E-Commerce Customer Churn Analysis and Prediction

By:

Priyanka Bagchi
Priyanka.Bagchi.pb@gmail.com

Supervisor: Dr. Sedef Akinli Kocak, PhD
sedef.akinlikocak@ryerson.ca

Ryerson University

Table of Contents

Introduction

Customer churn was, and still is, a very important concept in contemporary marketing that should not be ignored (Jahromi, Stakhovych, & Ewing, 2014). Customer churn or customer attrition refers to the loss of customers. It is the percentage of customers that have stopped using a company's product or service over a certain period of time. Lost customers also mean lost revenue, which is why it is so important for companies to know in advance which customers will churn in the near future.

Many small and relatively new small businesses are struggling due to the economic effects of COVID-19 and are wanting to find new/innovative ways to hold onto their loyal customer. With data science being such a hot topic, they might want to use it to help them with their decision-making. It is best to use machine learning with large datasets, which these companies might not have.

The focus of this data analysis project will be to predict customer churn of an up-and-coming e-commerce company that has a relatively small dataset. They want to use this analysis/prediction to plan what incentives and/or other retention offers to use to prevent this from happening. This will need to be done keeping the chances of overfitting in mind and what can be done if such happens.

For predictive modelling, three algorithms will be used to predict which customers will churn in the near future:

- Logistic Regression
- Random Forest
- Decision Tree

Literature Review

Many companies, from pre-pandemic times, were starting to realize that they should focus more on retaining their current customers while attracting new ones. For this, customers that are about to leave/churn need to be identified so that they can be targeted with tailored incentives (points, discounts, etc.) or other retention offers (coupons, etc.) (Jahromi, Stakhovych, & Ewing, 2014).

With so many companies losing customers during the COVID-19 pandemic due to economic concerns, such as high unemployment rate (Ranchhod & Daniels, 2021), more of such analysis and predictions are needed to keep current customers to be able to walk towards the path of recovery (Mulcahy, 2020). With many countries being in lockdown and preventing consumers to shop in store, all shopping is being operated digitally.

Many businesses have seen significant increases during the COVID-19 pandemic, along with a high churn rate (Rachmawati, 2021). This is possibly because searching and comparing products/offers/deals at competing stores simultaneously is much easier to do online. Loyal customers that would not have bothered or were not able to check what competitors are offering pre-pandemic, are doing so now. These are also the customers that produce higher revenue and margins than new customers. This makes it even more crucial to understand loyal customers and prevent them from churning by developing innovative marketing strategies and improve customer satisfaction (Cao, Yu, & Zhang, 2015).

Dataset

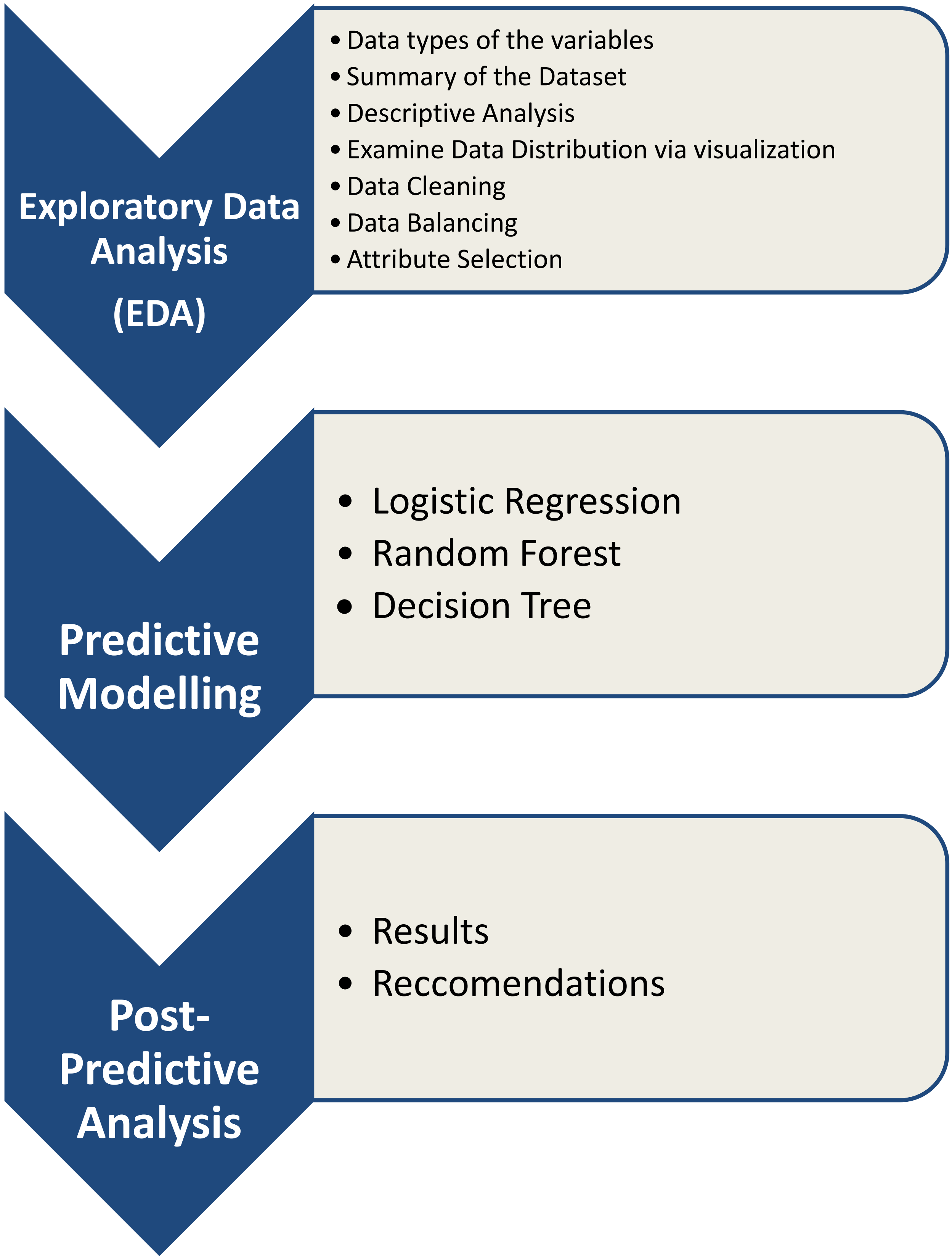
The dataset used for this project was curated by Ankit Verma and obtained from Kaggle. This small sample of 5630 rows was taken from the database of an e-commerce company sometime during 2019 (when the company was approx. 4 years old). Each row representing a separate customer. It was then modified by the curator to give it the current shape. The dataset has information on a wide range of customers including, but not limited to, those who have been customers since the inception of the company to the ones recently acquired.

