5-2022

Customer Churn prediction in ECommerce Sector

Abdulrahman Alshamsi
aya8777@rit.edu

Customer Churn prediction in E-Commerce Sector

by

Sultan Abdulrahman Alshamsi

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research
Rochester Institute of Technology

RIT Dubai

May 2022

RIT

Master of Science in Professional Studies:
Data Analytics

Graduate Capstone Approval

Student Name: Sultan Abdulrahman Alshamsi

Graduate Capstone Title: Customer Churn prediction in E-Commerce Sector

Graduate Capstone Committee:

Name: **Dr. Sanjay Modak** **Date:**
Chair of committee

Name: **Dr. Khalil Al Hussaeni** **Date:**
Member of committee

ACKNOWLEDGMENTS

I would like to express my thanks for all of people who contributed to my success in my learning journey. My sincere appreciation to Dr. Khalil for his guidance and support throughout the project. I would like also to thank my colleagues in the program for their cooperation and help during our study. Finally, I would like to thank Rochester Institute of Technology for offering this program and creating a great learning environment.

ABSTRACT

With the increasing popularity of e-commerce and the rapid development. The competition among e-commerce companies becomes fiercer (Zhang, 2015). As known, the most significant matter is to retain customers by providing the best services along with a suitable price. Therefore, with the rapid increase of e-commerce transaction volume and intense competition in the market to meet the high demand from customers, it is necessary to attract customers through customised services and targeted strategies to increase customer loyalty. On the other hand, e-commerce customer churn shows nonlinear changes and asymmetrical customer categories. E-commerce customer churn data has a typical sample imbalance which means that the number of churn samples is significantly larger than the number of non-churned samples or vice versa. This research proposes the methods or models imposed on e-commerce to proactively reduce customer churn.

Key words: Customer behaviour, Churn prediction, Customer churn, Digital marketing.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

1.1. Background of the Problem

In E-commerce, it turns out it is much more expensive for a company to attract new customers than to retain current ones (Saghir et al., 2019). For this specific reason, knowing in advance which customers will leave the company will enable the businesses to create offers or reduce the consumption of its products or services in a relevant way is crucial to increase their retention, build a good Customer Relationship and save acquisition costs. In today's competitive market, there is an immense variety of products and services to choose from. Accordingly, most consumers got used to walk freely from one brand to another, from one supplier to another, looking for the product or service that suits their needs. E-commerce Companies have been suffering from this phenomenon, known as customer "churn,". Therefore, instead of focusing on retaining their current customers, they often invest effort and allocate huge amount of money in attracting new customers.

In response to the above problem, existing research on customer churn mainly includes predictions related to traditional statistics, artificial intelligence, predictions based on statistical learning theory, predictions based on combined classifiers. This paper proposes research on e-commerce customer churn prediction based on customer segmentation, using improved SMOTE for data balance, and then using three different machine learning algorithms for prediction. Finally, predictors importance is identified to help decision makers in choosing the proper decisions on behalf of the organization.

1.2. Statement of the problem

Retaining customers is a challenging issue that is encountering most of organizations, particularly businesses operating in e-commerce sector. According to Wu et al., (2017), it is much more difficult to retain the existing customers as compared to attract new ones because existing customers provides high value in ecommerce; however, to attract new customers, companies need to invest a lot of money for making them as loyal customers. This study will develop a prediction model for E-commerce sector to correlate the key attributes leading to churn.

1.3. Project Goals

The project is important because Churn Prediction and analysis will allow e-commerce companies to anticipate and determine which clients are susceptible to migrate. As Churn predicted in e-commerce will help to know the real value of the potential loss of those clients to take the necessary retention measures to reduce or avoid their migration. The project goals are as follows.

- To propose commercial actions aimed at maintaining clients that are showing signs of churn and offer them customised offers.
- To develop prediction models for the customer churns in Ecommerce Company, analysing different attributes related to customer churn.
- To follow the CRISP-DM data mining methodology in a structured way based on the modelling, evaluation, and implementation of predicted models.
- To apply three different machine learning models which are Decision tree, Logistic Regression and Random Forest.

1.4. Methodology

Data mining methodologies were introduced to provide a more rounded view to the knowledge discovery process, beyond the phase of applying machine learning models or algorithms. Their aim is to provide a generic workflow of a data mining project (Huber et al., 2019). The proposed research will work with the CRISP-DM methodology that consists of the cyclical stages starting with understanding the business where we identify the problem and investigate its business opportunity. Then, we go to Data understanding in which we retrieve our data from several sources. After that, we must prepare the data in which collected data will go through different stages

including cleansing the data by removing the missing values, outliers, and unnecessary columns. Finally, we will end up with modelling, evaluation, and deployment where we will implement our models test it to measure its cost and accuracy (Huber et al., 2019). The CRISP-DM methodology generated new data mining processes that respond to the business problem of sending the right messages at the right time and place. The proposed research will develop the CRISP-DM data mining methodology, intending to guide in a structured way the knowledge discovery process based on the tasks of data understanding, data preparation, modelling, evaluation, and implementation. The technique focused on our problem is also explained. The technique identified above as the best is applied with the variables identified for our problem.

CRISP-DM data mining methodology was selected for this project due to due to its cross-industrial nature, CRISP-DM can be easily implemented on any data mining project regardless to its domain. Furthermore, it provides a road map for the researcher while carrying out the data mining project, it gives a uniform framework for planning and managing a project. Finally, it is proven that it is one of the most time and cost-effective methodologies.

1.5. Limitations of the study

There were some limitations which might hindered the project, few points are listed below:

1. Lack of similar previous projects as most of customer churn projects are directed towards the telecommunication sector.
2. Difficulty to get a rich dataset for an E-Commerce business based in the UAE due to confidentiality. In fact, obtaining a data set from a n E-commerce enterprise in the UAE would have helped in understanding customers behaviour in this region.

CHAPTER 2 - LITERATURE REVIEW

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 telecommunication sector. E-commerce 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 retain this type of customer group.

The value of e-commerce customer churn prediction is to merge e-commerce customer data over some time and establish e-commerce customer churn prediction models by analysing customer purchase behaviours (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 cost of as the new customers want to bring high profit in ecommerce. Comparatively, the customer purchase behaviour differs for both existing and new customers; however, it is essential to identify the reasons leading for the customer's loss. In support, Lu et al., (2018) stated that in the e-commerce sector, it is extremely important to analyse 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 behaviour 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 has been widely used in the customer relationship management of e-commerce companies, such as customer segmentation, customer churn prediction, and fraud analysis.

