MEF UNIVERSITY

MACHINE LEARNING APPLICATIONS TO INCREASE CUSTOMER SATISFACTION IN FINANCE SECTOR

Capstone Project

Leyla Yiğit

İSTANBUL, 2019

MEF UNIVERSITY

MACHINE LEARNING APPLICATIONS TO INCREASE CUSTOMER SATISFACTION IN FINANCE SECTOR

Capstone Project

Leyla Yiğit

Advisor: Prof. SEMRA AGRALI

İSTANBUL, 2019

MEF UNIVERSITY

Name of the project: Machine Learning Applications to Increase Customer Satisfaction in Finance Sector

Name/Last Name of the Student: Leyla Yiğit

Date of Thesis Defense: 09/09/2019

I hereby state that the graduation project prepared by Leyla Yiğit has been completed under my supervision. I accept this work as a "Graduation Project".

09/09/2019
Prof. Semra Ağralı

I hereby state that I have examined this graduation project by Leyla Yiğit which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

09/09/2019

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of Leyla Yiğit and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Prof. Semra Ağralı	……..
2. ……..	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Leyla Yiğit 09/09/2019 Signature

EXECUTIVE SUMMARY

MACHINE LEARNING APPLICATIONS TO INCREASE CUSTOMER SATISFACTION IN FINANCE SECTOR

Leyla Yiğit

Advisor: Prof. Semra Ağralı

AUGUST, 2019, 44 Pages

In this project, consumers' complaints about financial data are analyzed. After the analysis, we aim to provide a tool for financial companies such as banks, Lenders that will help them in managing communication with the consumers. Our main aim is to answer the question "How do consumers feel?" This analysis will give a complete picture of consumers' feedback.

We start the project by clustering the customers into different groups. In order to classify customers, we use classification algorithms XGBOOST and Random Forest. XGBOOST is used to predict the probability of getting a complaint. XGBOOST is also tested as an ensemble learning technique. By Using Random Forest the comparison of Bagging and Boosting is performed.

This kind of model is very useful for a customer service department that wants to classify the complaints they receive from their customers. These kinds of models can also be expanded into a system that can recommend automatic solutions to future complaints as they come.

The topic is motivated by the researcher's experience in finance where she intends to increase credit sell numbers by anticipating customer feelings. The data set that we use has many measures and dimensions that facilitate to use more than 3 machine learning algorithms. The complaints database is published by the Consumer Financial Protection Bureau (https://www.consumerfinance.gov/). It provides consumers' feedback in a string format.

We also aim to analyze consumers' complaints dataset from the perceptive of a consumer dispute.

Key Words: Complaints about Financial Products, XGBOOST, Exploratory Data Analysis, Random Forest.

ÖZET

FİNANS SEKTÖRÜNDE MÜŞTERİ MEMNUNİYETİNİ ARTTIRMAK İÇİN MAKİNE ÖĞRENME UYGULAMALARI

Leyla Yiğit

Tez Danışmanı: Prof. Semra Ağralı

AĞUSTOS, 2019, 34 Sayfa

Bu projede, tüketicilerin finansal verilerle ilgili şikayetleri analiz edilmiştir.

Uygulanan makina öğrenmesi algoritmaları sonucunda, şirketler şikayet takibini daha iyi yapacaklardır. Otomatik ve etkili çözümler sunan bir sistem bu algoritmaların çıktıları ile oluşturulabilir. Şikayet edecek müşteriler yada şikayet konuları önceden tahmin edilerek aksiyonlar alınabilir.

Şikayet analizleri için temelde şu makina öğrenmesi algoritmaları kullanılacaktır: XGBOOST, Rassal Orman ve Lojistik Regresyon. Bir şirketin şikayet alma yüzdesini ya da bir müşterinin tekrar şikayette bulunma yüzdesini tahmin etmek için XGBOOST kullanılacaktır. Bu veri seti için, literatür taramalarında SVM, Regresyon, Naive Bayes gibi tekniklerin kullanıldığı görülmektedir. XGBOOST ile ansambl bir algoritma kullanılmış olacaktır. Rassal Orman kullanılarak ise Bagging ve Boosting teknikleri karşılaştırılıyor olacaktır.

