UNIVERSITY OF GHANA, LEGON
ACCRA, GHANA
DEPARTMENT OF COMPUTER SCIENCE
COLLEGE OF BASIC AND APPLIED SCIENCES
FINAL YEAR PROJECT REPORT

LIGHT GRADIENT BOOSTING MACHINE FOR E-COMMERCE CUSTOMER CHURN PREDCTION

PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE BACHELOR OF SCIENCE DEGREE IN COMPUTER SCIENCE

BY

VANESSA ATTA-FYNN - 10665130
SEPTEMBER, 2022

TABLES OF CONTENTS

DECLARATION

I declare that this is my original work with due referencing included and has never been submitted before to any institution for the award of Certificate, Degree or Diploma.

Name: Vanessa Atta-Fynn

Signature:

Date: September 29, 2022

This project has been submitted for examination purposes with the approval of the supervisor

Name: Solomon Mensah, PhD

Signature:

Date: September 29, 2022

ACKNOWLEDGEMENT

I would like to thank the God Almighty for the grace, speed and provision given to me to undertake this project despite the time and season. I would also like to acknowledge our supervisor, Dr Solomon Mensah, for the patience, support and guidance he gave to undertake this study.

ABSTRACT

Customer churn is what happens when the relationship of a customer with a company comes to the end. Customer churn is super expensive and impacts brand negatively. Losing customers is a serious problem that impacts all industries. Some of the most common reasons customer's churn are, incorrect pricing of products or services, inadequate understanding of customer needs and lack of brand loyalty. This study focuses on churn in the retail industry as it has one of the highest churn rates and traditional statistical analysis relies on high amount of assumption on customer data. In retail, this project will be narrowed further down to e-commerce as predicting churn in the offline setting will require more extensive data collection. In this project, machine learning algorithms will be used to predict customer churn in e-commerce. The studied dataset on customer churn was extracted from Kaggle. The studied dataset was segmented into groups or clusters using the K-means clustering algorithm. The Light Gradient Boosting Machine (LGBM), Support Vector Machine (SVM) and XGBoost algorithms are used for the training and validation of the segmented dataset and their evaluation performance compared to find the best fit for prediction. After series of empirical analysis, it was observed that the LGBM emerged as the best predictive model for customer churn.

Keywords: *Customer Churn prediction, E-commerce, Customer Segmentation, Machine learning.*

CHAPTER 1

INTRODUCTION

1.1 Background & Motivation

Customer churn or customer attrition is what happens when the relationship of a customer with a company or service comes to an end. Customer churn is a phenomenon that affects all kinds of industries and sectors of work from telecommunication, retail and e-commerce, finance to airline markets.

In the early days customer acquisition was the order of the day. Organizations focused market strategies and resources on acquiring new customers. However, in the current setting, due to the heavily saturated markets and cut-throat competition, the trends have changed toward customer retention.

The work of *Van Den Poel and Larivi_ere* clearly measures the priority of customer retention over customer acquisition in recent years with surveys [1]. The goal of businesses now is to keep existing customers happy, and ensure a positive and long-lasting relationship with customers, thus benefiting both company and customer.

Survey research carried out by Statista Research Department in 2020 on industry customer churn rates in the U.S. revealed general retail to have an astounding churn rate of 24% and online retail with a churn rate of 22%, as shown in Figure 1. These rates are only lagging slightly behind that of the financial and cable industries with a churn rate of 25% each. Though e-commerce and retail services records one of the highest churn rates as shown in Figure 1, not much attention is paid to customer churn on e-commerce websites [2].

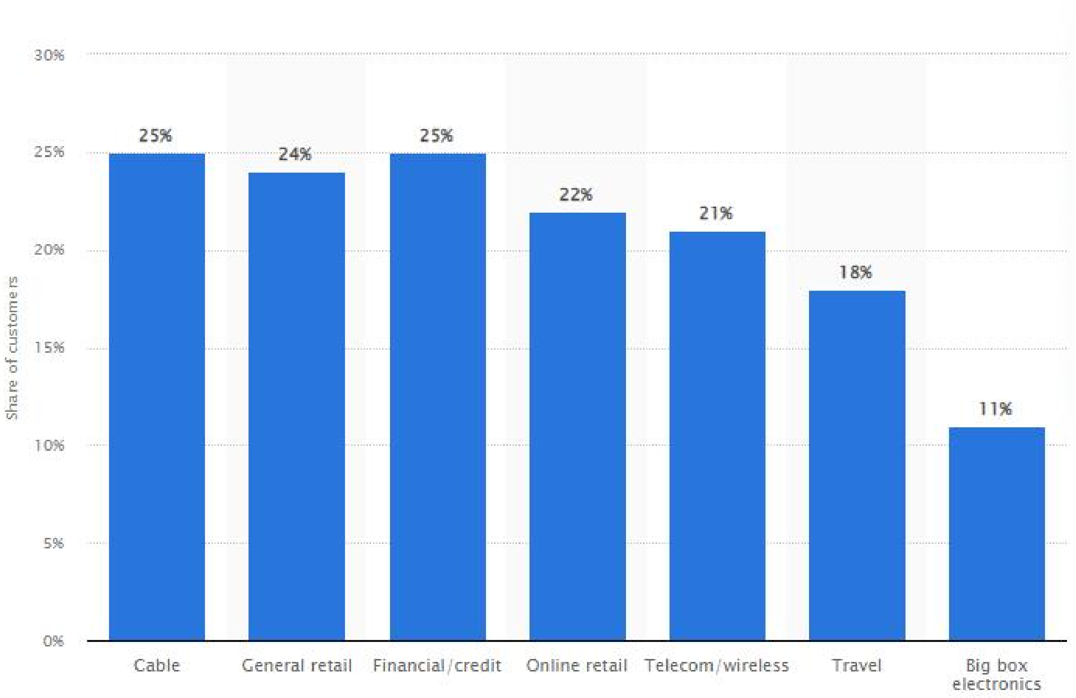


Figure 1: Customer Churn Rate by Industry U.S. in 2020

Customer churn prediction is one of the many methods developed to improve customer retention. It predicts the likelihood of customer churn and improves retention by honing in on likely abandoners. With this knowledge businesses are given a chance to take remedial actions to reduce or prevent churning.

This project seeks to illustrate how e-commerce churn prediction works with a suitable machine learning algorithm for customer churn prediction. It also seeks to propose an alternative algorithm for e-commerce customer churn prediction with better performance than others according to experimental results.

1.2 Problem Statement

Customer churn is very common but also very expensive for businesses. The lack of understanding of customers and their needs makes it difficult for them to maintain customer loyalty. Due to the non-contractual relationship in e-commerce, it is relatively difficult to predict customer churn and reasons behind it through traditional statistical analysis.

1.3 Research Objective

This study seeks to build and test algorithms for predicting customer churn in e-commerce sector by using a sample data set within this field.

1.4 Contribution of Study

At the time of publication, there was a noticeable research gap in the analysis and prediction of customer churn in the retail and e-commerce sector in Ghana. Most retail business owners and e-commerce stakeholders focus more on marketing strategies but are quite oblivious to this innovation. This paper serves as a stepping stone to breach that gap. This study is also geared towards contributing to the retail and e-commerce sector of the economy of the country by creating an alternative model to predict customer churn, thereby giving end-users the chance of making active decisions and remedial actions.

1.5 Scope of Study

The scope of this study is in e-commerce as a representation of online retail with notable references dating up to present time of publication.

1.6 Limitation of Study

This study focuses primarily on the e-commerce sector of the retail industry, as it is rather tedious to apply it to the offline world as this study is based on customer behavior on the web.

1.7 Organization of Study

This study will follow the following steps in the given order: the introduction which introduces the problem at hand and solution, review of essential literature, method and implementation, evaluation and results, a summary of accomplishments and conclusion, appendices, references and finally glossary.

CHAPTER 2

LITERATURE REVIEW

Customer churn basically refers to when customers of an enterprise seize to purchase goods and services of the enterprise and rather start to purchase that of the enterprise's competitors [3]. As stated earlier, it is noticed that research on customer churn seems to be prominent in the telecommunications sector. In e-commerce, customer churn specifically refers to when regular customers stop patronizing products of an e-commerce site usually based on certain common reasons like low product quality, frequent products shortages, delayed delivery, high cost of products etc. E-commerce customer churn is a type of customer churn that is based on no type of agreed customer relationship or is non contractual in nature. [4] also adds that "*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*". With this kind of customer relationship, it's difficult to detect and predict a customer churn in the future.

According to Alshamsi [5], customer churn prediction is important because it is used to record and merge data over a specific period of time and form e-commerce customer models through customer purchase behavior analysis.

As one of the world's oldest professions, the retail industry boasts of a variety of services and channels for selling and distributing goods to customers with the aim of making profit. The retail industry is a broad term for companies that sell goods and services to consumers. [6] writes that this consists of many kinds of retail sales and stores around the world such as grocery stores, department stores, convenience stores, electrical stores and many more. Some examples of retail companies are Amazon, Shop rite, Game, Walmart, Canadian tire and many more.

In the early days, the concept of retail was limited to selling goods and rendering services from a physical store. Over the past two decades, retail has seen some exponential growth and transformation since the introduction of the internet in the early 1980s even though in the 1960s companies used to conduct business with computer networks according to [6]. The vast adoption of online retail started in the 1990s. This industry has seen extreme growth for the past years with the recent advancements in technology especially in the years of COVID-19.[7]

Online retail or e-commerce has become the biggest shift in business worldwide. Customer churn is a major blind spot of most companies around the world and it's really affecting businesses. Most companies don't see the need to analyze and predict customer churn which could collapse their businesses. Companies wait till their establishments are on the brink of collapsing before they employ specialist to handle matters in regards to the customer churn their experiencing.

Having a customer churn model for your company will give them the ability to identify any change in customer behavior with regards to purchasing of goods and services and also helps you improve your customer relations exceptionally.

According to [6] electronic commerce or e-commerce is the buying and selling of goods and services over the internet. E-commerce can take place in so many ways, such as ordering goods, purchasing a service, buying a subscription and many more. The forecasting methods of customer churn can be summarized into three types:
Forecasting methods based on traditional statistical analysis, Prediction methods based on machine learning, and Prediction methods based on combinatorial classifiers according to [8].

[3] goes ahead to state that traditional statistical analysis has been used for centuries even before the introduction of computers. It has relied on minimum samples of data and makes prior assumptions about data. Therefore, traditional demand forecasting statistical analysis is preset with the "*self-evident proposition*" that past demand periods are a good predictor of demand in the years to come.

Machine learning is also a discipline of artificial intelligence that provides machines with the ability to automatically learn from data and past experiences while identifying patterns to make predictions with minimal human intervention. Machine Learning methods give computers the ability to work autonomously without any emphatic programming. Since machine learning applications are continuously fed with new data, they are able to independently learn, grow, develop and adapt according to [14].

[9] then compared machine learning to traditional statistics analysis and found that machine learning techniques and approach heavily relies on computing power but the traditional statistics analysis techniques do not require any thing of such nature. [9] goes on to state that "*Traditional statistics analysis heavily relies on small samples and heavy assumptions about data and its distributions when machine learning techniques tend to make less pre-assumptions about the problem and is liberal in its approaches and techniques to find a solution, many times using heuristics*".

It's hard to imagine life without e-commerce these days. It would not only be inconvenient but also much more complex. E-commerce is completely part of our lives now one way or another, it wasn't long ago when it was not even in existence. The origins of E-commerce go all the way back to about 40 years ago when "*teleshopping*" first came out as the father to the modern version. As we all know e-commerce got its real start when retail giant Amazon launched one of the first e-commerce websites back in the early 1990s. Since the inception of Amazon's website numerous companies have followed suit [3].

Today, online-retail and e-commerce are extremely popular because of the convenience of use in many aspects as compared to retail in physical stores for buyers and sellers alike. For one reason, e-commerce stores provide sellers a cost-effective way to advertise and sell products to a wide range of customers without the expenses of owning or renting a physical store. Similarly, the ease of use of e-commerce websites and stores gives consumers a wide range and variety of products from all around the world. With only a few clicks customers can easily pick and choose products and services with desired requirements in their comfort zones without much effort.

Due to the rapid increased in patronage of the e-commerce sector from all over the world it is very vital for companies in the e-commerce sector to be able to identify the high valued customers who are about to churn. It is also important for them to monitor purchasing habits or behaviors of customers who have not churned to be able to keep this group of customers.

This will make available e-commerce customer churn retention strategies to optimize customer churn and distinguish high value non-churn e-commerce customers and do an exceptional job of customer retention.