Customer Segmentation

Customer segmentation is used for the recognition of the value of customer relationship, a key step prerequisite for more efficient targeted marketing activities (Feng et al., 2018). According to the famous 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 (Sun et al., 2020). Therefore, to perform customer segmentation, we must first identify and tap customer value. If companies can focus on the actual value of customers and allocate resources for targeted marketing, they can improve their core competitiveness.

With the rapid development of e-commerce, corporate business activities carried out through the internet are more real-time and interactive, which also transforms the product-centric business type into a customer-centric model for ecommerce business (Dhote et al., 2020). In the view of Agrawal et al., (2018), customers become enterprises to gain profits and an important resource for competitive advantage. The churn rate of e-commerce customers is relatively high. If companies want to establish long-term alliances with customers and develop stable and continuous customer relationships. The e-commerce customer base is extensive and complex, and its value varies. The study of Saghir et al., (2019) raised a question about how to accurately identify high-value customers, predict churn and retain them in advance has become a hot spot in the field of e-commerce. In response, Wu et al., (2021) studied that the evaluation of customer value helps companies identify valuable customers among many consumers and implement different customer management according to different customer values so that the limited resources of the company can maximize the effect. Purchasing value as the main explicit value of e-commerce customers is selected by the text as the main measure of customer value. The amount of purchase is directly related to the sales volume of enterprise products or services. It is the guarantee for achieving the profit target of the enterprise and the value of customers. The most direct manifestation of. Here we mainly use historical transaction data to identify the value of customers and use this as a basis for customer segmentation.

Random Forest

This algorithm was introduced by a researcher named Tin Kam Ho (Ho, 1995). The concept of Random Forest algorithm is to create a prediction about the uncorrelated forests of trees, and the prediction of is known to be more accurate as compared to any kind of individual tree. Geetha et

al. (2020) mentioned that Random Forest utilises randomness of features and bagging, whenever there is a requirement of creating an individual tree from the uncorrelated forests of trees. Furthermore, the researchers proposed to employ Random Forest Classifier along with Support Vector Machine for computing customer churn. It was revealed in the study that the mentioned algorithms boosted the accuracy to 95 per cent. Hence, making them a more suitable and efficient system for computing the customer churns. Ullah (2019) proposed a customer churn prediction model by employing random forest algorithm in the telecom sector. The researchers classified the churned customers data, and the computed Random Forest accuracy was around 87 per cent.

AdaBoost

The Adaptive Boosting algorithm method is a mix algorithm that was proposed by Freund et al., (1995) which is one of the best Boosting algorithms. It uses an adaptive resampling method, creates a strong classifier by combining a set of weak classifiers. This technique will increase the sampling probability of the wrongly divided samples. It enhances the classification accuracy by sequential learning from previous classifiers. The balanced dataset is input into the integrated learning algorithm AdaBoost to train the weak classifier and create alterations to enhance churn prediction.

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 minority, searches its k nearest neighbours, which randomly selected k -nearest neighbour 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 dataset was imbalanced.

CHAPTER 3 - PROJECT DESCRIPTION

This project will go through several steps to build a customer churn prediction model. First, the dataset will go through preprocessing where the dataset is cleaned and to get best performance during modelling. After that, we will go through data visualization to get some insights about the data set in addition to the common aspects of churned customers. Finally, machine learning algorithms will be utilized to build customer churn prediction model.

3.1. Data Collection

The data is collected for this project from Kaggle for an e-commerce website. The empirical study starts in June 2021, and the observation ends in November 2021. The consumption data of customers who purchased goods on the website is selected for analysis and prediction. The dataset contains customer's consumption records in addition to historical behavioural interactions while using the platform.

3.2. Data Description

Data is collected form a leading E-commerce platform, it is a historical data containing customer details and experience and its outcome is customer churn flag (churn= 1, no churn = 0). The dataset shows more than 5000 customers and their interaction and preferences in the platform. The effectiveness of this data is that it contains some specific detailed attributes which will help in customer segmentation such as: preferred login device, Satisfaction store and other attributes. These attributes will help us in studying the causes of churn in each customer's segment to identify the triggers leading to customer churn. (Table 2 includes the name of the attributes, description, and type).

#	Attribute	Description	Type
1	CustomerID	Unique customer ID	Numeric
2	Churn	Churn flag	Numeric
3	Tenure	Tenure of customer in organization	Numeric
4	PreferrdLoginDevice	Preferred login device of customer	Character
5	CityTier	City tier	Numeric

6	WarehouseToHome	Distance in between warehouse to home of customer	Numeric
7	PreferedPaymentMode	Preferred payment method of customer	Character
8	Gender	Gender of customer	Character
9	HourSpendOnApp	Number of hours spend on mobile application or website	Numeric
10	NumberOfDeviceRegistered	Total number of devices registered per customer	Numeric
11	PreferedOrderCat	Preferred order category of customer in last month	Character
12	SatisfactionScore	Satisfactory score of customers on service	Numeric
13	MaritalStatus	Marital status of customer	Character
14	NumberOfAddress	Total number of added addresses per each customer	Numeric
15	Complain	Any complaint has been raised in last month	Numeric
16	OrderAmountHikeFromlastYear	Percentage increases in order from last year	Numeric
17	CouponUsed	Total number of coupons has been used in last month	Numeric
18	OrderCount	Total number of orders placed in last month	Numeric
19	DaySinceLastOrder	Day Since last order by customer	Numeric
20	CashbackAmount	Average cashback in last month	Numeric

Table 1: Data dictionary

CHAPTER 4 - PROJECT ANALYSIS

4.1. Exploratory Data Analysis

The first step in data exploration is to import several libraries in R to explore and visualize the data. then the numerical and categorical columns will be explored in addition to identification of missing data.