There are 20 unique numeric, qualitative and binary variables within this dataset:

1. **CustomerID:** Unique customer ID's
2. **Churn:** Churn class attribute with binary values (1 for churn and 0 for not churn)
3. **Tenure:** Tenure of customer in organization
4. **PreferredLoginDevice:** Preferred login device
5. **CityTier:** City tier
6. **WarehouseToHome:** Distance in between warehouse to home of customer
7. **PreferredPaymentMode:** Preferred payment method
8. **Gender:** Gender of customer
9. **HourSpendOnApp:** Number of hours spend on mobile application or website
10. **NumberOfDeviceRegistered:** Total number of devices registered per customer
11. **PreferedOrderCat:** Preferred order category of customer in the previous month
12. **SatisfactionScore:** Satisfactory score of customer on service
13. **MaritalStatus:** Marital status of customer
14. **NumberOfAddress:** Total number of addresses
15. **Complain:** Any complaint raised in the previous month
16. **OrderAmountHikeFromlastYear:** Percentage increase in order from last year
17. **CouponUsed:** Total number of coupons used in the previous month
18. **OrderCount:** Total number of orders placed in the previous month
19. **DaySinceLastOrder:** Days since last order
20. **CashbackAmount:** Average cashback in the previous month

I am interested in analyzing the "Churn" attribute given "1", which identifies those customers that have churned to try to predict what lead them to churn and why.

Approach


Exploratory Data Analysis
(EDA)
Data types of the variables
Summary of the Dataset
Descriptive Analysis
Examine Data Distribution via visualization
Data Cleaning
Data Balancing
Attribute Selection
Predictive Modelling
Logistic Regression
Random Forest
Decision Tree
Post-Predictive Analysis
Results
Reccomendations

[Link to my Github]

Step 1: Exploratory Data Analysis (EDA)

Data types of the Variables

```
CustomerID                 : int [1:5630] 50001 50002 50003 50004 50005 50006 50007 50008 50009 50010 ...
Churn                      : num [1:5630] 1 1 1 1 1 1 1 1 1 1 ...
Tenure                     : num [1:5630] 4 NA NA 0 0 0 NA NA 13 NA ...
PreferredLoginDevice       : chr [1:5630] "Mobile Phone" "Phone" "Phone" "Phone" ...
CityTier                   : num [1:5630] 3 1 1 3 1 1 3 1 3 1 ...
WarehouseToHome            : num [1:5630] 6 8 30 15 12 22 11 6 9 31 ...
PreferredPaymentMode       : chr [1:5630] "Debit Card" "UPI" "Debit Card" "Debit Card" ...
Gender                     : chr [1:5630] "Female" "Male" "Male" "Male" ...
HourSpendOnApp             : num [1:5630] 3 3 2 2 NA 3 2 3 NA 2 ...
NumberOfDeviceRegistered   : num [1:5630] 3 4 4 4 3 5 3 3 4 5 ...
PreferedOrderCat           : chr [1:5630] "Laptop & Accessory" "Mobile" "Mobile" "Laptop & Accessory" ...
SatisfactionScore          : num [1:5630] 2 3 3 5 5 5 2 2 3 3 ...
MaritalStatus              : chr [1:5630] "Single" "Single" "Single" "Single" ...
NumberOfAddress            : num [1:5630] 9 7 6 8 3 2 4 3 2 2 ...
Complain                   : num [1:5630] 1 1 1 0 0 1 0 1 1 0 ...
OrderAmountHikeFromlastYear: num [1:5630] 11 15 14 23 11 22 14 16 14 12 ...
CouponUsed                 : num [1:5630] 1 0 0 0 1 4 0 2 0 1 ...
OrderCount                 : num [1:5630] 1 1 1 1 1 6 1 2 1 1 ...
DaySinceLastOrder          : num [1:5630] 5 0 3 3 3 7 0 0 2 1 ...
CashbackAmount             : num [1:5630] 160 121 120 134 130 ...
```

Figure 1

Summary of the Dataset

```
   CustomerID        Churn             Tenure      PreferredLoginDevice    CityTier      WarehouseToHome
 Min.   :50001   Min.   :0.0000   Min.   : 0.00   Length:5630          Min.   :1.000   Min.   :  5.00
 1st Qu.:51408   1st Qu.:0.0000   1st Qu.: 2.00   Class :character     1st Qu.:1.000   1st Qu.:  9.00
 Median :52816   Median :0.0000   Median : 9.00   Mode  :character     Median :1.000   Median : 14.00
 Mean   :52816   Mean   :0.1684   Mean   :10.19                        Mean   :1.655   Mean   : 15.64
 3rd Qu.:54223   3rd Qu.:0.0000   3rd Qu.:16.00                        3rd Qu.:3.000   3rd Qu.: 20.00
 Max.   :55630   Max.   :1.0000   Max.   :61.00                        Max.   :3.000   Max.   :127.00
                                  NA's   :264                                          NA's   :251
 PreferredPaymentMode    Gender          HourSpendOnApp  NumberOfDeviceRegistered PreferedOrderCat
 Length:5630          Length:5630        Min.   :0.000   Min.   :1.000            Length:5630
 Class :character     Class :character   1st Qu.:2.000   1st Qu.:3.000            Class :character
 Mode  :character     Mode  :character   Median :3.000   Median :4.000            Mode  :character
                                         Mean   :2.932   Mean   :3.689
                                         3rd Qu.:3.000   3rd Qu.:4.000
                                         Max.   :5.000   Max.   :6.000
                                         NA's   :255
 SatisfactionScore MaritalStatus      NumberOfAddress     Complain      OrderAmountHikeFromlastYear
 Min.   :1.000     Length:5630        Min.   : 1.000   Min.   :0.0000   Min.   :11.00
 1st Qu.:2.000     Class :character   1st Qu.: 2.000   1st Qu.:0.0000   1st Qu.:13.00
 Median :3.000     Mode  :character   Median : 3.000   Median :0.0000   Median :15.00
 Mean   :3.067                        Mean   : 4.214   Mean   :0.2849   Mean   :15.71
 3rd Qu.:4.000                        3rd Qu.: 6.000   3rd Qu.:1.0000   3rd Qu.:18.00
 Max.   :5.000                        Max.   :22.000   Max.   :1.0000   Max.   :26.00
                                                                        NA's   :265
   CouponUsed        OrderCount     DaySinceLastOrder CashbackAmount
 Min.   : 0.000   Min.   : 1.000   Min.   : 0.000   Min.   :  0.0
 1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.:145.8
 Median : 1.000   Median : 2.000   Median : 3.000   Median :163.3
 Mean   : 1.751   Mean   : 3.008   Mean   : 4.543   Mean   :177.2
 3rd Qu.: 2.000   3rd Qu.: 3.000   3rd Qu.: 7.000   3rd Qu.:196.4
 Max.   :16.000   Max.   :16.000   Max.   :46.000   Max.   :325.0
 NA's   :256      NA's   :258      NA's   :307
```

Figure 2

Examine Data Distribution

Box plots to show outliers, minimum value, lower quartile (Q1), median value (Q2), upper quartile (Q3), and maximum value in the data set.