Based on research, [10] states that it is very important to analyze the loss of customers, be able to have a prediction of customers who might be lost and then take correlating assessments to keep these customers and avoid losing them.

Currently companies are using data mining technologies to conduct in-depth analysis of customer behaviors and customer transaction traits data. They do this by using these technologies to establish and learn customer purchase behavior and other metrics to predict customer churn in their companies. These technologies are widely used in customer relationship management of e-commerce companies such as customer segmentation, fraud analysis and customer churn prediction.

The prediction methods based on traditional statistical analysis mainly include linear discriminant analysis, the naive Bayesian model, cluster analysis, and logistic regression. For example, Pınar et al. used a naive Bayes classifier to predict customer churn of a telecom company in 2011. [8]. Based on data collected and analysis, their results showed that the average call duration of customers was strongly correlated with customer churn.

Customer segmentation is basically the act of separating a company's customers into groups that exhibit similarity among customers in each group. The main aim of segmenting customers is to determine how to relate to customers in each group. This highlights the importance of customer relationship, which is essential for well-structured marketing activities. In reference to the Pareto principle, 80% of a company's profits are created by 20% of its customers and 50% of its profits are lost by the bottom 30% of non-profit customers [11].

The customer segmentation process requires the collection of data such as transactional customer data which is made up of their static (E.g., Age, Gender etc.) dynamic data (E.g., Purchase frequency etc.) from vendors according to [12].
The collected data then goes through a filtering process called the "*pre-processing of data*". It is through this process that the relevant data gets extracted from the collected data. Feature selection is also a data reduction technique used to extract relevant features needed for the input vector of a predictive model. These are the pre-processing steps for establishing a subset of original features by exempting features that are irrelevant. [12]

With clustering, features are not segmented with regard to their effect on a specific target variable. Rather, it is done without a fixed aim as it identifies patterns in the already existing data sets. There are a number of algorithms that can be used to perform clustering. After segmentation, by the use of the clustering method they can be further analyzed and made available for campaigns [12].

"*K-means is one of the most used clustering algorithms and it's easy and efficient to use. The K-means supports the partitioning of the 'n' number of observations into a named number of 'k' clusters. The K-means is an unsupervised learning algorithm and one of the simplest algorithms used for clustering tasks*" according to [12].

[8] predicted customer churn with an SVM classifier in combination with customer segmentation using K-means clustering algorithm. The k-means algorithm was first used to group customers into three subdivisions. Next prediction was made using the SVM and logistic regression models for these models and results were compared. The results for the two showed significant improvement in prediction index due to the

customer segmentation performed and also, showed the SVM model performed better at predicting e-commerce churn prediction as compared to the Logistic Regression model.[8]

Similarly, [13] uses a Recency Frequency Monetary (RFM) model and k-means clustering algorithm to segment customers and then uses a XGBoost model for prediction. After comparative analysis with other models, the XGBoost was found to be the most highly performing and prediction was more accurate for all models after customer segmentation was performed.

The best and most efficient way to build a customer churn prediction model is by the use of Machine Learning applications. Machine Learning applications have the capacity to handle and analyze great amounts of data at the same time. This leads to the most precise and comprehensive results based on [4]'s research.

The first thing you need to do when building any model is to collect the right data. To build a perfect customer churn prediction model for a company you need to get more data about their customers. This is what will lead to more accuracy in prediction. The data for the customer churn prediction model should include at least customer data, product data and purchase history. For example, customer id, location, purchase frequency, product types, recent purchase record, marketing emails opened, coupons used [4].

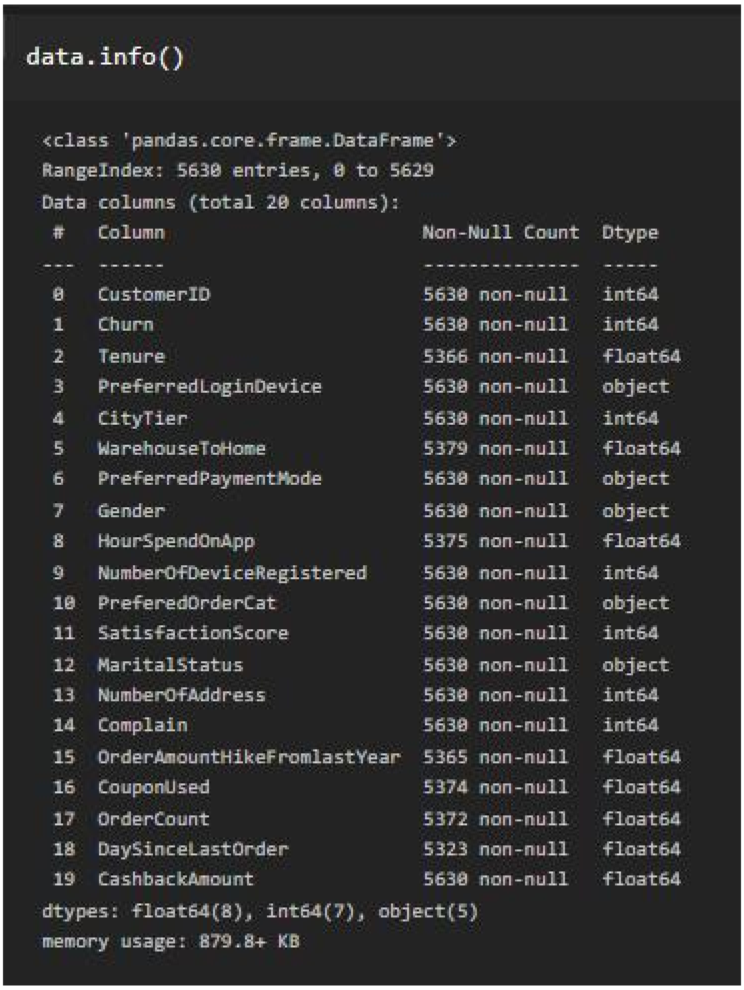
In conclusion, it has been found that there are a number of customer churn prediction models and techniques. Customer segmentation with the use of machine learning applications turns out to be the most effective technique per my research.After collecting the right data, the a machine learning model segments the data into groups with regards to their relevant similarities. The machine learning model finally is able to predict customer churn when all these important steps have been achieved. According to research the use of machine learning applications to predict customer churn is the fastest and most effective method and that is the method I have employed for this project.

CHAPTER 3

METHODOLOGY

3.1 Dataset

The data collected for this project is an e-commerce data set of historical data for an online retail store from Kaggle. The data set consist of a range of customer data collected as well as a churn flag showing which customers churned and which customers did not churn. It consists of a total of 20 columns and 5630 rows.



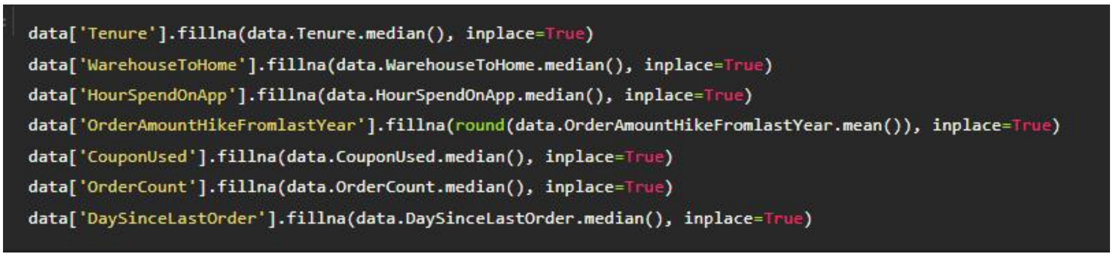
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5630 entries, 0 to 5629
Data columns (total 20 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   CustomerID                   5630 non-null   int64
 1   Churn                        5630 non-null   int64
 2   Tenure                       5366 non-null   float64
 3   PreferredLoginDevice         5630 non-null   object
 4   CityTier                     5630 non-null   int64
 5   WarehouseToHome              5379 non-null   float64
 6   PreferredPaymentMode         5630 non-null   object
 7   Gender                       5630 non-null   object
 8   HourSpendOnApp               5375 non-null   float64
 9   NumberOfDeviceRegistered     5630 non-null   int64
 10  PreferedOrderCat             5630 non-null   object
 11  SatisfactionScore            5630 non-null   int64
 12  MaritalStatus                5630 non-null   object
 13  NumberOfAddress              5630 non-null   int64
 14  Complain                     5630 non-null   int64
 15  OrderAmountHikeFromlastYear  5365 non-null   float64
 16  CouponUsed                   5374 non-null   float64
 17  OrderCount                   5372 non-null   float64
 18  DaySinceLastOrder            5323 non-null   float64
 19  CashbackAmount               5630 non-null   float64
dtypes: float64(8), int64(7), object(5)
memory usage: 879.8+ KB
```

3.2 Data Pre-processing & Feature Engineering

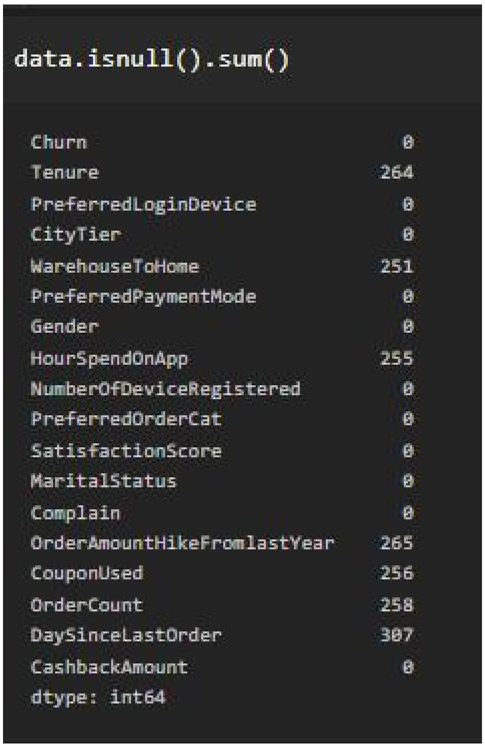
The discussed e-commerce data set was found to have a series of problems that were not ideal for our prediction later.

First of, the problem of missing values in column was identified and handled.



```
data['Tenure'].fillna(data.Tenure.median(), inplace=True)
data['WarehouseToHome'].fillna(data.WarehouseToHome.median(), inplace=True)
data['HourSpendOnApp'].fillna(data.HourSpendOnApp.median(), inplace=True)
data['OrderAmountHikeFromlastYear'].fillna(round(data.OrderAmountHikeFromlastYear.mean()), inplace=True)
data['CouponUsed'].fillna(data.CouponUsed.median(), inplace=True)
data['OrderCount'].fillna(data.OrderCount.median(), inplace=True)
data['DaySinceLastOrder'].fillna(data.DaySinceLastOrder.median(), inplace=True)
```

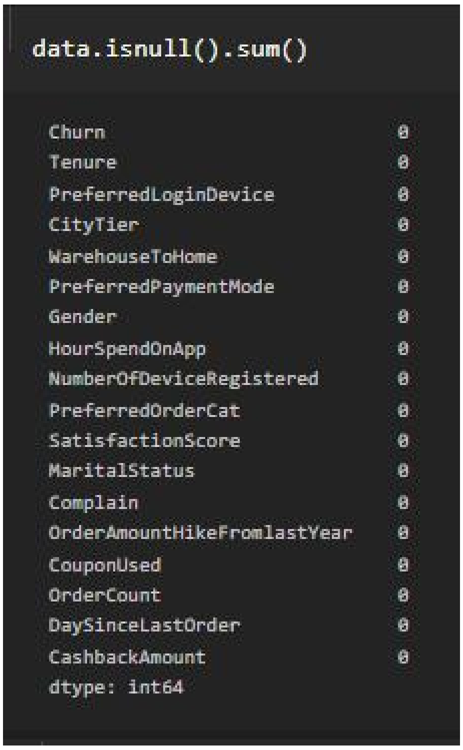
Original



```
data.isnull().sum()