The outcome of our data set is Churn, and there are no missing values in "churn" column. However, the outcomes variables are imbalanced due to the high number of retained customers in comparison to churned customers as shown in the table below.

churn	frequency
0	4,682
1	948

Table 2 : Churned versus retained customers

The following table contains the summary statistics of all numeric columns in our data. any column that has n=5630 shows that there are no missing values as we have 5630 unique customer IDs.

variable	n	min	max	median	iqr	mean	sd
CashbackAmount	5,630	0	324.99	163.28	50.623	177.223	49.207
CityTier	5,630	1	3.00	1.00	2.000	1.655	0.915
Complain	5,630	0	1.00	0.00	1.000	0.285	0.451
CouponUsed	5,374	0	16.00	1.00	1.000	1.751	1.895
DaySinceLastOrder	5,323	0	46.00	3.00	5.000	4.543	3.654
HourSpendOnApp	5,375	0	5.00	3.00	1.000	2.932	0.722
NumberOfAddress	5,630	1	22.00	3.00	4.000	4.214	2.584
NumberOfDeviceRegistered	5,630	1	6.00	4.00	1.000	3.689	1.024
OrderAmountHikeFromlastYear	5,365	11	26.00	15.00	5.000	15.708	3.675
OrderCount	5,372	1	16.00	2.00	2.000	3.008	2.940
SatisfactionScore	5,630	1	5.00	3.00	2.000	3.067	1.380
Tenure	5,366	0	61.00	9.00	14.000	10.190	8.557
WarehouseToHome	5,379	5	127.00	14.00	11.000	15.640	8.531

Table 3: Summary statistics for numerical columns

The following table studies the relationship between each the frequency of each attribute with respective of the outcome which is either churn or no churn. In addition, it reflects the mean, median of each attribute, Further details are listed below:

Churn	variable	n	mean	median
0	CashbackAmount	4,682	180.635	166.115
1	CashbackAmount	948	160.371	149.660
0	CityTier	4,682	1.620	1.000
1	CityTier	948	1.827	1.000
0	Complain	4,682	0.234	0.000
1	Complain	948	0.536	1.000
0	CouponUsed	4,434	1.758	1.000
1	CouponUsed	940	1.717	1.000
0	DaySinceLastOrder	4,429	4.807	4.000
1	DaySinceLastOrder	894	3.236	2.000
0	HourSpendOnApp	4,485	2.926	3.000
1	HourSpendOnApp	890	2.962	3.000
0	NumberOfAddress	4,682	4.163	3.000
1	NumberOfAddress	948	4.466	3.000
0	NumberOfDeviceRegistered	4,682	3.639	4.000
1	NumberOfDeviceRegistered	948	3.935	4.000
0	OrderAmountHikeFromlastYear	4,431	15.725	15.000
1	OrderAmountHikeFromlastYear	934	15.627	14.000
0	OrderCount	4,442	3.047	2.000
1	OrderCount	930	2.824	2.000
0	SatisfactionScore	4,682	3.001	3.000
1	SatisfactionScore	948	3.390	3.000
0	Tenure	4,499	11.502	10.000
1	Tenure	867	3.379	1.000
0	WarehouseToHome	4,515	15.354	13.000
1	WarehouseToHome	864	17.134	15.000

Table 4 : Relationship between numerical columns and churn

According to table 4, churned customers, compared to retained customers have:

1. Lower mean and median CashbackAmount.

2. Higher mean CityTier, but equal median CityTier to non-churned customers.
3. Higher mean and median Complain.
4. Lower mean CouponUsed, but equal median.
5. Lower mean and median DaySinceLastOrder.
6. Nearly equal mean and median HourSpendOnApp.
7. Higher mean NumberOfAddress, but equal median.
8. Higher mean NumberOfDeviceRegistered, but equal median.
9. Lower mean and median OrderAmountHikeFromlastYear.
10. Lower mean OrderCount, but equal median.
11. Unexpectedly higher mean SatisfactionScore, but equal median.
12. Greatly lower mean and median Tenure.
13. Higher mean and median WarehouseToHome.

The following table shows the frequency and level percentage of all categorical columns in our dataset. the frequency of each level is shown to analyze the frequency of each level in each variable, more details are shown below:

variable	level	n	percentage
Gender	Female	2,246	40
Gender	Male	3,384	60
MaritalStatus	Divorced	848	15
MaritalStatus	Married	2,986	53
MaritalStatus	Single	1,796	32
PreferedOrderCat	Fashion	826	15
PreferedOrderCat	Grocery	410	7
PreferedOrderCat	Laptop & Accessory	2,050	36
PreferedOrderCat	Mobile	809	14
PreferedOrderCat	Mobile Phone	1,271	23
PreferedOrderCat	Others	264	5
PreferredLoginDevice	Computer	1,634	29
PreferredLoginDevice	Mobile Phone	2,765	49
PreferredLoginDevice	Phone	1,231	22
PreferredPaymentMode	Cash on Delivery	149	3
PreferredPaymentMode	CC	273	5
PreferredPaymentMode	COD	365	6
PreferredPaymentMode	Credit Card	1,501	27
PreferredPaymentMode	Debit Card	2,314	41
PreferredPaymentMode	E wallet	614	11
PreferredPaymentMode	UPI	414	7

Table 5: Statistical summary for categorical attributes

In the previous table, each categorical attribute was divided in two classes based on the outcome. In fact, this table will help us in highlighting the common aspects of churned customers. Categorical variables are complete and zero missing values were observed. Some changes can be made to the data set including the following, “Mobile” and “Mobile Phone” to be grouped under

"Mobile". Furthermore, In PreferredPaymentMode column, "COD" and "Cash on Delivery" to be classified as "Cash on Delivery" and "Credit Card" and "CC" to be labeled as "Credit Card".

variable	level	Churn	n	percentage
Gender	Female	0	1,898	85
Gender	Female	1	348	15
Gender	Male	0	2,784	82
Gender	Male	1	600	18
MaritalStatus	Divorced	0	724	85
MaritalStatus	Divorced	1	124	15
MaritalStatus	Married	0	2,642	88
MaritalStatus	Married	1	344	12
MaritalStatus	Single	0	1,316	73
MaritalStatus	Single	1	480	27
PreferedOrderCat	Fashion	0	698	85
PreferedOrderCat	Fashion	1	128	15
PreferedOrderCat	Grocery	0	390	95
PreferedOrderCat	Grocery	1	20	5
PreferedOrderCat	Laptop & Accessory	0	1,840	90
PreferedOrderCat	Laptop & Accessory	1	210	10
PreferedOrderCat	Mobile	0	1,510	73
PreferedOrderCat	Mobile	1	570	27
PreferedOrderCat	Others	0	244	92
PreferedOrderCat	Others	1	20	8
PreferredLoginDevice	Computer	0	1,310	80
PreferredLoginDevice	Computer	1	324	20
PreferredLoginDevice	Mobile Phone	0	2,417	87
PreferredLoginDevice	Mobile Phone	1	348	13
PreferredLoginDevice	Phone	0	955	78
PreferredLoginDevice	Phone	1	276	22
PreferredPaymentMode	Cash on Delivery	0	386	75
PreferredPaymentMode	Cash on Delivery	1	128	25
PreferredPaymentMode	Credit Card	0	1,522	86
PreferredPaymentMode	Credit Card	1	252	14
PreferredPaymentMode	Debit Card	0	1,958	85
PreferredPaymentMode	Debit Card	1	356	15
PreferredPaymentMode	E wallet	0	474	77
PreferredPaymentMode	E wallet	1	140	23
PreferredPaymentMode	UPI	0	342	83
PreferredPaymentMode	UPI	1	72	17

Table 6: Relationship between categorical attributes and churn column

It is observed from the table above that churned customers are associated with:

1. Male gender.
2. Single marital status.
3. Mobile PreferedOrderCat.
4. Phone and computer PreferredLoginDevice.
5. Cash on Delivery PreferredPaymentMode.