Bu tür bir model, müşterilerinden aldıkları şikayetleri sınıflandırmak istedikleri bir müşteri hizmetleri departmanı için çok faydalı olacaktır.

Şikayet veritabanı Tüketici Finansal Koruma Bürosu tarafından sağlanmakta olup, bu veri seti tüketicilerin ipotek hizmetleri, ön ödemeli kart hizmetleri, öğrenci kredisi vb. bu linkten elde edilebilir: (https://www.consumerfinance.gov/)

Projenin amacı, müşterilerin şikayet analizleri ile finansal kurumların şikayetlere hızlı, sistemli ve doğru aksiyon almalarına yardımcı olmaktır.

Anahtar Kelimeler: Finansal Ürünler Şikayet, Veri Analizi, XGBOOST, Rassal Orman.

TABLE OF CONTENTS

TABLES

FIGURES

1. INTRODUCTION

When a company understands how a customer feels about their products/services after an analysis performed using the complaints database, the company can find out the issues related to the products and customer expectations. The companies, also, can take necessary precautions. They may achieve these through a model that predicts their complaint numbers. “Since, how and when” consumer complaints are resolved are among commonly used industry metrics for measuring customer satisfaction. In addition, customers are really affected by complaint comments on the internet while they prefer a new product or system.

1.1. Aim of the Project

The purpose of the project is to create a system that will automatically predict whether consumers dispute company response or not.

1.2. Dataset Source

Consumer Complaints dataset includes data about financial products and services. These complaints are sent to companies. Then, complaints are published at a certain time. In this way, consumer complaints help to improve the financial marketplace. (https://data.world/cfpb/consumer-complaints)

The Consumer Complaint Database is a group of complaints. These complaints can be about a service or product. This data is sent to the companies as feedback on where they went wrong but that can help them to improve as well. This database is updated daily. Each complaint comes from USA states. When the complaint is submitted to the company,
the company is given a warning that the complaint is forwarded. The CFPB publishes this data set in order to make financial services better. Personal data is deleted before this data is published.

Complaint Database is queried by using SQL. Then this queried data is downloaded. Data set source format is CSV. In all analysis, CSV file is uploaded to R and Python environment.

2. LITERATURE REVIEW

One of the ways through which consumers communicate their dissatisfaction is complaints. Therefore, it is crucial for institutions to understand the behavior of consumer complaints. In the world of business, receiving complaints from consumers takes place every day. Even though nobody likes complaints, a consumer's complaint gives an institution a chance to identify the problem and resolve a particular issue about a product or service. Such complaints help in creating a good relationship with consumer because when their concerns are taken seriously and they are provided with solutions, it demonstrates that the company cares and values their customers. It is therefore important for an institution to have a service complaint management system to help result into the best relationship with consumers. According to Bloemer et al (2003) a good relationship with customers is essential because it increases efficiency of the total complaining process. Normally, the complaining process takes two directions; the submission of complaints by the consumers, institution's response and the consumer's decision to accept or dispute the response. In this case, the algorithm techniques are applied to both directions to help predict the probability of the disputes by consumers as a way of ensuring better customer service management.

The consumer complaint database is introduced as a way of ensuring that consumers are protected as well as analyzing consumer's behaviors, financial services institutions and market activities. Currently, there are over 100,000 consumer complaints collected for various financial institutions. This makes it a rich resource for CFPB analysts as well as those financial institutions that are looking for upcoming trends concerning the consumer complaints that are related to financial services' products including the resolutions undertaken to solve the issues presented (McCoy, 2012). There are a number of trending complaints from consumers. There is also customer misunderstanding that brings about more complaints. Moreover, another likelihood of a complaint comes from an affluent and established neighborhood, where those consumers from wealthier areas present high chances of filing complaints. A closer view of these observations helps institutions to be able to get a deeper understanding of their own internal complaint problems and databases since they correlate data from consumer complaint database. The resulting insight is used to facilitate improvements in their regulatory complaint measures,

the customer satisfaction and their own effectiveness when it comes to institutional operations.