Churn                            0
Tenure                         264
PreferredLoginDevice             0
CityTier                         0
WarehouseToHome                251
PreferredPaymentMode             0
Gender                           0
HourSpendOnApp                 255
NumberOfDeviceRegistered         0
PreferedOrderCat                 0
SatisfactionScore                0
MaritalStatus                    0
Complain                         0
OrderAmountHikeFromlastYear    265
CouponUsed                     256
OrderCount                     258
DaySinceLastOrder              307
CashbackAmount                   0
dtype: int64
```

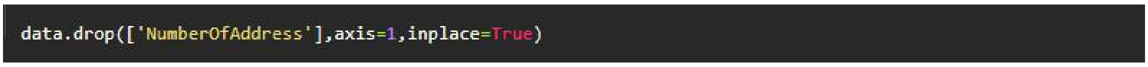
Fixed



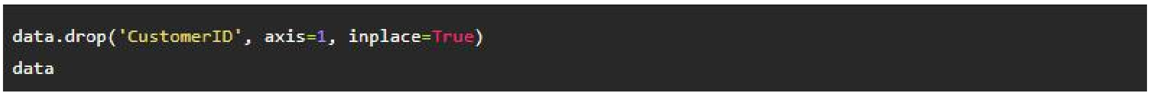
```
data.isnull().sum()

Churn                          0
Tenure                         0
PreferredLoginDevice           0
CityTier                       0
WarehouseToHome                0
PreferredPaymentMode           0
Gender                         0
HourSpendOnApp                 0
NumberOfDeviceRegistered       0
PreferedOrderCat               0
SatisfactionScore              0
MaritalStatus                  0
Complain                       0
OrderAmountHikeFromlastYear    0
CouponUsed                     0
OrderCount                     0
DaySinceLastOrder              0
CashbackAmount                 0
dtype: int64
```

Next, the data set was observed to have some irrelevant columns that would not play any role in bringing insights, hence these columns were drops.



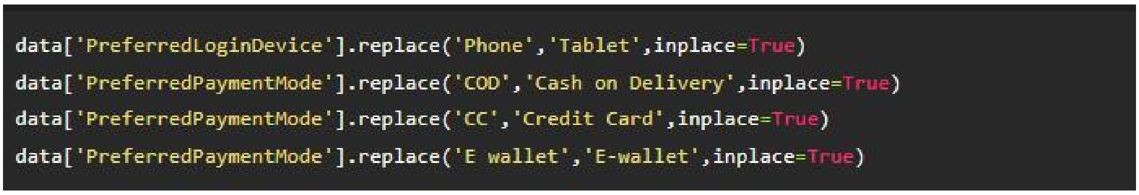
```
data.drop(['NumberOfAddress'],axis=1,inplace=True)
```



```
data.drop('CustomerID', axis=1, inplace=True)
data
```

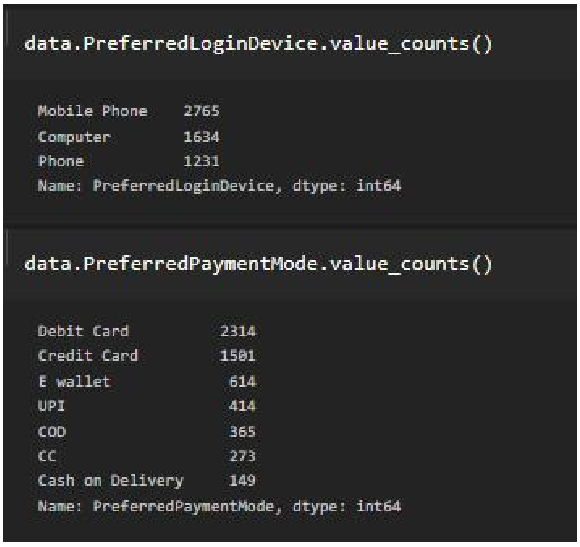
Next after observing the certain columns it was observed that there were variations of the same value such as in the *PreferredPaymentMethod* column, there was *Cash on*

Delivery and *COD* which represent the same thing. Columns with such issues were appropriately fixed.



```
data['PreferredLoginDevice'].replace('Phone','Tablet',inplace=True)
data['PreferredPaymentMode'].replace('COD','Cash on Delivery',inplace=True)
data['PreferredPaymentMode'].replace('CC','Credit Card',inplace=True)
data['PreferredPaymentMode'].replace('E wallet','E-wallet',inplace=True)
```

Original



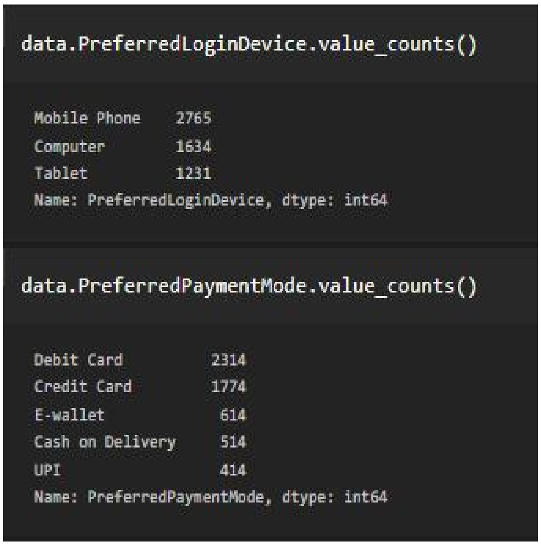
```
data.PreferredLoginDevice.value_counts()

Mobile Phone    2765
Computer        1634
Phone           1231
Name: PreferredLoginDevice, dtype: int64
```

```
data.PreferredPaymentMode.value_counts()

Debit Card          2314
Credit Card         1501
E wallet             614
UPI                  414
COD                  365
CC                   273
Cash on Delivery     149
Name: PreferredPaymentMode, dtype: int64
```

Fixed



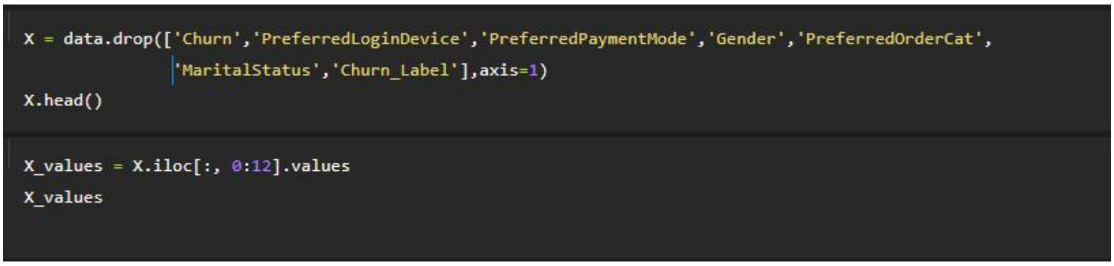
```
data.PreferredLoginDevice.value_counts()

Mobile Phone    2765
Computer        1634
Tablet          1231
Name: PreferredLoginDevice, dtype: int64
```

```
data.PreferredPaymentMode.value_counts()

Debit Card          2314
Credit Card         1774
E-wallet             614
Cash on Delivery     514
UPI                  414
Name: PreferredPaymentMode, dtype: int64
```

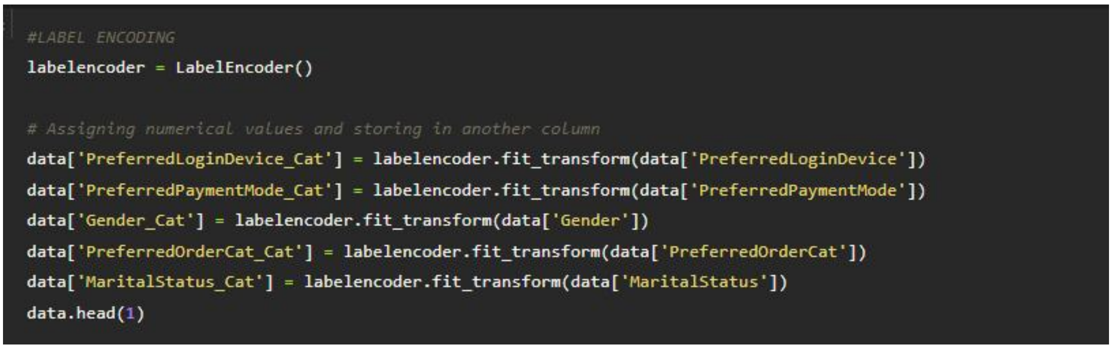
Coming to our customer segmentation, the K-means algorithm works with only numeric data. Hence, a new pandas dataframe was created containing only our numeric columns.



```
X = data.drop(['Churn','PreferredLoginDevice','PreferredPaymentMode','Gender','PreferredOrderCat',
               'MaritalStatus','Churn_Label'],axis=1)
X.head()
```

```
X_values = X.iloc[:, 0:12].values
X_values
```

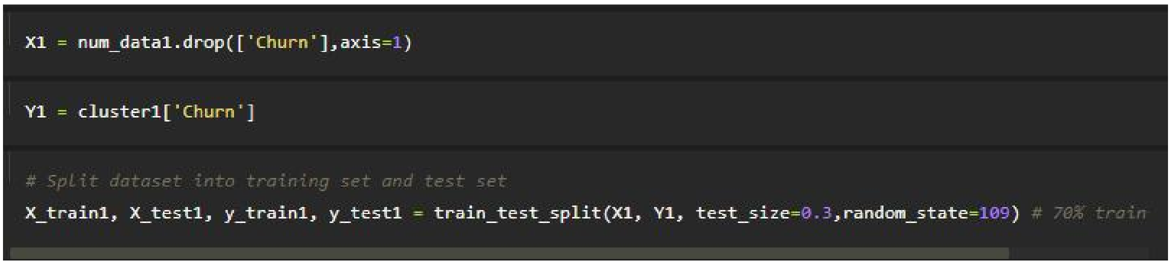
Finally, for our classification, the models worked well with numeric column hence our non-numeric categorical column were encoded using sci-kit learn's label encoder and stored in new columns.



```
#LABEL ENCODING
labelencoder = LabelEncoder()

# Assigning numerical values and storing in another column
data['PreferredLoginDevice_Cat'] = labelencoder.fit_transform(data['PreferredLoginDevice'])
data['PreferredPaymentMode_Cat'] = labelencoder.fit_transform(data['PreferredPaymentMode'])
data['Gender_Cat'] = labelencoder.fit_transform(data['Gender'])
data['PreferredOrderCat_Cat'] = labelencoder.fit_transform(data['PreferredOrderCat'])
data['MaritalStatus_Cat'] = labelencoder.fit_transform(data['MaritalStatus'])
data.head(1)
```

The resulting dataset was then separated based on their clusters and within each cluster, the data was split into train and test with sci-kit learn’s train_test_split function with a proportion of 0.7 train data and 0.3 test data.