The bar plot below illustrates the missing values in the data set, it shows the number of missing values per attribute.

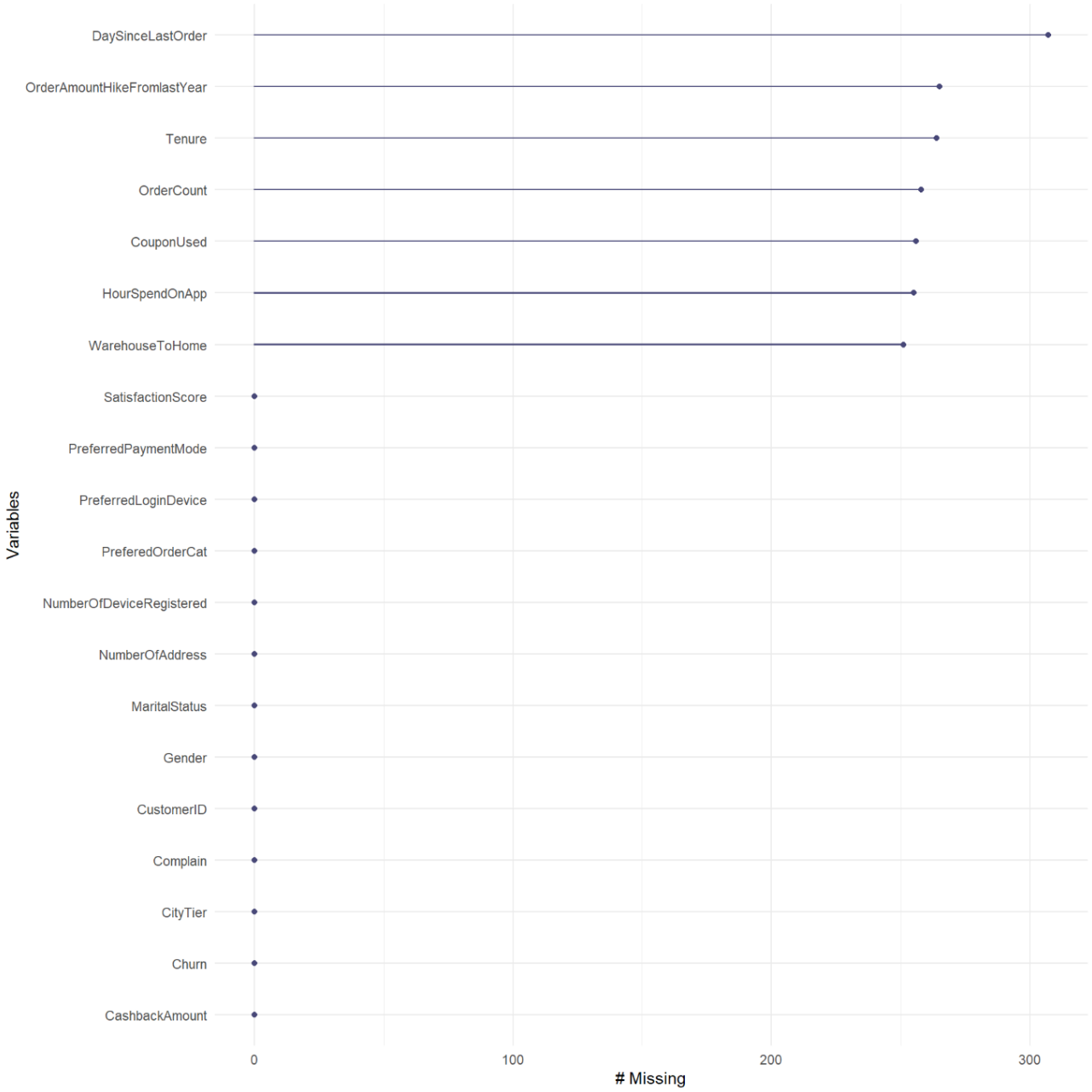

Figure 1: Missing values bar plot

The bar plot above illustrates the number of missing values per numerical attributes. Total of seven numerical attributes observed with missing values and the highest attribute in missing values is DaySinceLastOrder column, the percentage of missing values per column doesn't exceed 6%, for instance, DaySinceLastOrder column has 307 missing values which equals

5.54%. Hence, all rows containing missing values have been omitted to avoid generating errors when training or testing data.

4.2. Data Cleaning

Some steps were performed after data exploration which will help in creating more accurate machine learning models and eliminate bias. and after data cleaning, our data is composed of 3774 rows and 20 columns.

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". Finally, all missing values have been omitted to eliminate errors while building the models.

4.3. Data Visualization

In the bar charts below, we illustrated all the numerical attributes in our dataset to study their distribution, further description is shown below.

