling is giving too high feature importance to one particular feature .

6.3 Model Tuning

However the models which are built can be fine tuned to improve the models performance.

Here we used RandomSearchCV for hyperparameter tuning.It searches space as a bounded domain of hyperparameter values and randomly sample points in that domain.

The F1 score after hyperparameter tuning is 0.975.There isn't much difference in the F1 score after hypertuning.

6.4 Model Deployment

We hosted the website using flask, which will take the features as user input ,do all preprocessing in the backend, predict using the XGB model and show the result. All data transformations and the selected model is saved using joblib. These are then loaded using joblib into app.py file .The user input fields are extracted from html page and converted to a Pandas DataFrame, applied transformations and predicted whether the customer will churn or not using the loaded XGB model.

6.4.1 Home Page

6.4.2 Prediction Page

7. Result

The final model we chose is XGB classifier with hyperparameter tuning and it gave an f1 score of 97.5%. This also means that the feature vectors available as per our dataset are more representative of the actual input space. We also analyzed the feature importance for the models with scores >90%. This can help the organizations to control their churn rate by taking preventive measures. Finally ,we hosted the website using flask, which will take the features as user input ,do all preprocessing in the backend, predict using the XGB model and show the result in an html page.

8. Conclusion

E-commerce businesses are allocating huge amount of money to acquire new customers. However, customers' lifetime depends on a lot of variables and this study was about building customer churn prediction model for e-commerce businesses. The dataset used for this project is for a leading e-commerce platform which was taken from Kaggle. The study started with exploratory analysis and data visualisations to increase our understanding to churned customers. Then,the data was cleaned for applying different machine algorithms to predict customer churn which are Logistic Regression, Random Forest, Decision tree,Extra Trees Classifier, Bagging Classifier, Gaussian Naive, SVM, Gradient Boosting Classifier, ADA boost classifier, XGB classifier .It was found that XGB has the best accuracy at 98% and is used to predict whether the customer churn or not.

•

References

1) Ecommerce Customer Churn Analysis.
https://www.analyticsvidhya.com/blog/2022/06/e-commerce-customer-churn-prediction/

2) XGboost https://www.nvidia.com/en-us/glossary/data-science/xgboost/

3) Math behind XG boost
https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

4) Parameter tuning in XGBoost.
https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

5) Data Transformation to better analyze data - Case study .
https://www.isixsigma.com/implementation/case-studies/data-transformations-helped-one-company-better-analyze-their-process-data/

6) Detecting and treating Outliers.
https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/

7)**Stephanie Glen**. "Winsorize: Definition, Examples in Easy Steps" From **tatisticsHowTo.com**: Elementary Statistics for the rest of us!
https://www.statisticshowto.com/

8) RobustScaler
.https://www.geeksforgeeks.org/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/

9) SMOTE using Python.
https://towardsdatascience.com/applying-smote-for-class-imbalance-with-just-a-few-lines-of-code-python-cdf603e58688