```
X1 = num_data1.drop(['Churn'],axis=1)
```

```
Y1 = cluster1['Churn']
```

```
# Split dataset into training set and test set
X_train1, X_test1, y_train1, y_test1 = train_test_split(X1, Y1, test_size=0.3,random_state=109) # 70% train
```

3.3 Models Used

For customer segmentation, the K-mean clustering algorithm was implemented to create our customer groups. The following models were then compared to find the best fit for this study:

i. Support Vector Machine (SVM)

ii. Extreme Gradient Boosting Trees (XGBoost)

iii. Light Gradient Boosted Machine Algorithm (lightgbm) - proposed model

3.4 Data Analysis

As inspired by [4] a data set with adequate customer details was collected for this study. This data set had the following features:

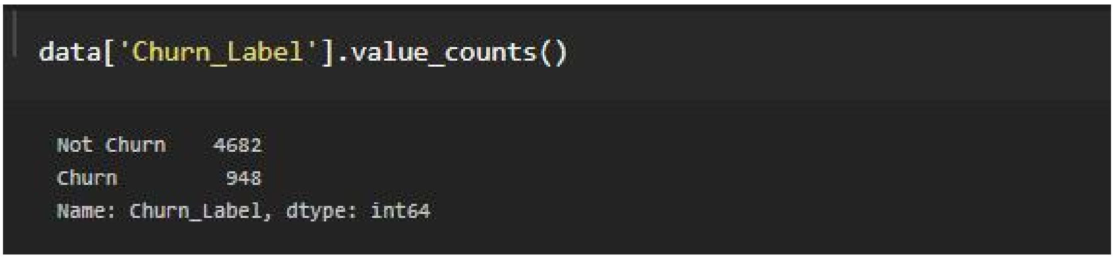
Customer ID	Unique customer ID
Tenure	Tenure of customer in organization
PreferredLoginDevice	Preferred login device of customer

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
PreferedOrderCart	Preferred order category of customer in last month
SatisfactionScore	Satisfactory score of customers on service
MaritalStatus	Marital status of customer
NumberOfAddress	Total number of 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 coupons 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
Churn	Churn Flag

Exploratory Data Analysis and Results

Exploratory data analysis revealed a class imbalance within the Churn (target) column of the dataset. Out of the 5630 records of this dataset, there are 4682 Not Churn records as against 948 Churn records. This class imbalance will need to be resolved as it may after the models ability to predict Churn customers.

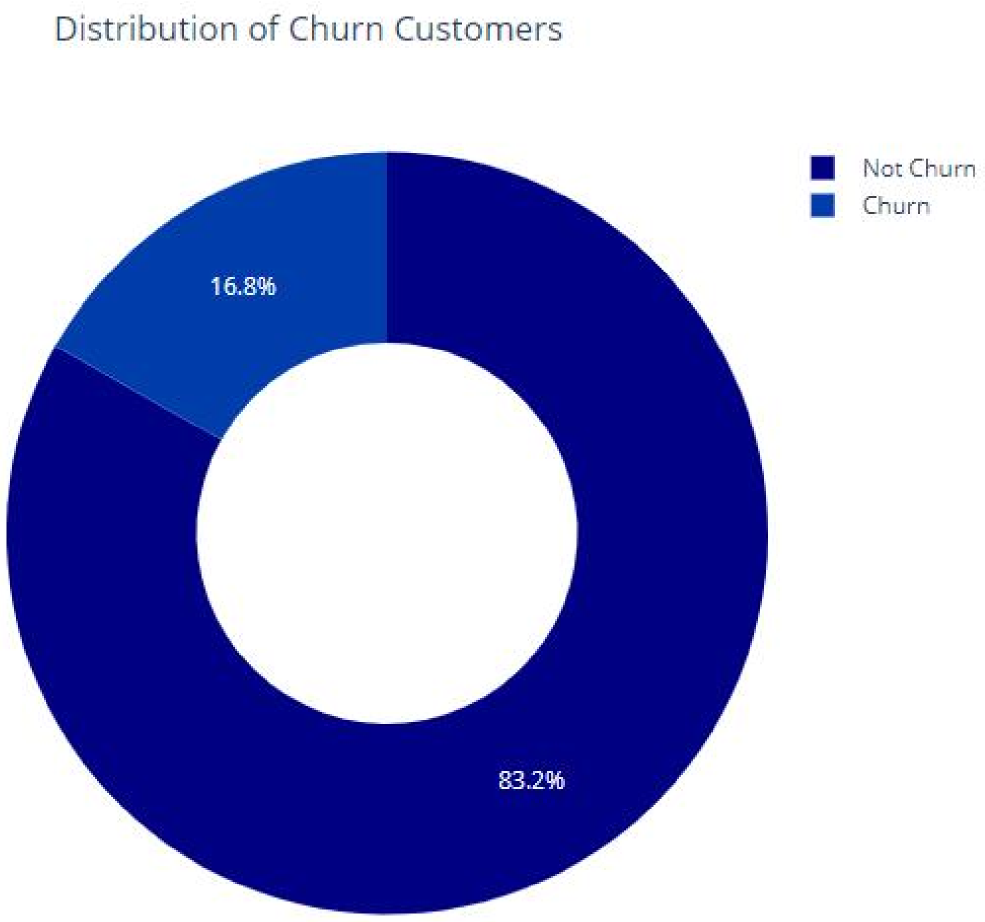
Uni-variate Analysis



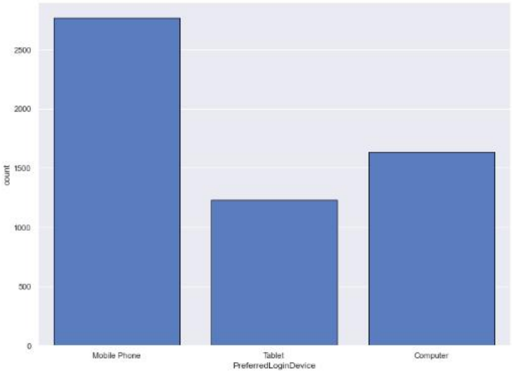
```
data['Churn_Label'].value_counts()
```