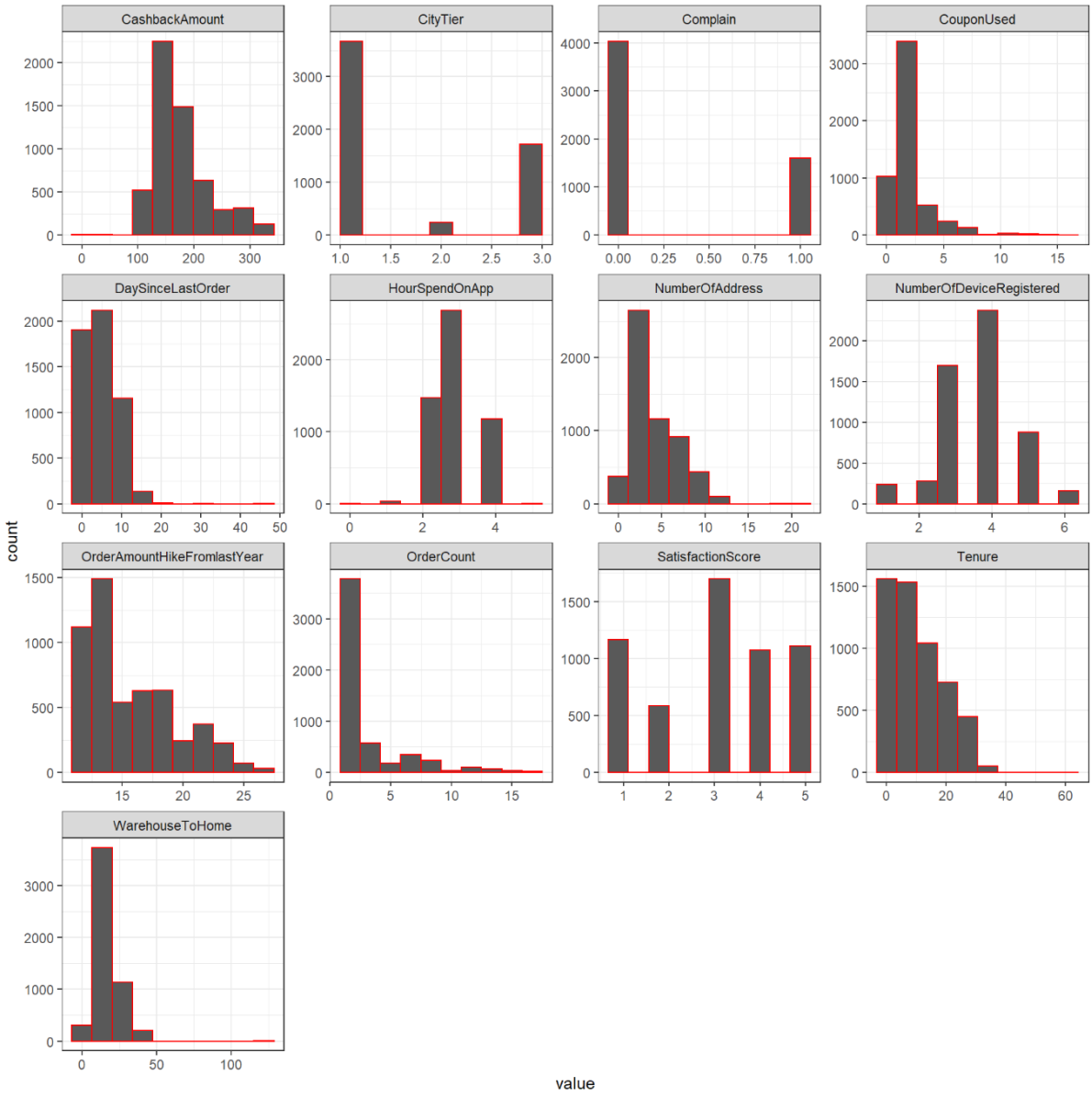

Figure 2: Numerical variables bar plot

Some observations on the visualizations listed above:

1. Most of the columns are right-skewed such as CouponUsed, DaySinceLastOrder and Tenure.

2. Some variables have limited values like CityTier (3 values) and Complain (2 values). Box plots give an indication of the distribution of the values while taking in consideration mean, median and outlier. in the following box plot, numerical attributes were plotted in box plot and description is mentioned as below:

Figure 3: Numerical variables box plot

the above figure shows great difference between churned and retained customers in distribution in the following attributes Complain, SatisfactionScore, Tenure, NumberOfDeviceRegistered.