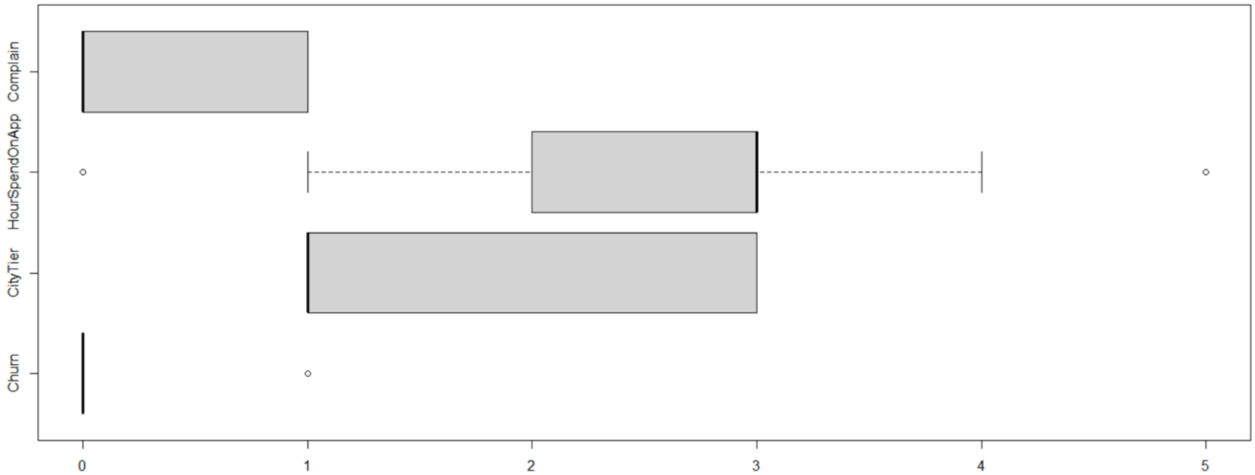


Figure 3

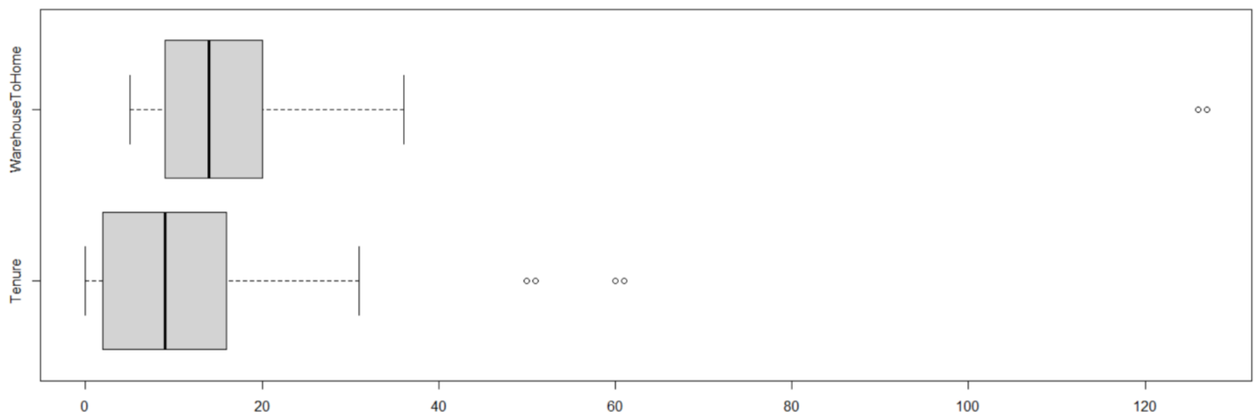


Figure 4

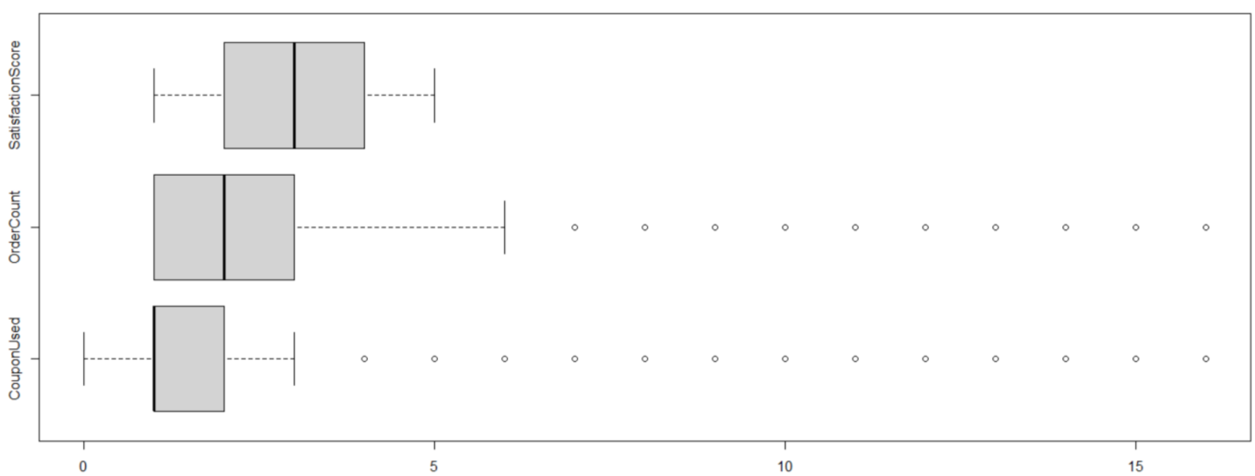


Figure 5

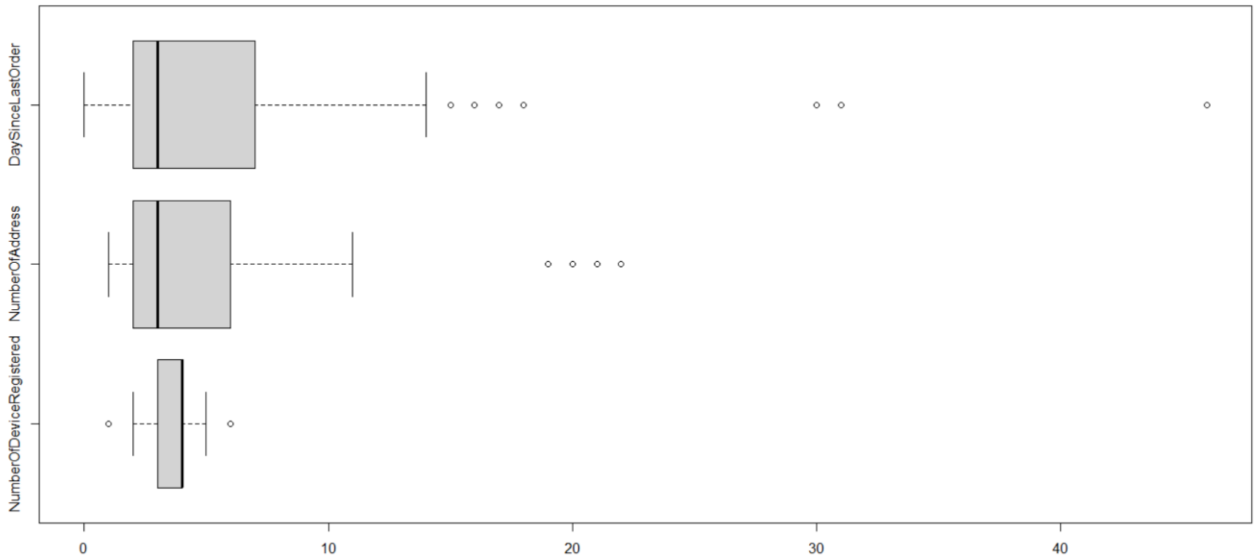
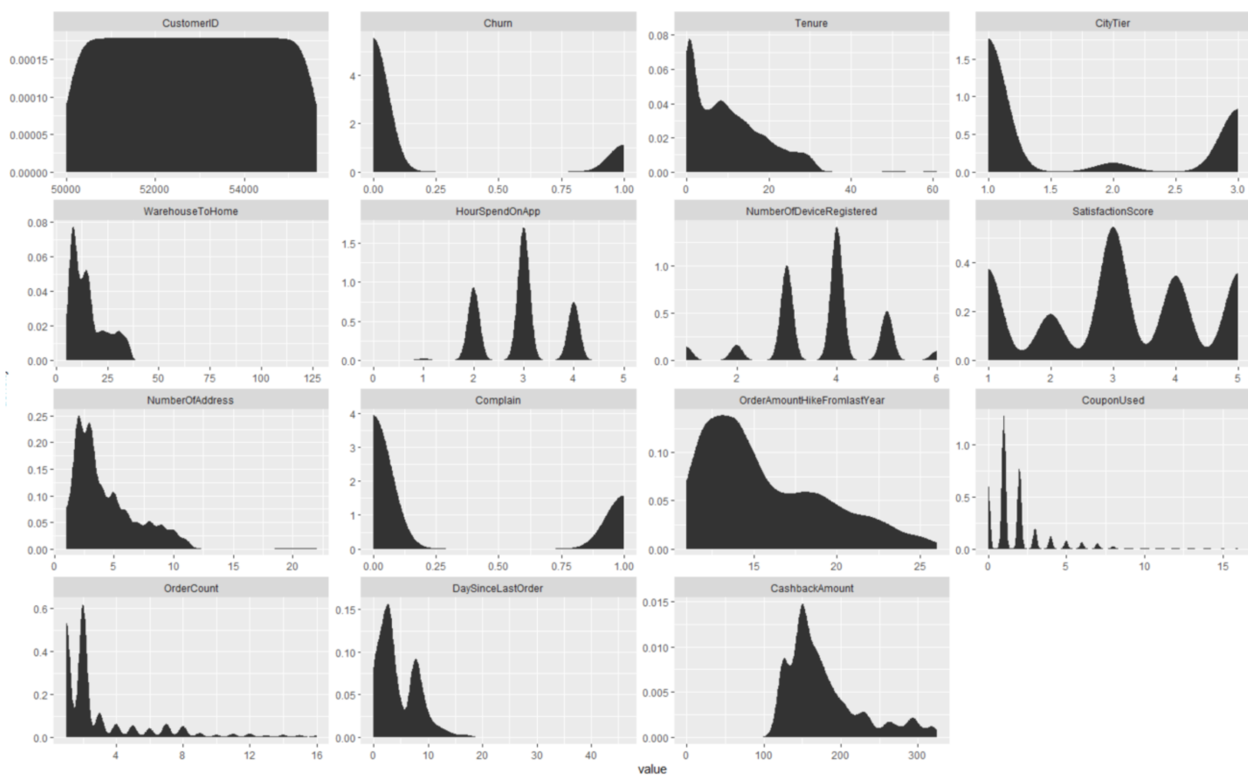


Figure 6

Density plots to see the Distribution of all the Variables



Frequency table of all the columns that have character data types

```
Character_Data %>% count(PreferredLoginDevice)
A tibble: 3 x 2
PreferredLoginDevice     n
<chr>                <int>
Computer              1634
Mobile Phone          2765
Phone                 1231
Character_Data %>% count(PreferredPaymentMode)
A tibble: 7 x 2
PreferredPaymentMode     n
<chr>                <int>
Cash on Delivery       149
CC                     273
COD                    365
Credit Card           1501
Debit Card            2314
E wallet               614
UPI                    414
Character_Data %>% count(Gender)
A tibble: 2 x 2
Gender     n
<chr>  <int>
Female  2246
Male    3384
Character_Data %>% count(PreferedOrderCat)
A tibble: 6 x 2
PreferedOrderCat       n
<chr>              <int>
Fashion              826
Grocery              410
Laptop & Accessory  2050
Mobile               809
Mobile Phone        1271
Others               264
Character_Data %>% count(MaritalStatus)
A tibble: 3 x 2
MaritalStatus     n
<chr>         <int>
Divorced        848
Married        2986
Single         1796
```