```
Not Churn    4682
Churn         948
Name: Churn_Label, dtype: int64
```

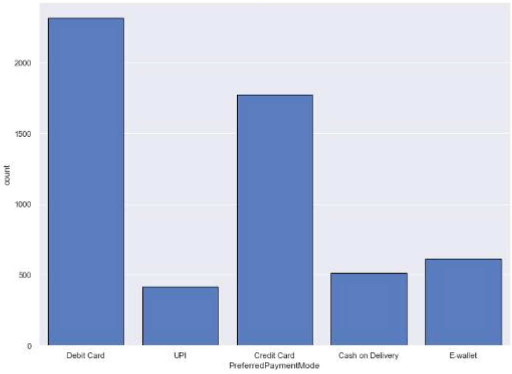
Churn customers make up 16.8% of that of the data set as observed in the pie chart of the Distribution of Churn Customers.



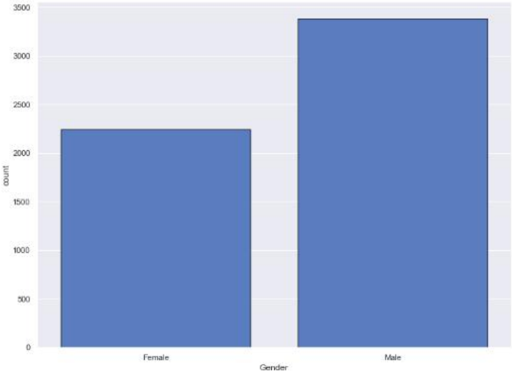
Preferred Login Device



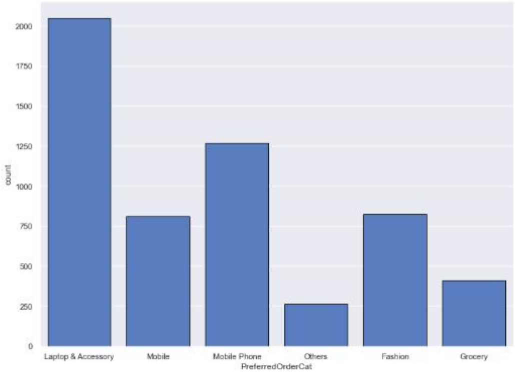
Preferred Payment Method



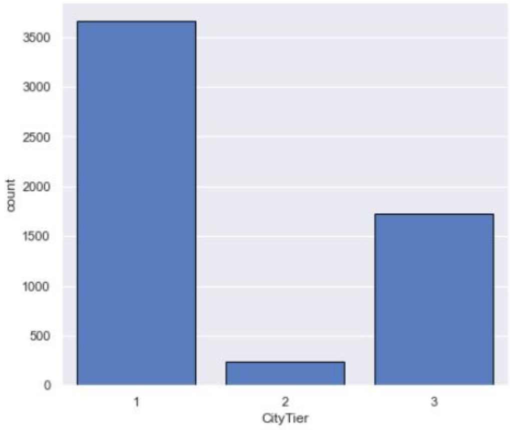
Gender



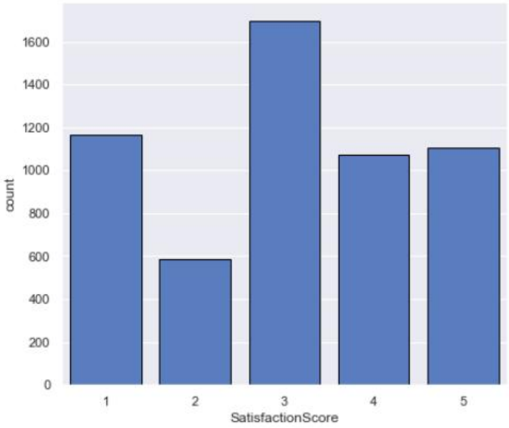
Preferred Order Category



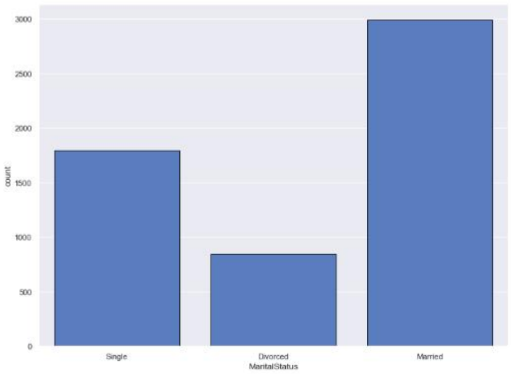
City Tier



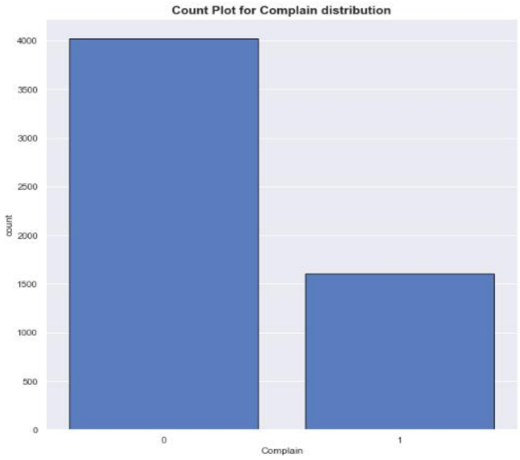
Satisfaction Score



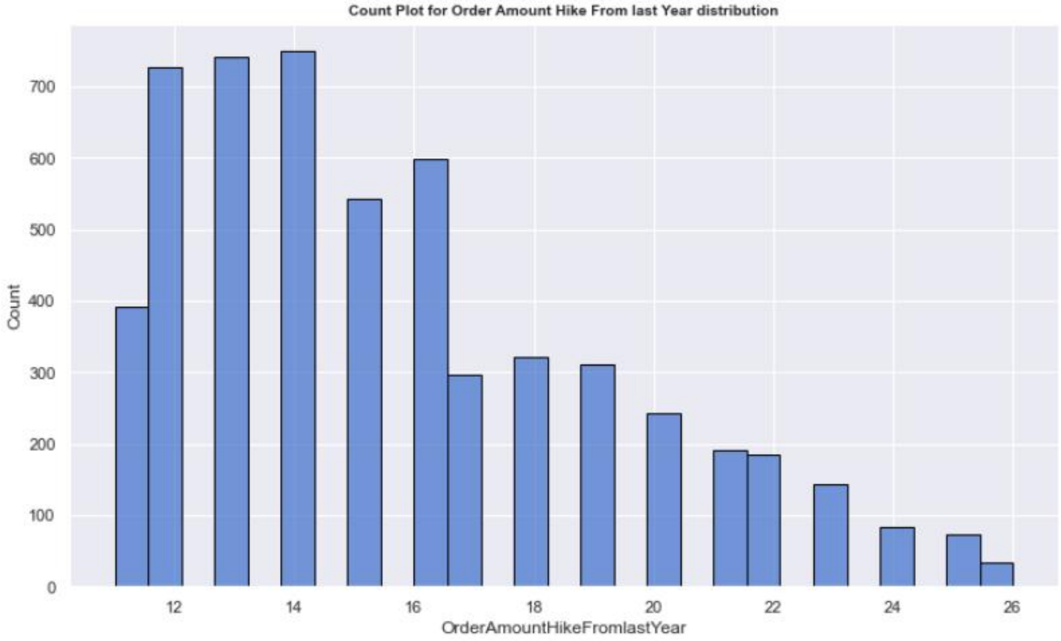
Marital Status



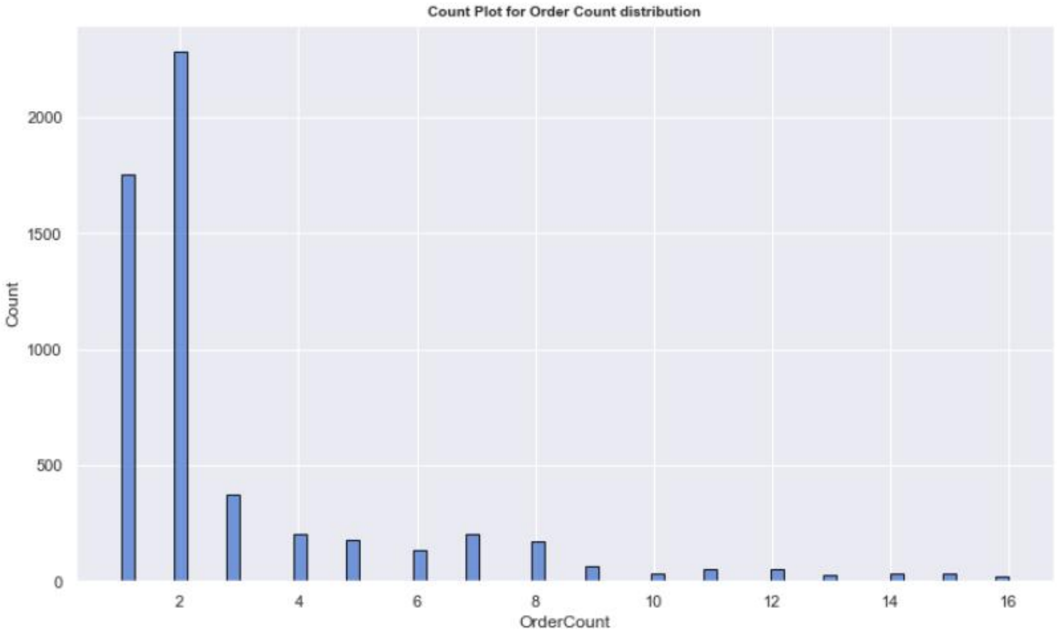
Complain



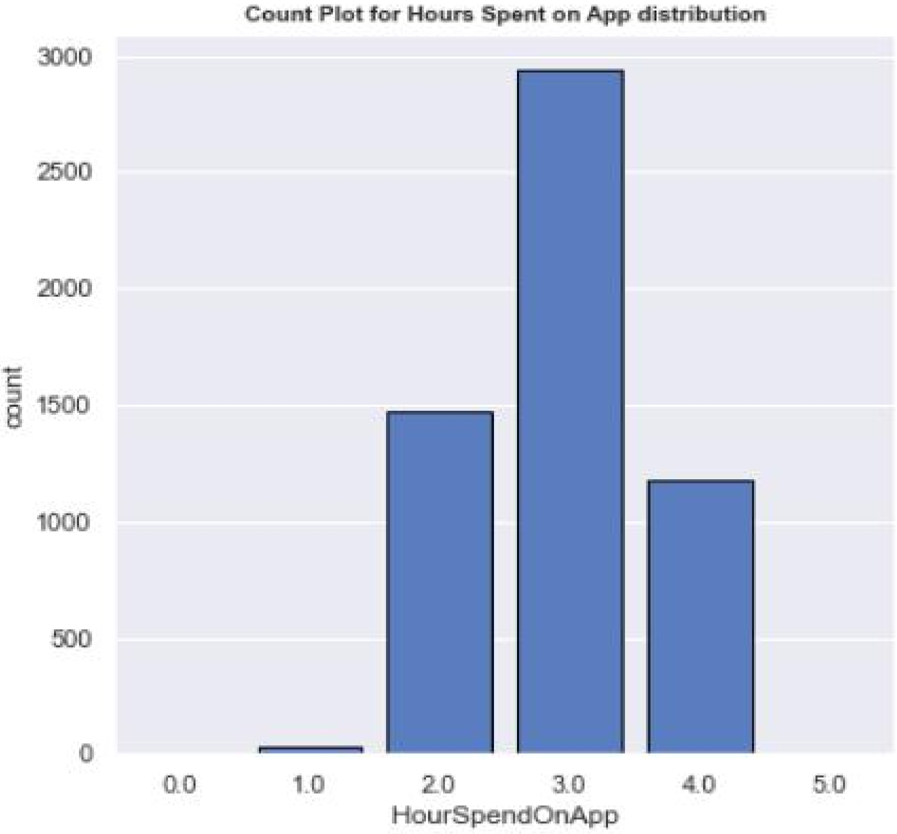
Count plot distribution for Order Amount Hike from Last Year



Count plot distribution for Order Count

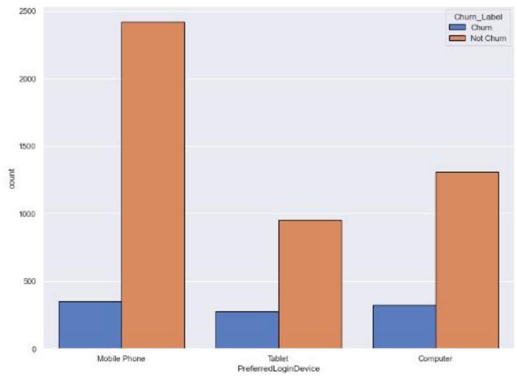


Count plot distribution for Hours Spent on App

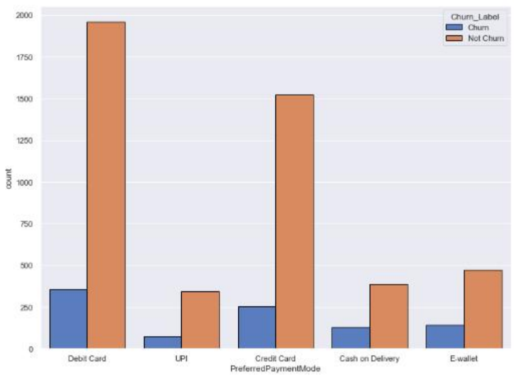


Bi-variate Analysis

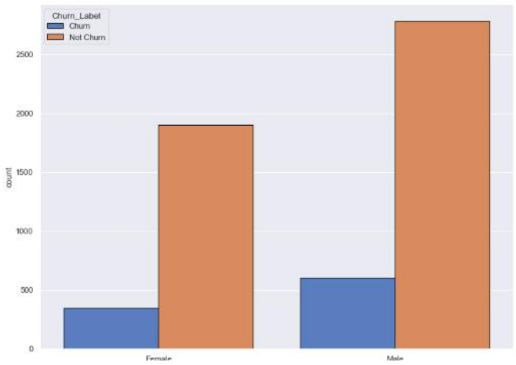
Preferred Login Device vs Churn



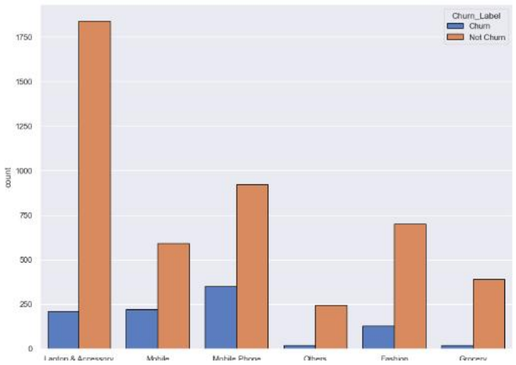
Preferred Payment Method vs Churn



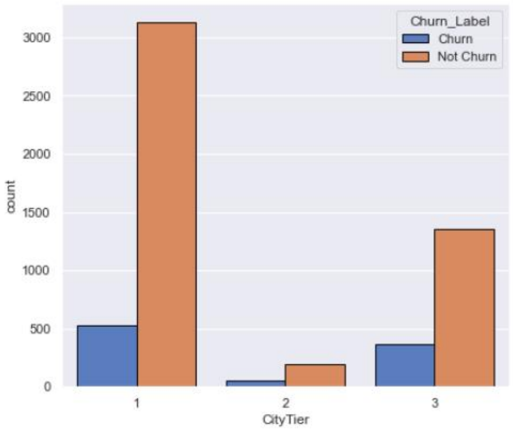
Preferred Login Device vs Churn



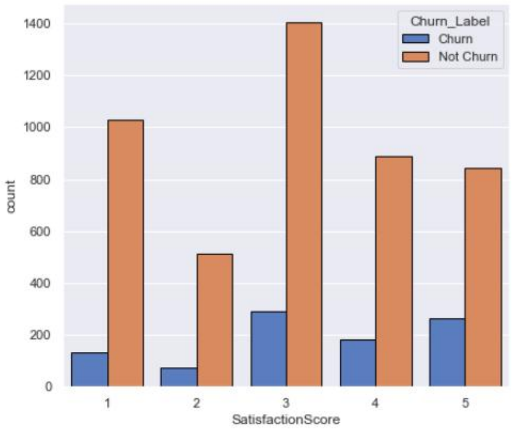
Preferred Payment Method vs Churn



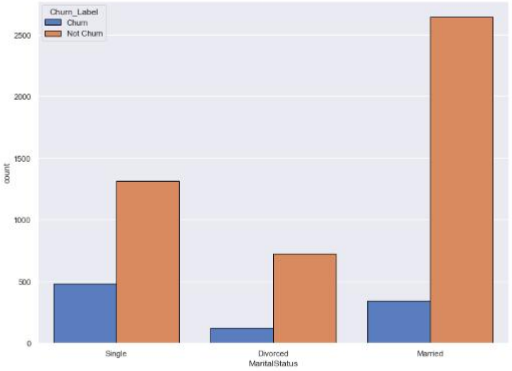
City Tier vs Churn



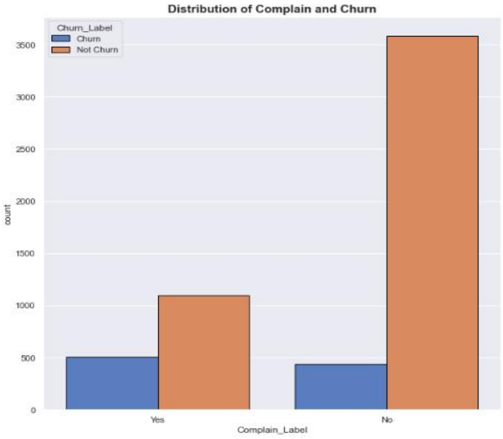
Satisfaction Score vs Churn



Marital Status vs Churn

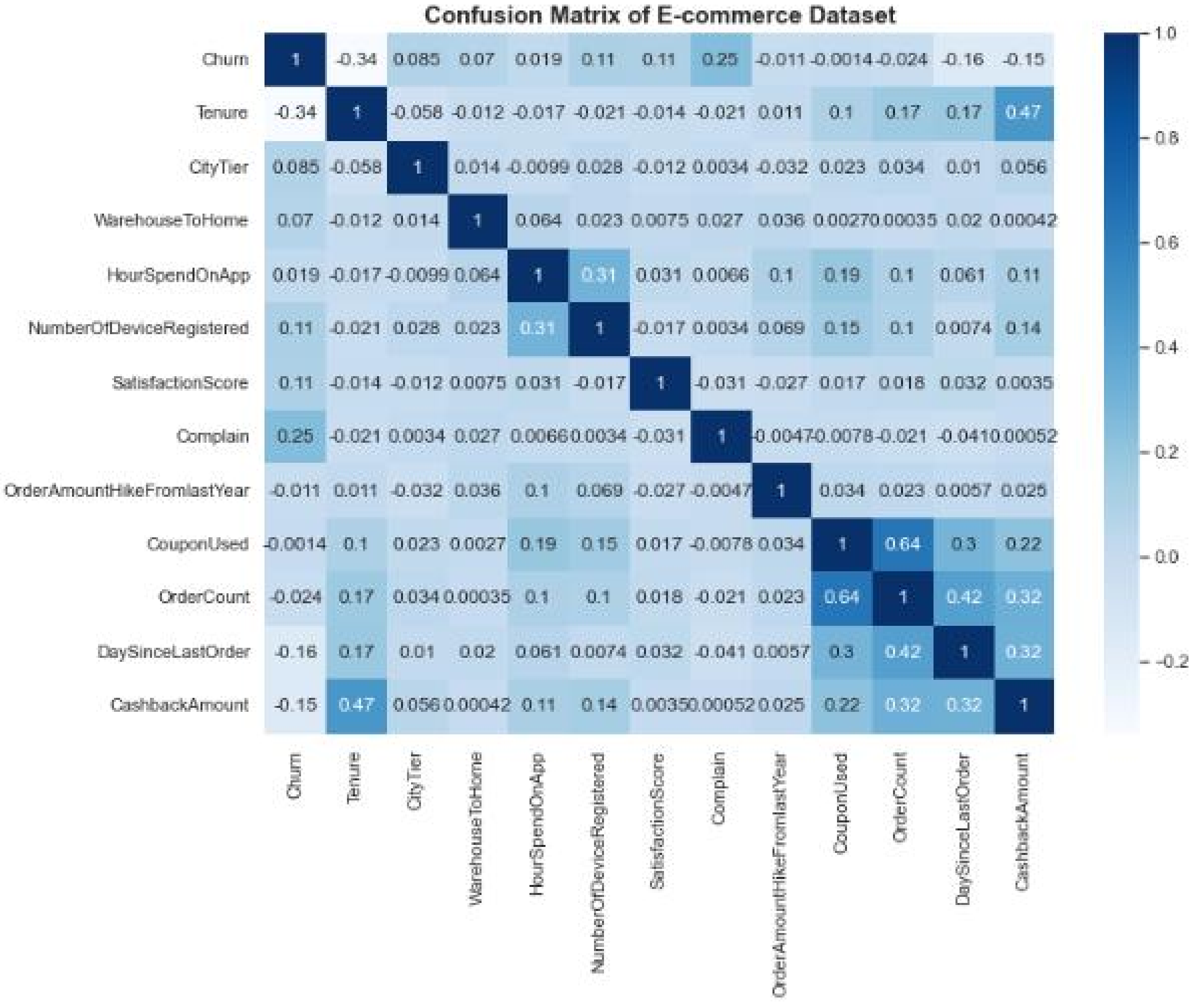


Complain vs Churn

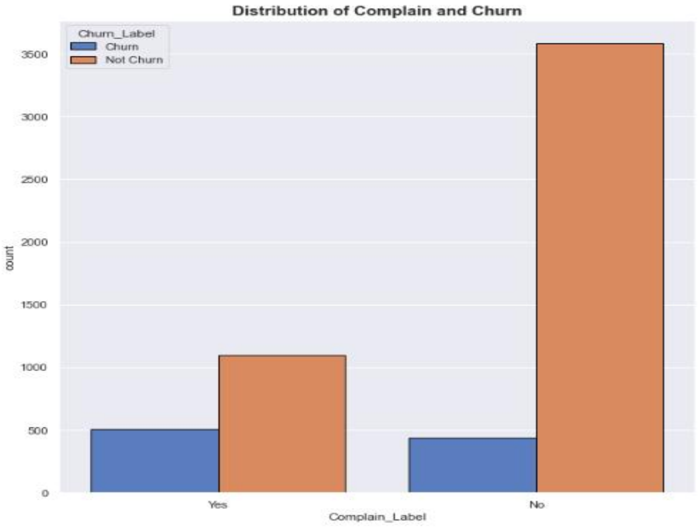


Multivariate Analysis

With the help of a confusion matrix, it was revealed that the target column Churn is has a positive and its highest correlation with the Complain column. Those the correlation is not very strong, it shows a positive correlation implying an increase in the number of churn customers with an increase the number of customers have have made a complaint.



We can see the relationship of Churn and Complaint better with the help of this grouped bar chart.



3.5 Experimental Setup

• Pandas – for dataset manipulations.

• Numpy - numeric manipulations

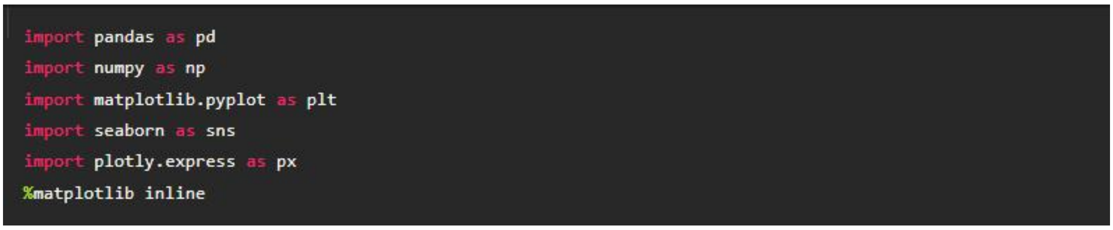
• Sklearn - model building

• Mathplotlib – for charts and graph visualizations

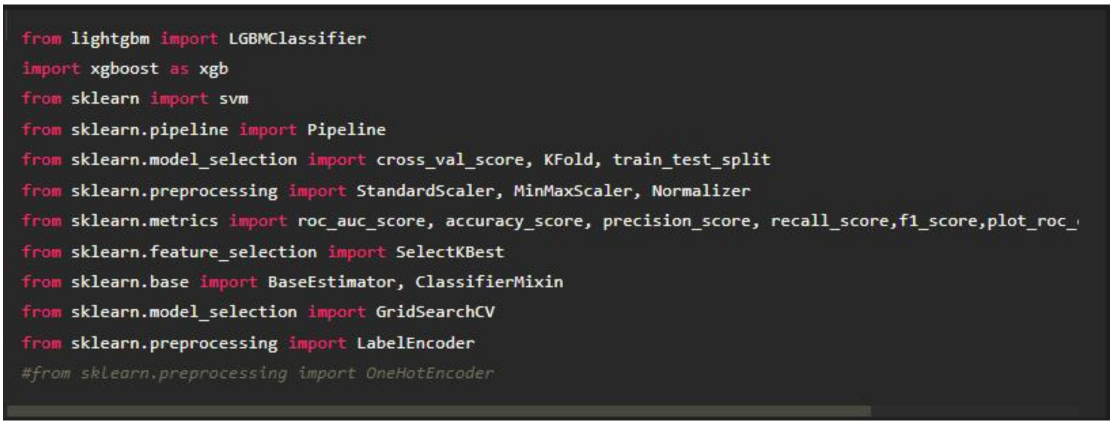
• Seaborn – for charts and graph visualizations

• Plotly – for charts and graph visualizations

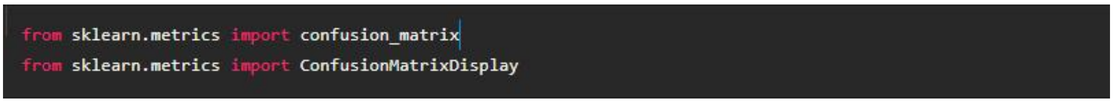
Library Imports



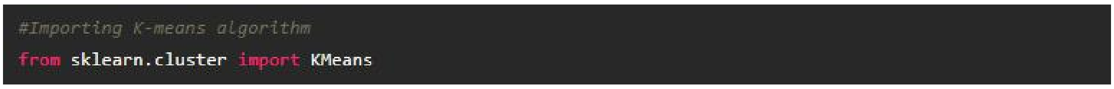
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
%matplotlib inline
```



```
from lightgbm import LGBMClassifier
import xgboost as xgb
from sklearn import svm
from sklearn.pipeline import Pipeline
from sklearn.model_selection import cross_val_score, KFold, train_test_split
from sklearn.preprocessing import StandardScaler, MinMaxScaler, Normalizer
from sklearn.metrics import roc_auc_score, accuracy_score, precision_score, recall_score,f1_score,plot_roc_
from sklearn.feature_selection import SelectKBest
from sklearn.base import BaseEstimator, ClassifierMixin
from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import LabelEncoder
#from sklearn.preprocessing import OneHotEncoder
```



```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import ConfusionMatrixDisplay
```



```
#Importing K-means algorithm
from sklearn.cluster import KMeans
```