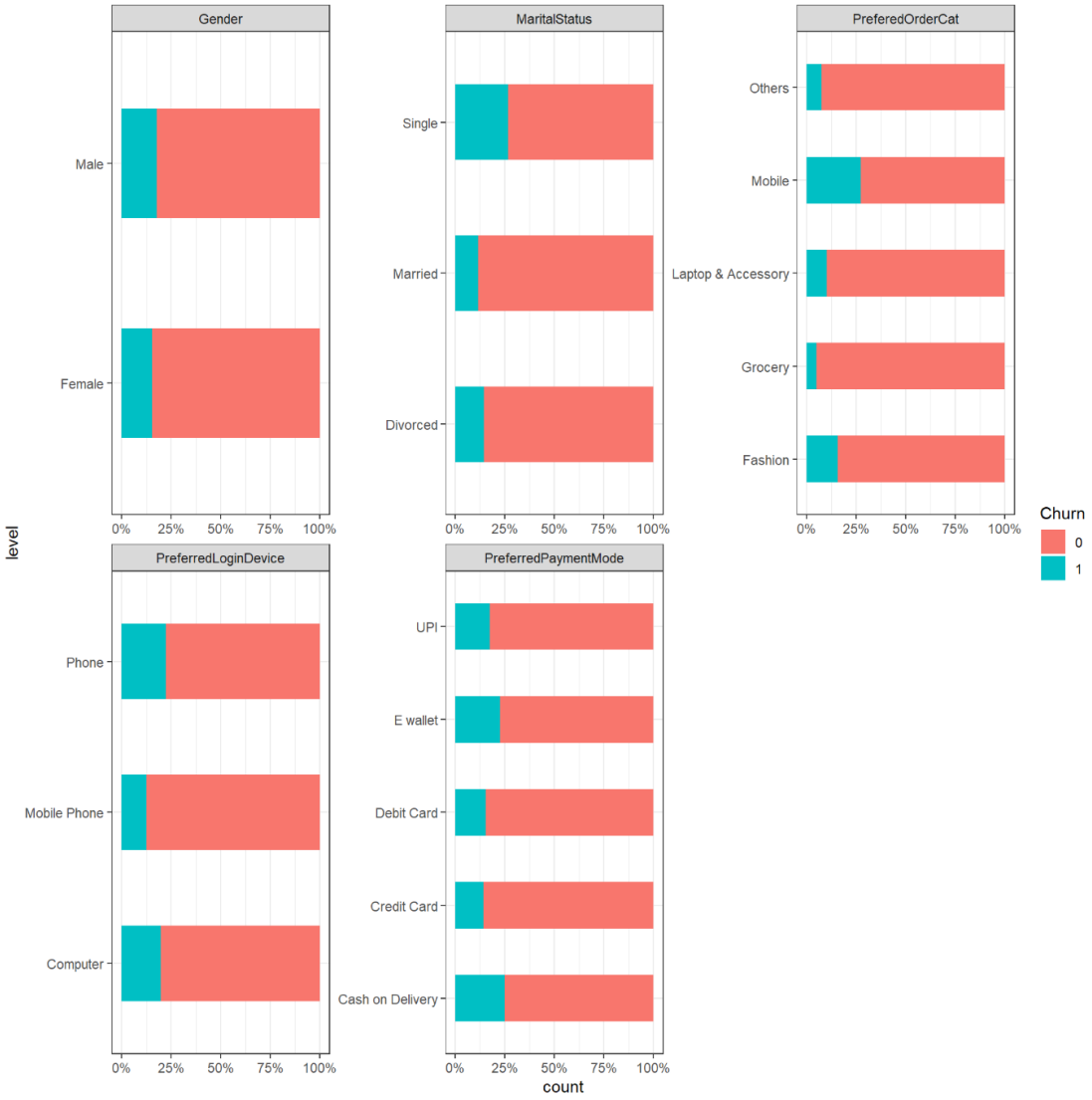

Figure 4: Categorical attributes bar plot

From the above plot we can find out some characteristics of churned customers. Churned customers gender is mostly "Male" and Marital status is "Single". In addition, Churned customers tend to prefer "Mobile phone "for shopping and "cash on delivery" payment method. On the other hand, retained customers prefer "Grocery" Order category the most.

In the following figure, compared the satisfaction score with customer churn to understand the relationship between churn and satisfaction score.

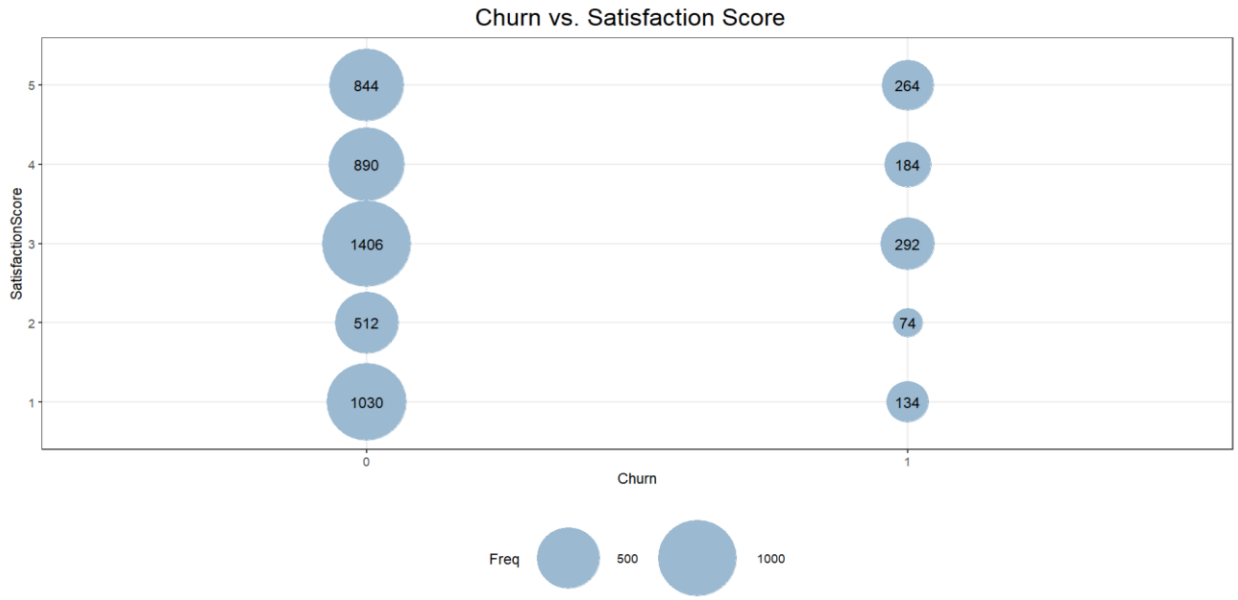

Figure 5: Churn Versus Satisfaction score

the figure above showed remarkable result that contradicts satisfaction score's objective, 80 % of the churned customers gave a satisfaction score from 3 to 5 which is unreasonable. In fact, this outcome is complying with table 4 as it shows higher satisfaction score mean for churners in comparison with retained customers.

4.4. Results – Exploratory Data Analysis

Several exploratory analysis and visualizations were listed to enhance the understanding of the used data set. The most important insights are going to be described in the following paragraph.

The main objective was to distinguish between churned and retained customers in addition to finding the associated attributes leading to churn. At first, it was observed that single male customers are having slightly higher probability of churn. In addition, Mobile preferred order category is related to customer churn as well. Furthermore, the churned customers are slightly higher in phone/Mobile phone preferred login device which might be caused by the E-commerce's customer user experience phone version of the ecommerce. Also, it was found that chumed customers are having higher mean in complain, city tier Number of addresses and number of registered devices. However, our study shows that satisfaction score is higher in churned customers which was not expected. On the other hand, Tenure, and count of number of orders is lower for churned customers which is reasonable.

4.5. Model Building

The following stage consists of several steps that must be taken starting with splitting the data into training and testing data. Then, three machine learning models are going to be built to compare their accuracy. Further description of the selected algorithms is as mentioned below:

1. **Decision tree**

Decision trees are a form of non-supervisory machine learning algorithms which can be used for classification or regression. the scope of decision tree is to create a model that predicts the value of the outcome by learning simple decision rules assumed from data features (Decision Trees, 2022).

2. **Logistic regression**

Logistic regression is a process where probability of discrete outcome is modelled which is usually used for binary outcomes. Logistic regression is widely used on classification problems especially when the aim of the study is to determine if a sample appropriately fitting into a class. and considered as one of the main analytical algorithms (Thomas W. Edgarm, 2017).

3. **Random forest**

Random forest is machine learning algorithm which combines the output of various decision aiming for a single output, it overcomes some overfitting and bias issues associated with decision trees and gives accurate predictions especially when individual trees are uncorrelated with each other (Random Forest, 2020).

Data Preprocessing

Before going deep into building the models, certain steps must be followed to make sure that models are built to give the best performance. As mentioned earlier, the dataset consists of 3774 rows following data cleaning step. Therefore, data must be split into training data and test data. Training data will take 75% of the total dataset which means that it has 2830 rows while test data has 944 rows (25%).

The outcome of the models will be churn, Customer ID column will be removed from predictors column. Furthermore, all categorical predictors will be converted to numerical values. Also, all predicters with zero or low variability will be eliminated.

Two extra steps are going to be carried out exclusively for logistic regression model which are:

- Data normalization for all predictors.
- Highly correlated variables in all predictors are going to be removed.