Figure 7

Total number of missing values in each column

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

Figure 8

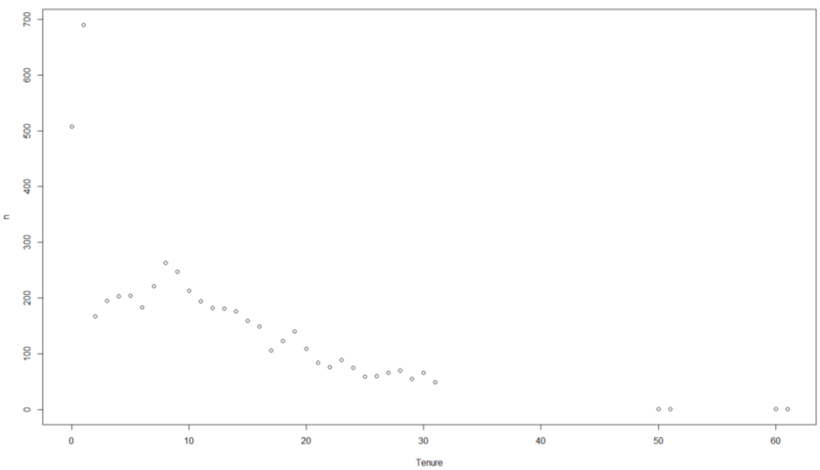
Data Cleaning: Character Variables

Since the following categories have the same meaning, they will be cleaned by choosing one of the two:

1) Mobile Phone and Phone: Mobile Phone
2) CC and Credit Card: Credit Card
3) Cash on Delivery and COD: Cash on Delivery

Data Cleaning: Numeric Variables

When trying to remove all rows with missing values, only 3774 rows remained and 1856 rows were removed. With the dataset already being too small and skewed, I decided to clean it and keep as many of the rows as possible, if not all, by replacing missing values with the median value.



	Tenure	n
1	0	508
2	1	690
3	2	167
4	3	195
5	4	203
6	5	204
7	6	183
8	7	221
9	8	263
10	9	247
11	10	213
12	11	194
13	12	182
14	13	181
15	14	176
16	15	159
17	16	149
18	17	106
19	18	123
20	19	140
21	20	109
22	21	84
23	22	76
24	23	89
25	24	75
26	25	59
27	26	60
28	27	66
29	28	70
30	29	55
31	30	66
32	31	49
33	50	1
34	51	1
35	60	1
36	61	1
37	NA	264

Figure 9

For Tenure, which is the tenure of each customer with the organization since its inception. I decided to fill in the N/A's with 0's, keeping in mind that the company has only been around for 4 years and the tenure of majority of its customers are 1 year or less.

Density plots to see the Distribution of all the Variables after Cleaning

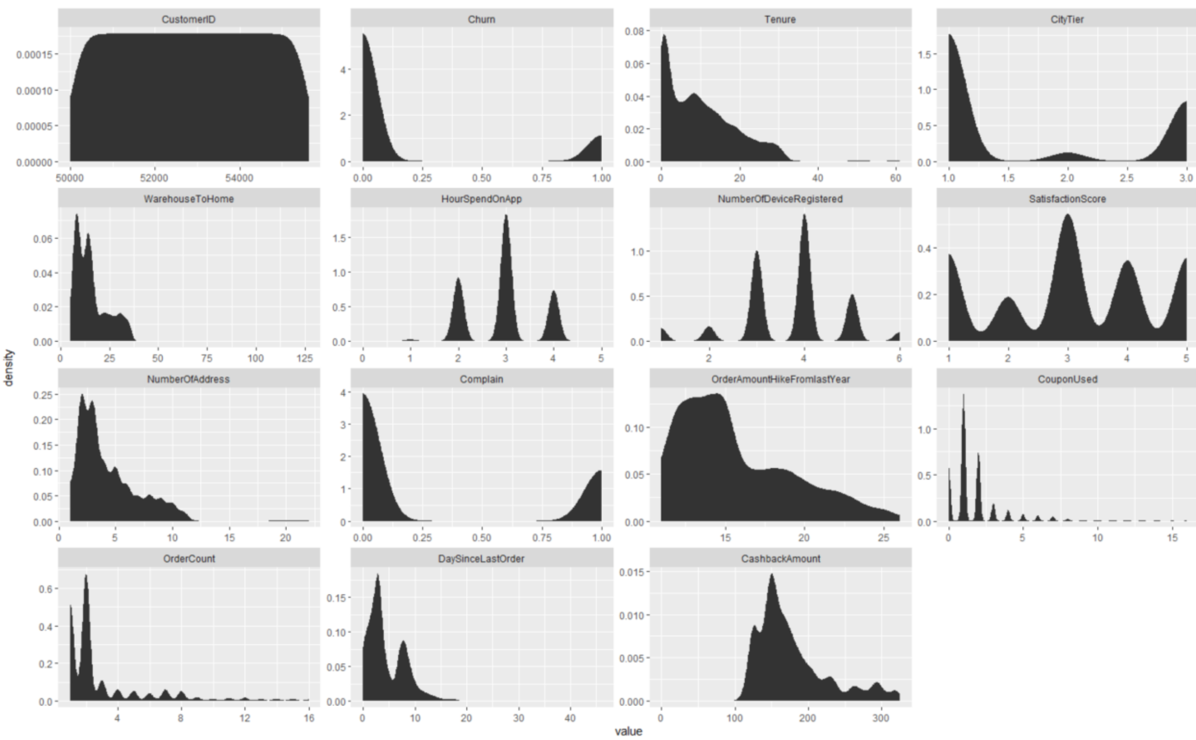


Figure 10

An known imbalance in the target class (Churn) can be seen.

Data Cleaning: Removal of Outliers

CustomerID is removed as it does not provide any information.

Variables that have outliers:

HourSpendOnApp: The 3 customers who spent 5 hours on the app or website did not churn and will not help in predicting if customers are leaving after spending hours on the app or website. So they are removed.

WarehouseToHome: The 2 customers who live the furthest from the warehouse did not churn, so this will not help in predicting if distance from the warehouse could be a reason for customers leaving. So they are removed.

Tenure: The 4 customers who have been with the company for over 50 months have not churned. This will not help in predicting why such loyal customers are leaving. Due to the company being only 4 years old and 50 – 60 months being 4.2 – 5 years, these accounts could be belonging to founders, employees or even test accounts belonging to the company. So they are removed.

OrderCount: There are quite a few customers who churned even after placing many orders. Only the customers who churned and are OrderCount outliers are kept, so we can try to predict what is causing customers to leave even after placing many orders.

CouponUsed: There are quite a few customers who churned after using 4 or more coupons. Only the customers who churned and are CouponUsed outliers are kept, so we can try to predict what is causing customers to leave even after using so many coupons. This will also help us to predict if customers are only purchasing from this company when they have a coupon and leaving when there are none available for a long period of time.

NumberOfAddress: 2 out of the 4 outlier customers who had bought from (or sent to) 19 or more addresses, churned. Neither one of them are the tenure outliers. The 2 NumberOfAddress outliers that churned are kept to help us predict why they churned after buying from or sending to so many addresses.

DaySinceLastOrder: Only 1 of the outlier customers churned, so this will not help us to predict a customer who is about to churn. So they are removed.