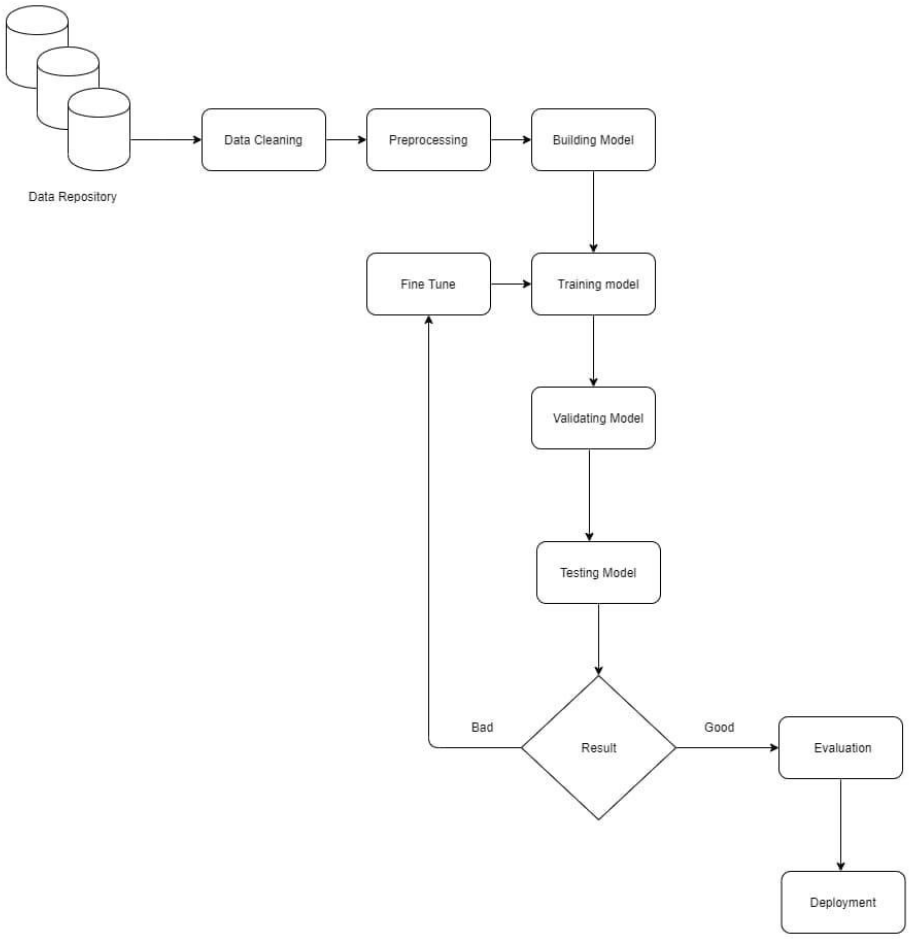


Figure 2: Diagram of General Architecture Overview

CHAPTER 4

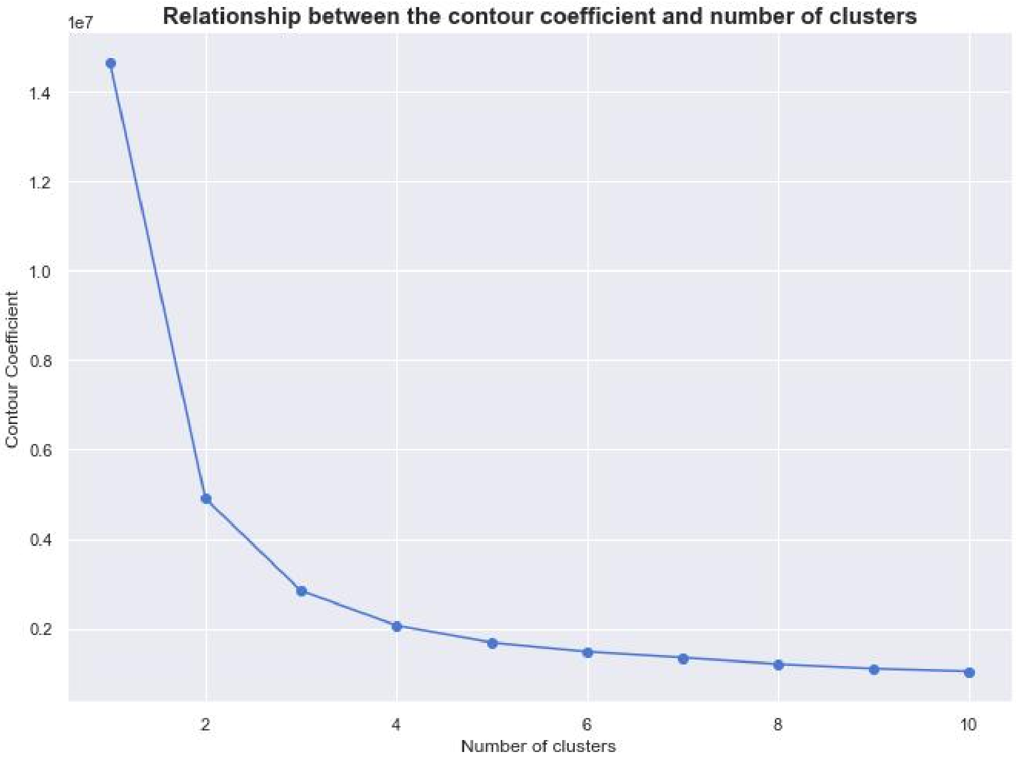
RESULTS & EVALUATION

This chapter discusses the various results products from implementation of various models used within this study. These are presented as follows:

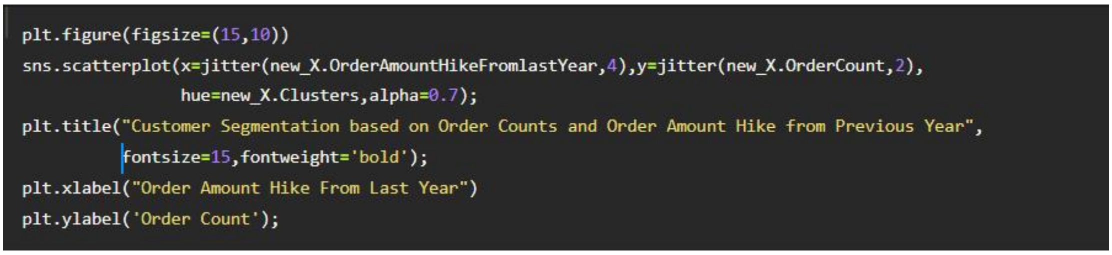
4.1. Customer Segmentation with K-means

Being inspired by [8], the customer segmentation was implemented using the K-means clustering algorithm. The clustering was done based on customer purchase behavior and the columns that were selected for this clustering where the Order Count column telling indicating the number of purchases the customer has made over the last month and OrderAmountHikeFromLastYear indicating the increase in purchases compare to the previous year, expressed as a percentage.