Finally**,** Synthetic minority oversampling technique (SMOTE) will be applied to balance the levels of Churn column by generating new samples of the minority class (Churned) using nearest neighbors.

After applying SMOTE, training data is currently composed of 4717 rows with equally divided classes of churners and non-churners. Summary of the previous step is shown in the following table.

	Retained Customers	**Churned Customers**
Before SMOTE	2357	473
After SMOTE	2357	2357

Table 7 : Churn Column Before and after SMOTE

4.6. Comparison of Different Models

The result of each algorithm has been analyzed and compared with other algorithms. This gave us the most reliable machine learning algorithm for our case. Comparison was based on accuracy and kappa values.

Accuracy:

Accuracy is the ratio of number of correct predictions and total number of predictions. In our case, as we are dealing with binary classification problem, accuracy will be calculated using the formula below.

$$\text{Accuracy} = \frac{True\ positive(TP)+True\ negative(TN)}{True\ Positive(TP)+True\ negative(TN)+False\ Positive(FP)+False\ negative(FN)}$$

Kappa:

Kappa is a useful metric that can help in overcoming multi-class classification problems in which accuracy measures might be misleading. In addition, kappa is beneficial when dealing with imbalanced classes such as our case. kappa can be computed by the following formula (Landis et al., (1977).

$$\text{kappa} = 1 - \frac{1-Observed\ agreement}{1-Expected\ agreement}$$

$$\text{Observed agreement} = \frac{True\ Positive\ (TP)+True\ Negative(TN)}{Total}$$

$$\text{Expected Agreement} = \frac{TN+FP}{Total} * \frac{TN+FN}{Total} + \frac{FN+TP}{Total} * \frac{FP+TP}{Total}$$

Decision tree:

		Actual		
		No	Yes	**Total**
Prediction	No	691	48	**739**
	Yes	95	110	**205**
	Total	**786**	**158**	**944**

Table 8: Decision tree confusion matrix

Accuracy = $\frac{110+691}{691+48+95+110} = 84.8\ \%$

Observed Agreement = $\frac{691+110}{944} = 0.8485$

Expected Agreement = $\frac{691+95}{944} * \frac{691+48}{944} + \frac{48+110}{944} * \frac{95+110}{944} = 0.688$

kappa = $1 - \frac{1-0.8485}{1-0.688} = 0.514$

The first model that was built is the decision tree and its accuracy and kappa metrics were computed. The accuracy of the model is estimated at 85% while kappa is 0.514.

Logistic regression:

		Actual		
		No	Yes	**Total**
Prediction	No	634	32	**666**
	Yes	152	126	**278**
	Total	**786**	**158**	**944**

Table 9: Logistic regression confusion matrix

Accuracy = $\frac{634+126}{634+32+152+126} = 80.5\%$

Observed Agreement = $\frac{634+126}{944} = 0.805$

Expected Agreement = $\frac{634+152}{944} * \frac{634+32}{944} + \frac{32+126}{944} * \frac{152+126}{944} = 0.6367$

Kappa = $1 - \frac{1-0.805}{1-0.6367} = 0.46$

The second machine learning algorithm used is logistic regression. Logistic regression had slightly lower accuracy and kappa numbers in comparison to decision tree which are 80.5% and 0.46 respectively.

Random forest:

		Actual		
		No	Yes	**Total**
Prediction	No	634	32	**666**
	Yes	152	126	**278**
	Total	**786**	**158**	**944**

Table 10: Random Forest decision matrix

Accuracy = $\frac{767+116}{767+42+19+116} = 93.5\%$

Observed Agreement = $\frac{767+116}{944} = 0.9353$

Expected Agreement = $\frac{767+19}{944} * \frac{767+42}{944} + \frac{42+116}{944} * \frac{19+116}{944} = 0.737$

Kappa = $1 - \frac{1-0.9353}{1-0.737} = 0.75$

The last machine learning algorithm applied for this project is the random forest. As shown in the table above, accuracy is higher than the rest of the models. In addition, significant difference was observed in kappa number at 0.75.