NumberOfDeviceRegistered: The outlier value 6 is not too far from the upper quartile of 4 only the outliers that churned are kept.

By keeping some of the outlier customers that churned, this also helped to slightly balance the dataset.

Correlation Between all Numerical Variables

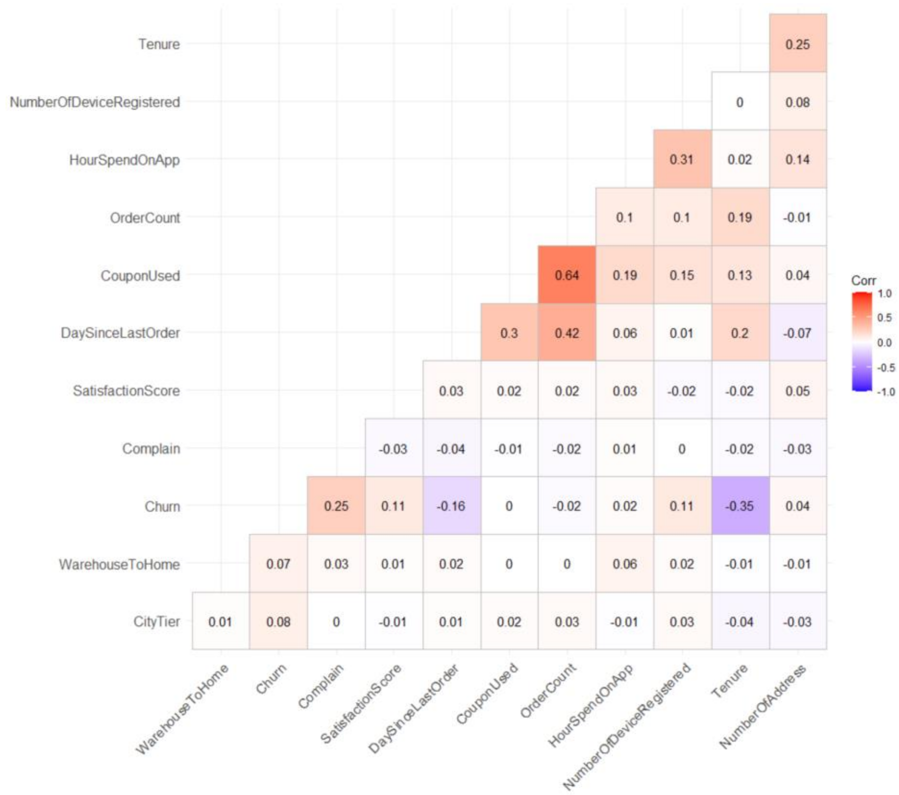


Figure 11

Feature Selection

Recursive Feature Elimination (RFE) along with the top 10 important variables predicted by the 3 models was used to select the final top 10 variables.

RFE is a commonly used feature selection algorithm. It is so popular because it is easy to configure and use. Also, because it is quite effective at selecting variables in a training dataset that are more or most relevant in predicting the target variable (Brownlee, Recursive Feature Elimination (RFE) for Feature Selection in Python, 2020).

The top 10 variables selected:

1. Churn
2. Tenure
3. OrderCount
4. Complain
5. NumberOfAddress

6. CashbackAmount
7. DaySinceLastOrder
8. NumberOfDeviceRegistered
9. WarehouseToHome
10. CityTier

Step 2: Predictive Modelling

Predictive modeling process of taking known results to create, process, and validate 3 models used to predict future outcomes.

- Logistic Regression
- Random Forest
- Decision Tree

Since the dataset does not have a balanced number for Churn (the target class). The dataset was split into train and test sets in a way that preserved the same proportions of examples in each target class as observed in the original dataset by using stratified train-test split (Brownlee, Train-Test Split for Evaluating Machine Learning Algorithms, 2020).

Logistic Regression

Logistic Regression has been used in the biological sciences since the early twentieth century. It has been used in many social science applications as well. It is the go-to method for binary classification problems (problems with two class values) (Brownlee, Logistic Regression for Machine Learning, 2016) and when the dependent variable (target) is categorical (Swaminathan, 2018). Logistic Regression is used to find what the probability is that a new instance belongs to a certain class. Since it is a probability, the outcome lies between 0 and 1. To use it as a binary classifier, a threshold needs to be assigned to differentiate the two classes. For example, a probability value higher than 0.50 for an input instance will classify it as ‘target class A’; otherwise, ‘target class B’. This generalized version of Logistic Regression is known as the multinomial logistic regression (Uddin, Khan, Hossain, & Moni , 2019).

For example,

- When predicting whether an email is spam (1) or (0)
- Or if the tumour is malignant (1) or not (0)
- And for this paper, predicting customer churn (1) or (0)

Random Forest

Random forest is a flexible, easy to use supervised machine learning algorithm that produces great results (even without hyper-parameter tuning). Due to its simplicity and diversity, it is one of the most commonly used algorithms and can be used for both classification and regression tasks. The "forest" it builds, is a collection of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forests are also hard to beat performance wise. (Donges, 2019).

The out-of-bag (oob) error estimate

“In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run, as follows:

Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the kth tree.

Put each case left out in the construction of the kth tree down the kth tree to get a classification. In this way, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, take j to be the class that got most of the votes every time case n was oob. The proportion of times that j is not equal to the true class of n averaged over all cases is the oob error estimate. This has proven to be unbiased in many tests.” (Breiman & Cutler) So cross-validation techniques will not be used for the random forest model.

Decision Tree

Decision tree (DT) is one of the earliest and prominent machine learning algorithms (Uddin, Khan, Hossain, & Moni , 2019). The Decision tree model is a non-parametric supervised learning method that develops classification systems that predict or classify future observations based on a set of decision rules (1.10. Decision Trees, n.d.). In its rule this process automatically only includes the variables that really matter in making a decision and the variables that do not contribute to the accuracy of the tree are ignored. This can extract very useful information about the data and can be used to reduce the data to relevant variables before training other learning techniques (IBM, 2017).

The models will be compared by:

1) Unbalanced Training Set Before Feature Selection
2) Unbalanced Training Set After Feature Selection
3) Balanced Training Set Before Feature Selection
4) Balanced Training Set After Feature Selection

Metrics used to Evaluate the 3 Models

As per the literature review, these are some of the best evaluation metrics to use for a dataset with an imbalanced target class.

ROC

An ROC is one of the fundamental tools for diagnostic test evaluation and is created by plotting the true positive rate against the false positive rate at various threshold settings (Fawcett, 2008).

AUC

The area under the ROC curve (AUC) is also quite commonly used to determine the predictability of a model. It is used for binary classification problems. High AUC value represents the superiority of a model and vice versa (Uddin, Khan, Hossain, & Moni , 2019).