According to Ayres (2013), there is a process that the consumer complaint takes. It begins when the consumer files a complaint with the CFPB. Here, the data base is used as a way of sorting what the complaint is and for whom it is intended. The CFPB then reviews the complaints presented by the consumers to check on the completeness, jurisdiction and whether it is non-duplicated. The complaints that meet the criteria are the presented to appropriate institution for resolution. Once it is received by an institution, it is then answered back as soon as possible. When the answer is sent back to the consumer, they are given an option of accepting or disputing the response. In his study, Ayres (2013) reveals that at the end of it all, the consumer is asked to write if they were satisfied with the resolution. One interesting aspect of database is that the consumers have the ability to provide a narrative as a way of explaining their reason for the registered complaint. The consumer complaint database is an important aspect as it provides the institutions a better understanding of what their consumers are going through and how they can effectively solve their issue.

There are three machine learning algorithms that this review focuses on as an experiment for sentiment analysis. They include XGBOOST, Random Forest and Regression analysis. The three algorithms have different philosophies; however every machine learning technique is depicted to be effective in previous studies that have been conducted on them.

Machine learning and approaches that are driven by data are becoming a crucial aspect in many areas. According to Coussement (2008), a factor that leads to such successful applications is the use of scalable learning systems that has the ability to learn the interest model from large datasets. One of the machines learning technique that is used is the XGBoost which is known to be the most shining in many applications. This is because the technique provides state of the art result in ranking issues. It is also used as a predictor that stands alone as it is connected into real world pipeline productions for as click by the means of rate prediction. When it comes to challenges such as Netflix prizing, this algorithm is a defector choice of ensemble technique. The influence of the system has been highly recognized in data mining challenges. Some of the challenges that can be addressed by this system in an institution include; prediction of store sales, prediction of

customer behaviours and risky hazard prediction. In his study, Chen (2018) asserts that the scalability of the XGBoost in all situations makes it possible for its success. This is because the system is able to run ten times more than the current solutions that exist on a particular memory limited place. The scalability is possible due to the various significant systems as well as algorithmic optimization in XGBoost. Furthermore, the XGBoost is able to exploit computation that is out-of-core thereby enabling data scientists to process as many examples as possible on a desktop (Morel, 1997)

Previous studies on market research reveals that regression analysis is the most frequently used tool because it allows market researchers to be able to analyze the correlation between dependent and independent variables (Ghazizadeh, 2014). When it comes to marketing applications, the dependent variables are normally the outcome intended whereas the independent variable is the tool that should achieve the outcome intended. Chagas (2018) found that the use of regression analysis helps in indicating if the independent variables have a significant correlation with dependent variables. It also makes it possible to indicate the relative strength of various independent variables' effects on a dependent variable. Moreover, regression analysis helps in predicting consumer complaints. Regression analysis is one of the most popular techniques that are used by researchers who work on predicting customer satisfaction. The regression has been trained to depict if a customer is going to churn. Here the accuracy depends majorly on the coefficients for the regression (Fornell, 1980)

Another great statistical learning model that is used to analyze consumer complaint database is Random forest. This algorithm technique works well with small medium data. It is composed of various decision tresses, where each has same nodes, with different data that results to different leaves. Datta (2000) indicates in his study that this technique can solve both regression and classification problems, thus a diverse model that is widely used by researchers. Random forest is fast to train when it comes to test data and it prevents overfitting data. When it comes to regression issues, Random forest is created by growing simple trees, each able of generating a response numerical value. Here, the predictor is selected randomly from similar distribution as well as for all trees (Kandasamy, 2018)

On the other hand, in classification issues, the Random forest defines a margin function that shows the extent to which the number of votes for the correct class surpasses the average vote in any class available in the dependent variable. This measure ,therefore,

makes it possible to make predictions as well as associating a confidence measure to the predictions (Vladislava, 2017). Basically, Random Forest can incorporate those data that are missing flexibly in the predictor variables. Usually, these missing data are seen in particular observation, where the prediction made is based on the previous node at the particular tree. This technique has two ways of replacing those data that are missing values (Ha, 2002). To start with, if the variable is not categorical, this method normally computes median values of the class variable. On the other hand, if the variable is categorical, the replacement becomes the non-missing value in variable class.In the previous studies for Financial Complaints dataset, XGBOOST or Random Forest are not used instead of Naive Bayes, SVM is used. Pang and Lee (2002), Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization.