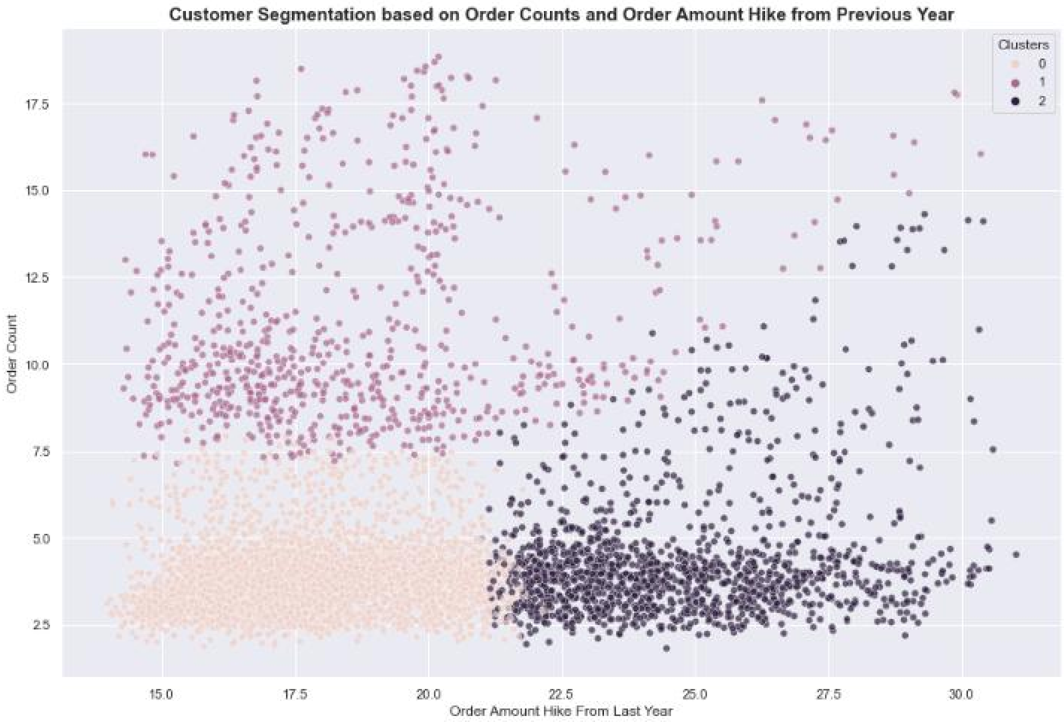
Clustering with the K-means is a bit tricky since it requires the number of clusters to be created to be explicitly stated. This may not always be very obvious in the real world application such as this hence, the Elbow method was employed to give a hint of the optimal number of clusters that should be ideally formed for this data set.



With the help of the Elbow method, three clusters were created with using the scatter plot function as seen in the diagram below. Here cluster 1 is Cluster I, cluster 2 is Cluster II and cluster 0 is Cluster III.



```
plt.figure(figsize=(15,10))
sns.scatterplot(x=jitter(new_X.OrderAmountHikeFromlastYear,4),y=jitter(new_X.OrderCount,2),
                hue=new_X.Clusters,alpha=0.7);
plt.title("Customer Segmentation based on Order Counts and Order Amount Hike from Previous Year",
          fontsize=15,fontweight='bold');
plt.xlabel("Order Amount Hike From Last Year")
plt.ylabel('Order Count');
```



The table below presents customer segmentation on the discussed E-commerce dataset. Group with the K-means algorithm on the features OrderCount and OrderAmountHikeFromLastYear

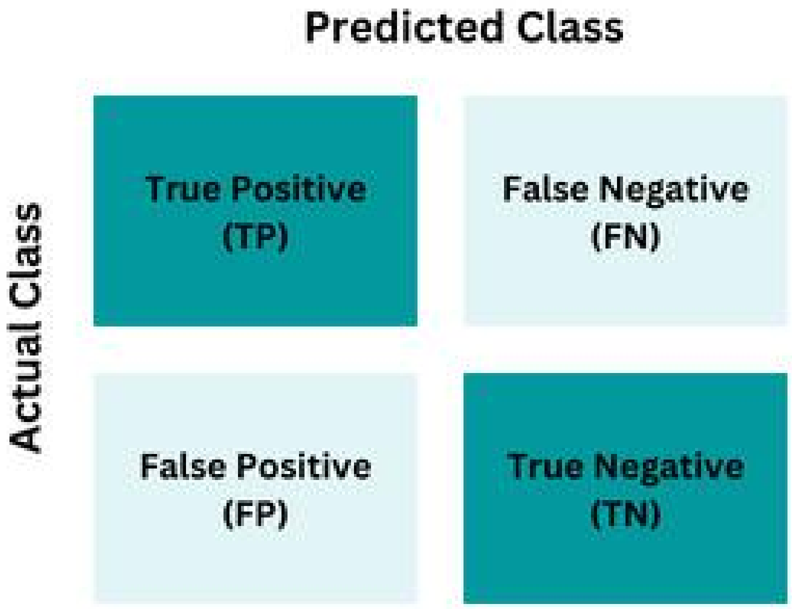
Table 1. Results of K-means Classification

Data set name	**Full data set**	**Churn**	**Non-Churn**
E-commerce data set	5630	948	4682
Cluster I	733	122	611
Cluster II	1477	238	1239
Cluster III	3420	588	2832

4.2 Evaluation Metrics

Confusion Matrix

The confusion matrix is a metric that gives on overview of model performance with respect to actual values versus predicted values. It can be used for both binary and multi-class classifications problems [16].



True Positive(TP) indicates the number of positive examples classified accurately.

False Positive(FP) indicates the number of negative examples classified as positive.

True Negative (TN) indicates the number of negative examples classified accurately.

False Negative (FN) indicates the number of positive examples classified as negative.

Accuracy:

Accuracy of an algorithm is represented as the ratio of correctly classified patients (TP+TN) to the total number of patients (TP+TN+FP+FN).

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Receiver Operating Characteristic (ROC) and Area Under Curve (AUC):

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate (TPR) = $\frac{TP}{TP+FN}$
- False Positive Rate (FPR) = $\frac{FP}{FP+TN}$

The Area Under Curve (AUC) metric measures the entire two-dimensional area underneath the entire ROC curve.

4.3 Support Vector Machine (SVM)

Based on the works of [8], the SVM model was implemented on the data set described above. This processes incorporated hyper-parameter tuning and class imbalance handling with SMOTE. At the end of the process, the experiment produced the following results.

Accuracy

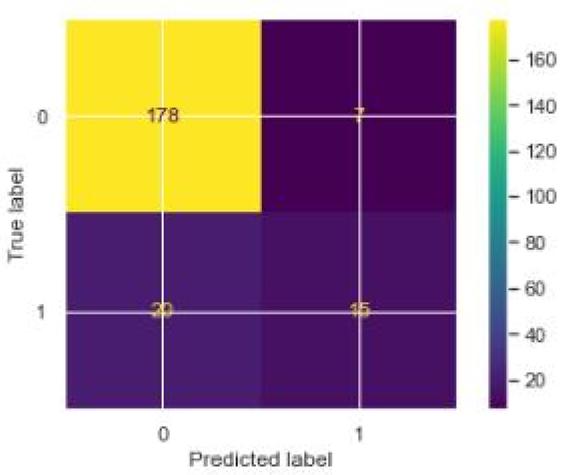
Customer Segments	**Training Set**	**Test Set**
Cluster I	0.8850	0.8773
Cluster II	0.9255	0.9167
Cluster III	0.8571	0.8567
Avg. Accuracy	**0.8892**	**0.8836**

AUC

Customer Segments	**Training Set**	**Test Set**
Cluster I	0.6975	0.6954
Cluster II	0.8016	0.8005
Cluster III	0.5848	0.5916
Avg. AUC	**0.6946**	**0.6958**

Confusion Matrices

Cluster I



Here the SVM model performs classification on 30% of customers (220 customers) from Cluster I.

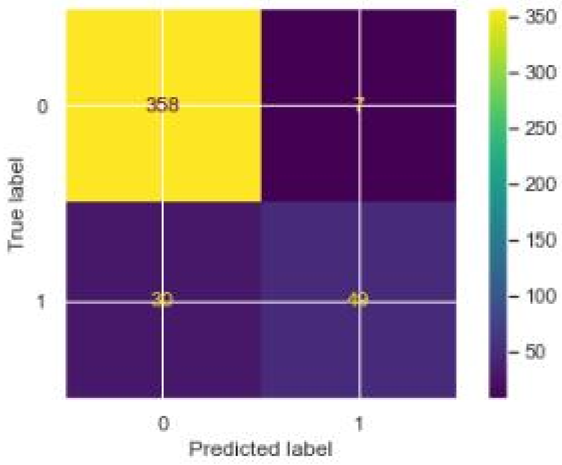
TP - It accurately classified 178 Non-Churn customers

FP - It wrongly classified 20 Churn customers as Non-Churn customers

TN - It accurately classified 15 Churn customers

FN - It accurately classified 7 Non-Churn customers as Churn customers

Cluster II



Here the SVM model performs classification on 30% of customers (444 customers) from Cluster II.

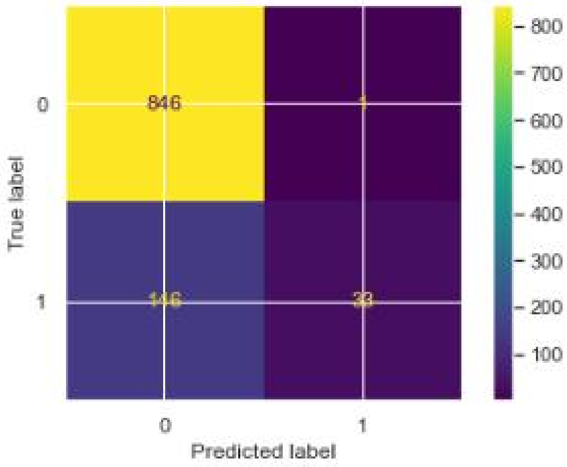
TP - It accurately classified 358 Non-Churn customers

FP - It wrongly classified 30 Churn customers as Non-Churn customers

TN - It accurately classified 49 Churn customers

FN - It accurately classified 7 Non-Churn customers as Churn customers

Cluster III



Here the SVM model performs classification on 30% of customers (1026 customers) from Cluster III.

TP - It accurately classified 846 Non-Churn customers

FP - It wrongly classified 146 Churn customers as Non-Churn customers

TN - It accurately classified 33 Churn customers

FN - It accurately classified 1 Non-Churn customers as Churn customers

4.4 Extreme Gradient Boosting Trees (XGBoost)

Based on the works of [13], the Extreme Gradient Boosting (XBoost) trees was implemented on the clusters obtained from customer segments performed with the K-means clustering algorithm. This processes incorporated hyper-parameter tuning and class imbalance handling. At the end of the process, the experiment produced the following results.

Accuracy

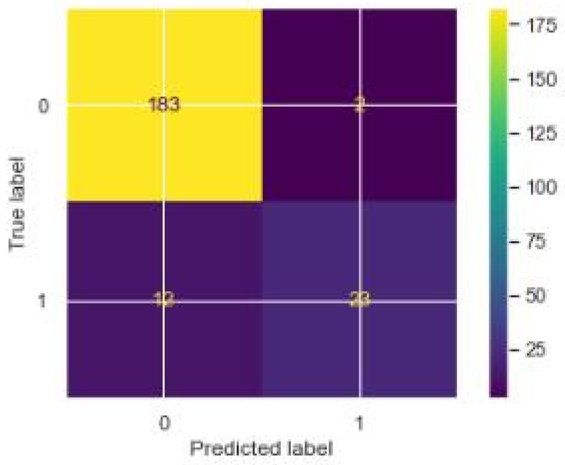
Customer Segments	**Training Set**	**Test Set**
Cluster I	0.9493	0.9364
Cluster II	0.9671	0.9167
Cluster III	0.9148	0.8977
Avg. Accuracy	**0.9437**	**0.9169**

AUC

Customer Segments	**Training Set**	**Test Set**
Cluster I	0.8506	0.8232
Cluster II	0.8957	0.7708
Cluster III	0.8001	0.7706
Avg. AUC	**0.8488**	**0.7882**

Confusion Matrices

Cluster I



Here the XGBoost model performs classification on 30% of customers (220 customers) from Cluster I.