F1

The F1 Score is used to measure a model's accuracy. It is the harmonic mean between precision and recall. The range of the F1 Score is [0, 1] and tells you how precise a model is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). The F1 Score tries to find the balance between precision and recall (Mishra, 2018). Due to an imbalanced dataset with a small positive target class, it makes more sense to use the F1 score.

(F1 = 2*Precision*Recall/(Precision + Recall)

Accuracy

Accuracy is the percentage of total items classified correctly overall (Swalin, 2018). Due to the imbalanced target class, accuracy was not the right metric to evaluate the models trained on this dataset. It will still be looked at and recorded for learning purposes.

(TP+TN)/(N+P)

Sensitivity

Sensitivity is the percentage of items correctly identified as positive out of total true positives (Swalin, 2018).

TP/(TP+FN)

Specificity

Specificity is the percentage of items correctly identified as negative out of total negatives (Swalin, 2018).

TN/(TN+FP)

Step 3: Post-Predictive Analysis

Results

Due to the dataset being relatively small, all the models are overfitting. Cross-validation, was used to validate the stability and generalizability of the Logistic Regression model. This helped to improve only the AUC results in all the 4 comparisons.

Unbalanced Training Set Before Feature Selection							
	AUC	ROC	Sensitivity	Specificity	Accuracy	F1 Score	Used Pythagoras Theorem sqrt((1-sensitivity)² +(1-specificity) ²) for ROC
Logistic Regression	0.91	0.93	0.62	0.94	0.88	0.91	
Logistic Regression 3 CV	0.78	0.93	0.62	0.94	0.88	0.91	
Logistic Regression 10 CV	0.78	0.93	0.62	0.94	0.88	0.91	
Random Forest	0.93	0.53	0.87	0.99	0.97	0.98	
Decision Tree	0.80	0.89	0.65	0.95	0.89	0.93	

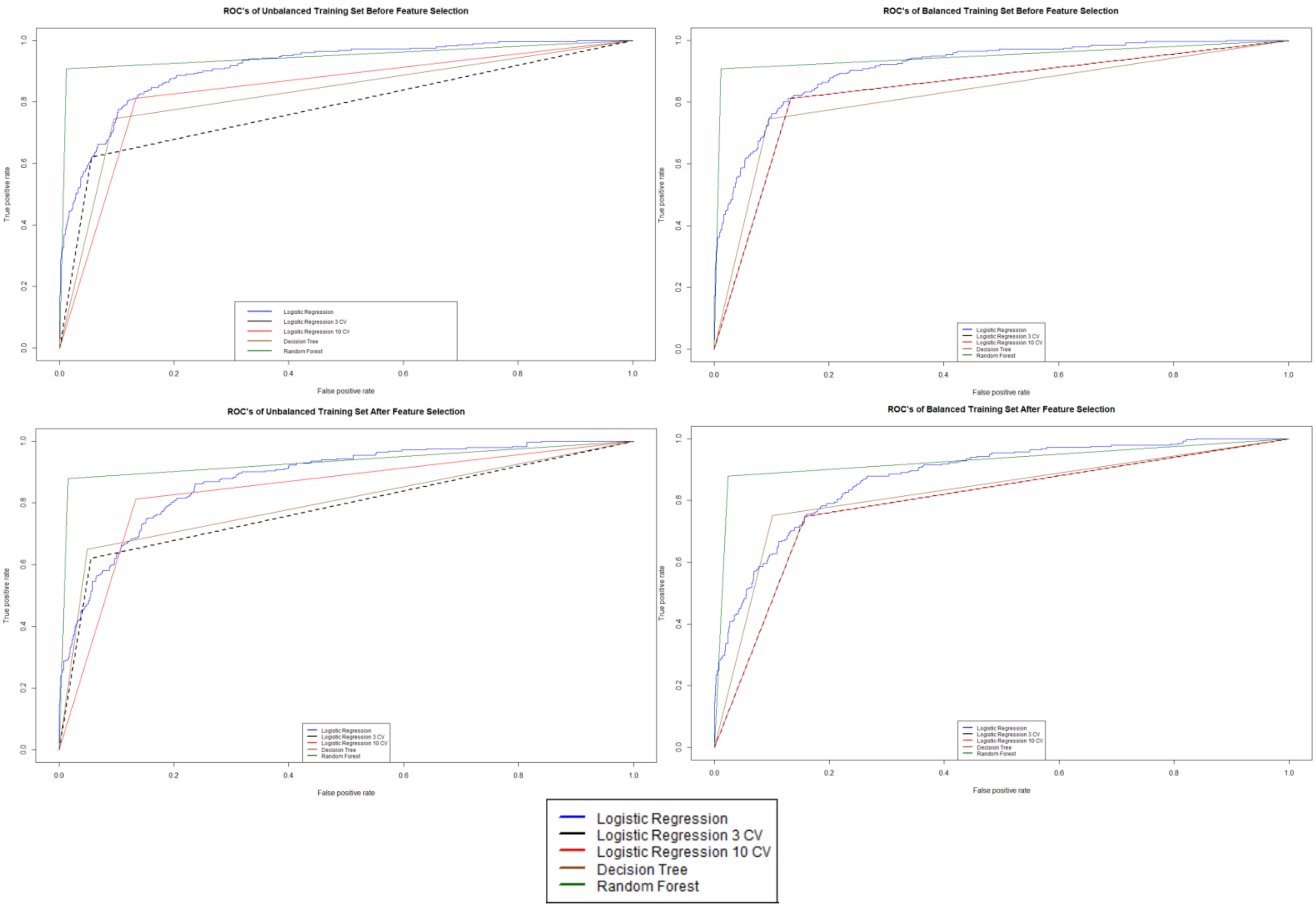
Unbalanced Training Set After Feature Selection							
	AUC	ROC	Sensitivity	Specificity	Accuracy	F1 Score	Used Pythagoras Theorem sqrt((1-sensitivity)² +(1-specificity) ²) for ROC
Logistic Regression	0.88	1.04	0.51	0.94	0.85	0.91	
Logistic Regression 3 CV	0.73	1.04	0.51	0.94	0.85	0.92	
Logistic Regression 10 CV	0.73	1.04	0.51	0.94	0.85	0.92	
Random Forest	0.93	0.53	0.87	0.99	0.96	0.98	
Decision Tree	0.80	0.89	0.65	0.95	0.89	0.93	

Due to the target class being unbalanced, oversampling was used to balance the train set to see if it helps with the overfitting and improve the performance of the 3 models.