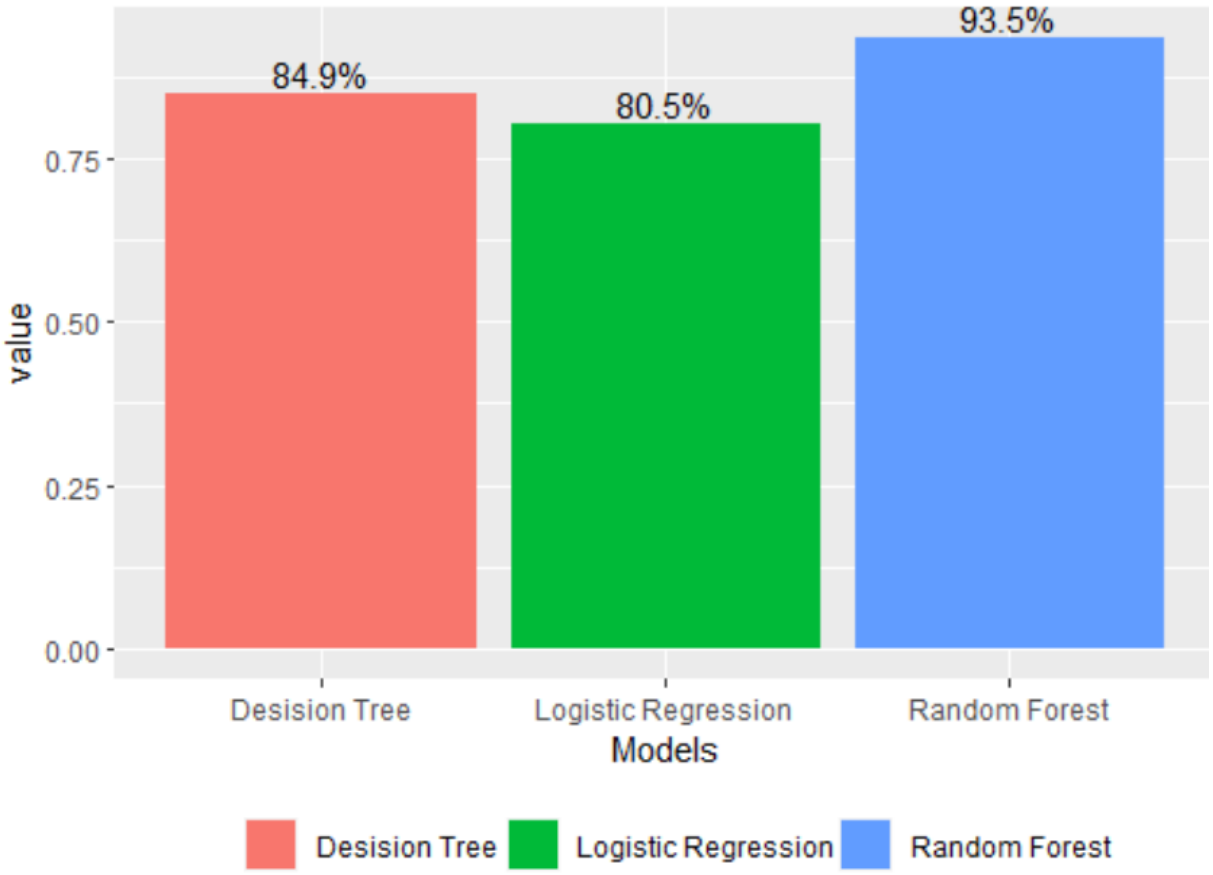

Figure 6: Model's accuracy comparison

In summary, Random Forest showed great performance surpassing decision tree and logistic regression in terms of accuracy. Moreover, Random Forest had approximately 47 % higher kappa number than both decision tree and logistic regression. Hence, Random Forest is the most suitable algorithm machine learning algorithms according to accuracy and kappa metrics. According to Landis et al., (1977), Achieving kappa score greater than 0.60 is considered as substantial result for the model. Thus, random forest model falls into the substantial category.

CHAPTER 5 - CONCLUSION

5.1. 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 foe 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. It was noticed that churned customers associated with male gender, single marital status. Then, three different machine algorithms were applied to predict customer churn which are Decision tree, Logistic regression, and random forest. It was found that random forest has the best accuracy and kappa score at 93.5% and 0.75 respectively.

5.2. Recommendations

Based on the analysis, Random Forest was utilized to find out the significance of each variable. Therefore, the 29 predictors were sorted in descending order according to their importance. It was found that the top 4 predictors are Tenure, Complain, NumberOfAddress, MaritalStatus_Single) which are associated with 90% of the impact on the outcome. Thus, these attributes were proved to be having the greatest influence on customer churn.

Variable	Importance
Tenure	100.0000000
Complain	42.3960392
NumberOfAddress	16.8245874
MaritalStatus_Single	13.0439810
WarehouseToHome	9.6616788
CashbackAmount	9.3012150
MaritalStatus_Married	9.1711003
PreferedOrderCat_Laptop...Accessory	8.3206381
CityTier	8.2065317
SatisfactionScore	8.0448382
DaySinceLastOrder	7.5919938
OrderAmountHikeFromlastYear	6.4732324
NumberOfDeviceRegistered	5.5588406

Table 11: List of most important variables

Some recommendations to business owner from the analysis are listed below:

1. Business must increase the tenure of their customer which can be done by initiating some loyalty programs or special pricings for loyal customers.
2. Since complains is coming in the second place in terms of importance, it must be handled carefully, and the organization must ensure that its customer service is qualified to deal with complains professionally.
3. The organization must eliminate the root cause of complains by enhancing the customer experience and conduct some surveys to get the customer's feedback. In fact, getting user's opinion and enhancing their experience will add a lot of value for the company.

4. Conduct A/B testing to enhance user experience and user interface which will reflect positively on the conversion rate.

5.3. Future Work

In future projects, I would like to build real time analysis for a local e-commerce platform and link it to mail marketing software. This integration will enable organizations to automate their offers with churners which is certainly going to minimise customer attrition. Also, I would like to go in depth in retained customers behaviour and most preferred goods. Hence, studying retained customer behaviour will reflect greatly on the company's income.

BIBLIOGRAPHY

1. Zhang, D. (2015). Establishment and application of customer churn prediction model. Beijing Institute of Technology.
2. Saghir, M., Bibi, Z., Bashir, S., & Khan, F. H. (2019, January). Churn prediction using neural network-based individual and ensemble models. In 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST) (pp. 634-639). IEEE.
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. Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. Procedia Cirp, 79, 403-408.
5. Shao, D. (2016). Analysis and prediction of insurance company's customer loss based on BP neural network. Lanzhou University
6. 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.
7. Huang, J. (2018). A Comparative Study of Social E-Commerce and Traditional E-commerce. Economic and Trade Practice, (23), 188-189.
8. Feng, X., Wang, C., Liu, Y., Yang, Y., & An, H. G. (2018). Research on customer churn prediction based on comment emotional tendency and neural network. Journal of China Academy of Electronics Science, 13(03), 340-345
9. 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.
10. Dhote, S., Vichoray, C., Pais, R., Baskar, S., & Shakeel, P. M. (2020). Hybrid geometric sampling and AdaBoost based deep learning approach for data imbalance in E-commerce. Electronic Commerce Research, 20(2), 259-274.
11. Agrawal, S., Das, A., Gaikwad, A., & Dhage, S. (2018, July). Customer churn prediction modelling based on behavioural patterns analysis using deep learning. In

2018 International conference on smart computing and electronic enterprise (ICSCEE) (pp. 1-6). IEEE.

12. Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. IEEE Access.
13. Ho, T. K. (1995). Random decisions forest. Proceedings of 3rd International Conference on Document Analysis and Recognition (pp. 278-282). New Jersey: IEEE.
14. Geetha, V., Punitha, A., Nandhini, A., Nandhini, T., Shakila, S. and Sushmitha, R., 2020, July. Customer Churn Prediction In Telecommunication Industry Using Random Forest Classifier. In *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1-5). IEEE.
15. Ullah, I., Raza, B., Malik, A.K., Imran, M., Islam, S.U. and Kim, S.W., 2019. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE Access*, *7*, pp.60134-60149.
16. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. In: European conference on computational learning theory. Heidelberg: Springer; 1995. p. 23–37.
17. Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310.
18. Decision Trees. (2022, 04 22). Retrieved from Sickit learn: https://scikit-learn.org/stable/modules/tree.html#:~:text=Decision%20Trees%20(DTs)%20are%20a,as%20a%20piecewise%20constant%20approximation.
19. Thomas W. Edgarm, D. O. (2017). Research Methods of Cyber Security.
20. Random Forest. (2020, December 7). Retrieved from IBM: https://www.ibm.com/cloud/learn/random-forest#:~:text=%20What%20is%20random%20forest%3F%20%201%20Decision,bagging%20method%20as%20it%20utilizes%20both...%20More%20