Financial institutional should realize the significance of consumer complaints for their survival. If the consumer complaints are handled carefully and managed appropriately, it will result to loyal customers to an institution. Through the application of the algorithm techniques, it is possible to get a complete picture of the consumers' feedback. These algorithms are used in this case to generate sentiments from the customer's complaints dataset and classify the sentiments accordingly depending on the consumer's complaints. The algorithms were also used to give a deeper understanding of the products and complaints diversification

3. PROBLEM STATEMENT AND METHODOLOGY

The machine learning algorithms that can give a further perspective for the problem studied in this project are XGBOOST, Random Forest and Logistic Regression Analysis by using Python and R. The diagram provided in Figure 1 shows the phases that the project will follow.

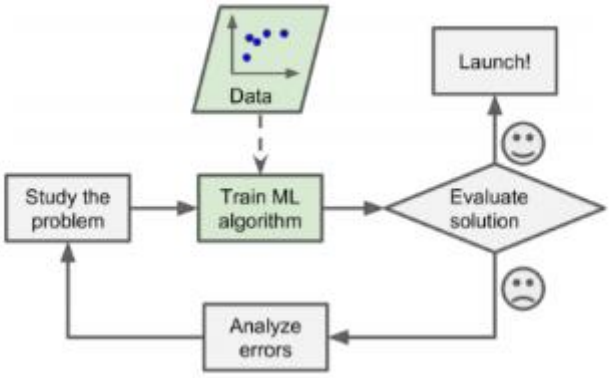

Figure 1: Machine Learning Flow (Aurélien, 2017)

Project phases are given below:

1. Study of Business Problem
2. Data Gathering
3. Exploratory data analysis
4. Data cleaning and Data Prepossessing
5. Random Forest, XGBOOST and comparison of bagging and boosting algorithms
6. Logistic Regression
7. Conclusion

4. EXPLORATORY DATA ANALYSIS AND DATA PREPROCESSING

In this section, data prepossessing and detailed exploratory data analysis are explained.

4.1. Data Shape

The Data frame has 1356310 rows and 18 Observations.

The columns, column data types and the data held by the columns are explained in the below table.

Table 1: Consumer Complaint Database Features and Data Types

COLUMN NAME	TYPE	DESCRIPTION
date_received	date	The date the CFPB received the complain
product	string	The type of product/service the consumer identified in the complaint
sub_product	string	The type of sub-product the consumer identified in the complaint
issue	string	The issue the consumer identified in the complaint
sub_issue	string	The sub-issue the consumer identified in the complaint
consumer_complaint_narrative	string	complaint narrative Actual feedback/complaint
company_public_response	string	The company's optional, public response to a consumer's complaint
company	string	The complaint is about this company
state	string	The consumer's reported mailing state for the complaint
zip_code	string	Mailing ZIP code provided by the consumer
tags	string	Data that supports easier searching by or on behalf of consumers
consumer_consent_provided	string	Finds if the consumer agreed to publish their complaint narrative
submitted_via	string	a How the complaint was submitted to CFPB
date_sent_to_company	date	The date the CFPB sent the complaint to the company
company_response_to_consumer	string	This is how the company responded
timely_response	boolean	Whether the company gave a timely response
consumer_disputed	string	Whether the consumer disputed the company's response
complaint_id	integer	The unique identification number for a complaint

Only 1 of the 18 features is in the numeric data type. There are 2 date columns. All other columns are of string data type.

Strings are encoded in the model phase of the project. But in this trial, when confusion matrix is evaluated, the model cannot learn to 1's (If consumer dispute, it is 1 in both train and test dataset.) because of this reason, string encoding is not preferred. In addition to this, Since the product and issue columns in Data have a lot of value, almost 600 features are created after encode. This leads to a performance problem in R. Therefore, One-hot Encode method is applied only for the mortgage product and after confusion matrix analysis, it is seen that the model could not learn enough.