Balanced Training Set Before Feature Selection							
	AUC	ROC	Sensitivity	Specificity	Accuracy	F1 Score	Used Pythagoras Theorem sqrt((1-sensitivity)² +(1-specificity)²) for ROC
Logistic Regression	0.91	0.80	0.87	0.81	0.86	0.90	
Logistic Regression 3 CV	0.84	0.80	0.81	0.87	0.86	0.90	
Logistic Regression 10 CV	0.84	0.80	0.81	0.87	0.86	0.90	
Random Forest	0.95	0.46	0.91	0.99	0.97	0.98	
Decision Tree	0.82	0.84	0.74	0.91	0.87	0.92	

Balanced Training Set After Feature Selection							
	AUC	ROC	Sensitivity	Specificity	Accuracy	F1 Score	Used Pythagoras Theorem sqrt((1-sensitivity)² +(1-specificity)²) for ROC
Logistic Regression	0.88	0.91	0.84	0.75	0.82	0.88	
Logistic Regression 3 CV	0.80	0.91	0.75	0.84	0.82	0.88	
Logistic Regression 10 CV	0.80	0.91	0.75	0.84	0.82	0.88	
Random Forest	0.93	0.54	0.88	0.98	0.96	0.97	
Decision Tree	0.83	0.84	0.75	0.90	0.87	0.91	

Oversampling on its own and along with feature selection seem to have helped improve the performance of all 3 models.



Recommendations

Seeing the results its evident that continuing to collect more data is the best option for this e-commerce company if they want to use machine learning models to predict customer churn. This will help in fixing the overfitting issue and improve the performance of the models. Consulting with a more experienced data scientist is also recommended to be able to troubleshoot error messages in a timely manner.

If they still want to use the current predictions, the top 8 variables to keep in mind would be:

1. Tenure
2. OrderCount
3. Complain
4. NumberOfAddress

5. CashbackAmount
6. DaySinceLastOrder
7. NumberOfDeviceRegistered
8. WarehouseToHome

Conclusion

If small and relatively new companies such as this up-and-coming e-commerce company want to take advantage of predictive analytics they will need to ensure that they have a big enough dataset to do so. As seen in the results, companies with small datasets will face the issue of overfitting. Since collecting more data was not possible for the scope of this project oversampling to balance the training set and feature selection was used to improve performance of the models and prove this point. Due to error messages that were taking too long to troubleshoot, time restrictions and the scope of this course:

- The planned use of the XGBoost model was replaced with Decision Tree
- Simple oversampling was used instead of Safe-Level-SMOTE
- Automatic feature selection method and Variable Importance from Machine Learning Algorithms, Recursive Feature Elimination (RFE) used instead of Step wise Forward and Backward Selection

Such small and up-and-coming businesses can use machine learning to see what predictions they get with their data that they have at the moment by replicating the steps taken in this project but will have to interpret the results with a grain of salt due to the high chance of overfitting. They can then try to use the analysis/prediction to help them figure out what aspects of their business they need to focus on first as they start to rebuild/recover their business from the economic effects of COVID-19 during this difficult time. More time should also be taken to try different machine learning models to see if they perform better. As well as, Consulting with a more experienced data scientist is also recommended to be able to troubleshoot error messages in a timely manner.

Lost customers mean lost revenue which is why it is so important for companies to know in advance which customers will churn in the near future and what to do/focus on to prevent this from happening.

References

1.10. Decision Trees. (n.d.). Retrieved from scikit learn: https://scikit-learn.org/stable/modules/tree.html

Breiman, L., & Cutler, A. (n.d.). *Random Forests Leo Breiman and Adele Cutler.* Retrieved from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr

Brownlee, J. (2016). *Logistic Regression for Machine Learning.* Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Brownlee, J. (2020, May). *Recursive Feature Elimination (RFE) for Feature Selection in Python*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/rfe-feature-selection-in-python/

Brownlee, J. (2020). *Train-Test Split for Evaluating Machine Learning Algorithms.* Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. *Advances in Knowledge Discovery and Data Mining, 5476*, 475-482. doi:10.1007/978-3-642-01307-2_43

Cao, J., Yu, X., & Zhang, Z. (2015, February 26). Integrating OWA and data mining for analyzing customers churn in E-commerce. *Journal of Systems Science and Complexity*, 381–392. doi:https://doi-org.ezproxy.lib.ryerson.ca/10.1007/s11424-015-3268-0

Donges, N. (2019). *A complete guide to the random forest algorithm.* Built In. Retrieved from https://builtin.com/data-science/random-forest-algorithm

Fawcett, T. (2008, June). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. doi:https://doi.org/10.1016/j.patrec.2005.10.010

IBM. (2017). *Decision Tree Models*. Retrieved from IBM: https://www.ibm.com/docs/en/spss-modeler/18.1.0?topic=trees-decision-tree-models

Jahromi, A. T., Stakhovych, S., & Ewing, M. (2014, October 15). Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, *43*(7), 1258-1268. doi:https://doi.org/10.1016/j.indmarman.2014.06.016

Mishra, A. (2018, February 24). Metrics to Evaluate your Machine Learning Algorithm. *Towards Data Science*. Retrieved from https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Mulcahy, L. (2020, October 28). Avoiding customer churn: How to secure repeat business for your brand as the pandemic continues. *MultiBriefs*. Retrieved May 2021, from http://exclusive.multibriefs.com/content/avoiding-customer-churn-how-to-secure-repeat-business-for-your-brand-as-the/marketing

Nnamoko, N., & Korkontzelos, I. (2020, April). Efficient treatment of outliers and class imbalance for diabetes prediction. *Artificial Intelligence in Medicine, 104*. doi:https://doi.org/10.1016/j.artmed.2020.101815

Rachmawati, I. (2021). Customer's Loyalty of Indonesia Cellular Operators in The Pandemic of COVID-19. *Jurnal Manajemen Teknologi, 19*(3), 220 – 238. doi:https://doi.org/10.12695/jmt.2020.19.3.1

Ranchhod, V., & Daniels, R. C. (2021, March). Labour Market Dynamics in South Africa at the Onset of the COVID-19 Pandemic. *South African Journal of Economics, 89*(1). doi: https://doi-org.ezproxy.lib.ryerson.ca/10.1111/saje.12283

Swalin, A. (2018, June). *Choosing the Right Metric for Evaluating Machine Learning Models—Part 2*. Retrieved from KD nuggets: https://www.kdnuggets.com/2018/06/right-metric-evaluating-machine-learning-models-2.html

Swaminathan, S. (2018). *Logistic Regression — Detailed Overview*. towards data science. Retrieved from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Uddin, S., Khan, A., Hossain, M. E., & Moni , M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making, 19*. doi:https://doi.org/10.1186/s12911-019-1004-8