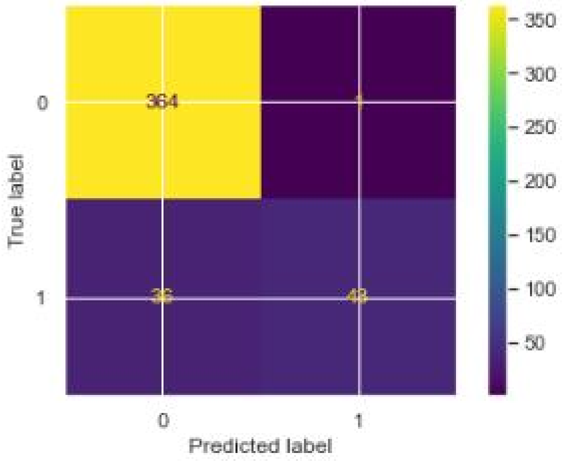
TP - It accurately classified 183 Non-Churn customers

FP - It wrongly classified 12 Churn customers as Non-Churn customers

TN - It accurately classified 23 Churn customers

FN - It accurately classified 2 Non-Churn customers as Churn customers

Cluster II



Here the XGBoost model performs classification on 30% of customers (444 customers) from Cluster II.

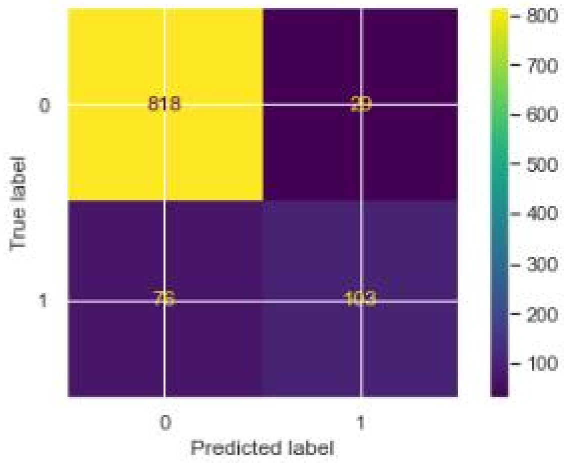
TP - It accurately classified 364 Non-Churn customers

FP - It wrongly classified 36 Churn customers as Non-Churn customers

TN - It accurately classified 43 Churn customers

FN - It accurately classified 1 Non-Churn customers as Churn customers

Cluster III



Here the XGBoost model performs classification on 30% of customers (1026 customers) from Cluster III.

TP - It accurately classified 818 Non-Churn customers

FP - It wrongly classified 76 Churn customers as Non-Churn customers

TN - It accurately classified 103 Churn customers

FN - It accurately classified 29 Non-Churn customers as Churn customers

4.5 Light Gradient Boosting Machine

One of the objectives of this study was to propose an alternative algorithm for e-commerce customer churn classification. The Light Gradient Boosting Machine (lgbm) algorithm was implemented for this reason. With hyper-parameter tuning and class imbalance handled within the algorithm, the following results was produced.

Accuracy

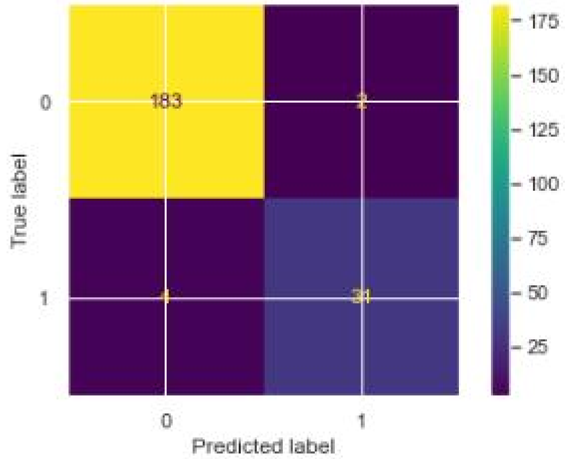
Customer Segments	**Training Set**	**Test Set**
Cluster I	0.9609	0.9091
Cluster II	0.9681	0.9212
Cluster III	0.9440	0.9094
Avg. Accuracy	**0.9577**	**0.9132**

AUC

Customer Segments	**Training Set**	**Test Set**
Cluster I	0.9818	0.9592
Cluster II	0.9887	0.9382
Cluster III	0.9755	0.9518
Avg. AUC	**0.9820**	**0.9497**

Confusion Matrices

Cluster I



Here the LGBM model performs classification on 30% of customers (220 customers) from Cluster I.

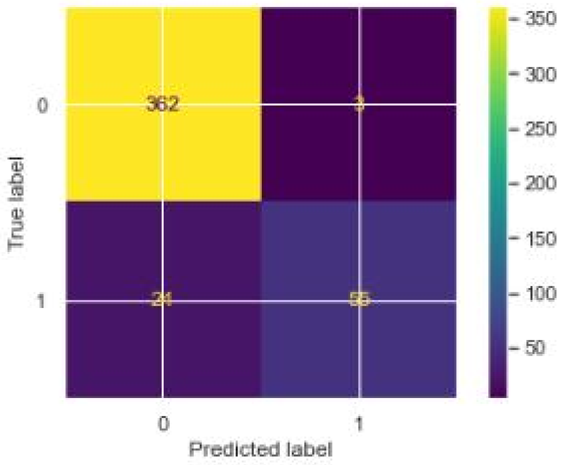
TP - It accurately classified 183 Non-Churn customers

FP - It wrongly classified 1 Churn customers as Non-Churn customers

TN - It accurately classified 31 Churn customers

FN - It accurately classified 2 Non-Churn customers as Churn customers

Cluster II



Here the LGBM model performs classification on 30% of customers (444 customers) from Cluster II.

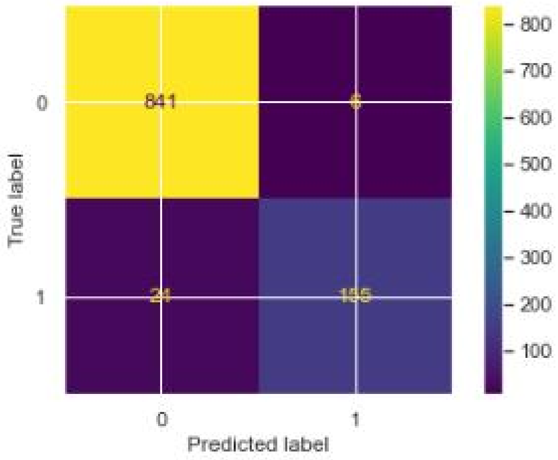
TP - It accurately classified 362 Non-Churn customers

FP - It wrongly classified 21 Churn customers as Non-Churn customers

TN - It accurately classified 55 Churn customers

FN - It accurately classified 3 Non-Churn customers as Churn customers

Cluster III



Here the LGBM model performs classification on 30% of customers (1026 customers) from Cluster III.

TP - It accurately classified 841 Non-Churn customers

FP - It wrongly classified 21 Churn customers as Non-Churn customers

TN - It accurately classified 155 Churn customers

FN - It accurately classified 6 Non-Churn customers as Churn customers

Result summary table

Model	Avg. Accuracy	Avg. AUC	TP Cluster (I+II+III)	FP Cluster (I+II+III)	TN Cluster (I+II+III)	FN Cluster (I+II+III)
SVM	0.8836	0.6958	1382	196	97	15
XGBoost	0.9169	0.7882	1365	124	169	32
LGBM	0.9132	0.9497	1386	43	241	11

Selecting Best Metrics and Model

Model	Avg. Accuracy	Avg. AUC	TP Cluster (I+II+III)	FP Cluster (I+II+III)	TN Cluster (I+II+III)	FN Cluster (I+II+III)
SVM	0.8836	0.6958	1382	196	97	15
XGBoost	**0.9169**	0.7882	1365	124	169	32
LGBM	0.9132	**0.9497**	**1386**	**43**	**241**	**11**

CHAPTER 5

CONCLUSION

Customer churn prediction plays an important role in the e-commerce sector. The results of this study also have some limitations. The study was carried out using an a single e-commerce customer data set with 5360 customers which may not reflex the entire behavior of customers world-wide. Ideally this research should have been verified with several data sets with a larger pool of customer data.

At the end of the entire study, the Support Vector Machine was found to be the least performing among the 3 models tested over the data set with an mean accuracy of 0.8836 and mean AUC of 0.6958 over the 3 customer segments. It had the most amount of mis-classifications (FP + FN = 196+15 = 211).

The Extreme Gradient Boosting Machine (XGBoost) took second place in model performance as it even had a higher accuracy than both SVM and LGBM models of 0.9169 and a second best AUC of 0.7882. The XGBoost model still had a substantial amount of mis-classifications (FP + FN = 124 + 32 = 156)

Finally, our proposed model for this study which is the Light Gradient Boosting Machine (LGBM) produced the best overall performance with a mean accuracy of 0.9132 and mean AUC of 0.9497. It had a drastically low amount is mis-classifications (FP + TP = 43 + 11 = 54) then the other 2 models.

Observing the false positives and false negatives for the all 3 models, we see that all models have a high number of false positives - 196, 124 and 43 for SVM, XGBoost and LGBM models respectively - and significantly lower number of false negatives - 15, 32 and 11 for SVM, XGBoost and LGBM models respectively - as compared to false positives for all 3 models.

Looking at the performance of our proposed model, Light Gradient Boosting Machine, it is observed that in spite of the class imbalance in the dataset, its effects are not as drastic as compared to the other 2 models used. The LGBM performs much better on imbalance data as compared to SVM and XGBoost models according to this test which is a big plus as most of real world data is naturally skewed in class or imbalanced.

With the help of the presented machine learning models and research performed, business owners and stakeholder in the e-commerce sector can now have the technical knowledge and tools to be able to easily adopt this technology into their business setting to forecast likely churners and quickly take remedial action to improve customer satisfaction and retention

BIBLOGRAPHY

[1] Van den Poel, D. and Larivière, B., 2022. Customer attrition analysis for financial services using proportional hazard models.

[2] X. Yu, S.Guo, J. Guo, and X.Huang, “An extended support vector machine forecasting framework for customer churn in e-commerce”

[3] Wu, X. J., & Meng, S. S., 2017. “Research on e-commerce customer churn prediction based on customer segmentation and Ada-Boost”. Industrial Engineering, 20(02), 99- 107

[4] Shao, D. 2016. “Analysis and prediction of insurance company’s customer loss based on BP neural network. Lanzhou University”

[5] Alshamsi, Abdulrahman. 2022. “Customer Churn prediction in E-Commerce Sector”. Thesis. Rochester Institute of Technology.

[6] McFerrin, Joe. 2021.“The History of eCommerce: from its Origins to Modern Day”. IWD.

[7] Shankar, V., Kalyanam, K., Setia, P., Golmohammadi, A., Tirunillai, S., Douglass, T., Hennessey, J., Bull, J. and Waddoups, R., 2021. How Technology is Changing Retail. Journal of Retailing, 97(1), pp.13-27.

[8] Xiahou, X.; Harada, Y. 2022. “B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM.” J. Theor. Appl. Electron. Commer. Res,17,458–475.

[9] Hassibi, Khosrow.2016. “Machine learning vs. Traditional: Different philosophies, Different Approaches. Data Science Central”

[10] Lu, N., Liu, X. W., & Lee, L. 2018. “Research on customer value segmentation of online shop based on RFM. Computer Knowledge and Technology”, 14(18), 275-276, 284

[11] Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. 2020. “Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting”. Information Fusion, 54, 128-144.

[12] Varad R Thalkar. 2021. Comput. Sci. Eng. Inf. Technol, Int. J. Sci. Res,7 (6) : 207-211

[13] Xiahou, X.; Harada, Y. 2022. “B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM.” J. Theor. Appl. Electron. Commer. Res,17,458–475.

[14] Kanade, Vijay. 2022.”Artificial Intelligence: What is Machine Learning?”. Spiceworks.

[15] Oriane, Ferrera. 2015.“The customers’ perception of servicescape’s influence on their behaviours in the food retail industry”.

[16] “Confusion matrix,” *Confusion Matrix - an overview | ScienceDirect Topics*. [Online]. Available: https;//www.sciencedirect.com/topics/engineering/conusion-matrix [Accessed: 28-Sep-2022].

APPENDIX

The repository to the notebooks for the various models:

https://github.com/VanessaAttaFynn/Customer_Churn_Prediction