Whether there is a duplicate row or not is controlled also.

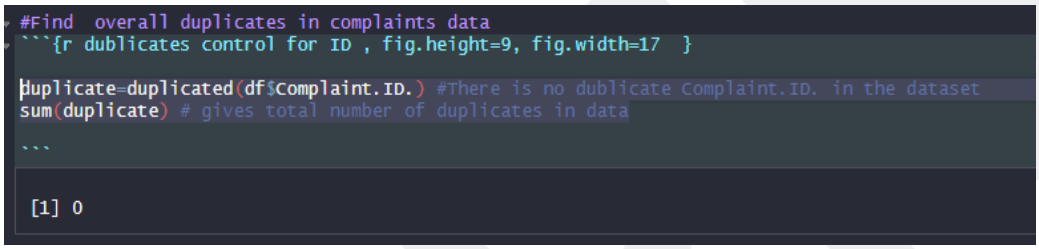

Figure 2: Duplicate row control

The following table shows the unique and null ratios of the columns.

Table 2: Consumer Complaint Database NA's and Unique Ratio

Variable Name	Variable Type	% of Missing	No. of Unique values
Date.received**	Date	0	2812
Product*	factor	0	18
Issue*	factor	0	168
Company*	factor	0	5372
Submitted.via*	factor	0	6
Date.sent.to.company**	Date	0	2761
Company.response.to.consumer*	factor	0	9
Timely.response.*	factor	0	2
Complaint.ID	integer	0	1356310
State*	factor	0.02	64
ZIP.code*	factor	0.09	23003
Sub.product*	factor	0.17	77
Sub.issue*	factor	0.4	220
Consumer.disputed.*	factor	0.43	3
Consumer.consent.provided.*	factor	0.44	5
Company.public.response*	factor	0.64	11
Consumer.complaint.narrative*	factor	0.69	400751
Tags*	factor	0.86	4

4.2. Missing Value Analysis

In the CSV file, NA's are not in a proper format. The code is written as follows to identify all null records.

```
#N/A's is not in a proper way in the data frame
df<-read.csv("complaint_data.csv",na.strings=c("","  ","N/A","NA"))

```

Figure 3: Get all NA's in CSV

The following figure shows NA's distribution in the dataset.

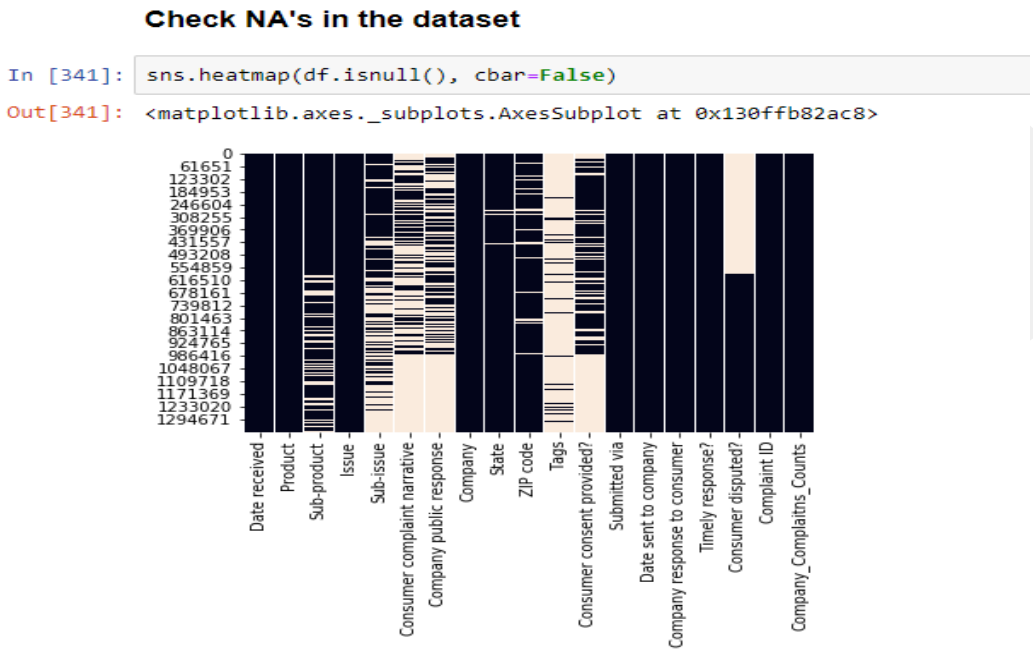

Figure 4: NA's Distribution in the Data Frame

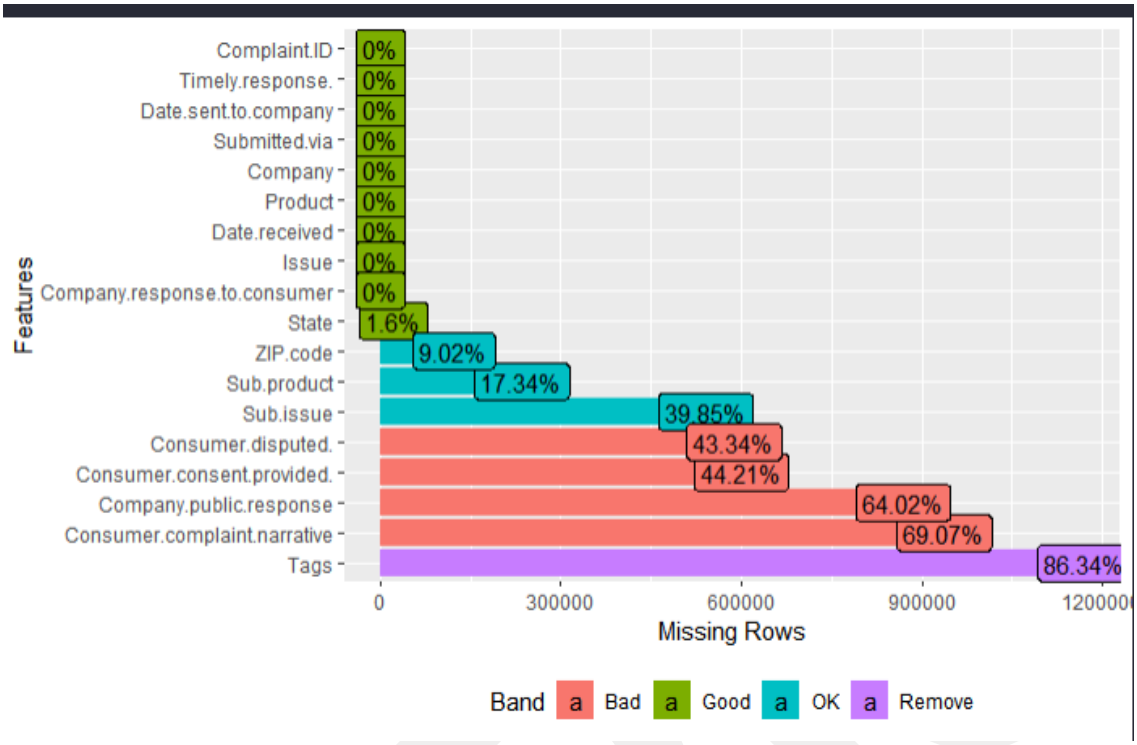

Figure 5: The features that cannot be used because of NA's rate.

Tags: It is deleted because NA's ratio is high to use for a machine learning algorithm.

Zip Code: It is deleted.

Consumer Complaint Narrative: It is deleted for both high NA's ratio and it is also irrelevant with the algorithm.

4.3. Check Zero and High Variance

These columns are omitted from the data frame. A List of these features is given below as uninformative categorical features. Complaint.ID column has also high variance.

Complaint.ID: Has high correlation. Is is deleted from dataset.

ZIP.code: It is deleted because State will be used only for location.

Consumer.complaint.narrative: It is deleted for prediction analysis. It is irrelevant for this kind of classification analysis.

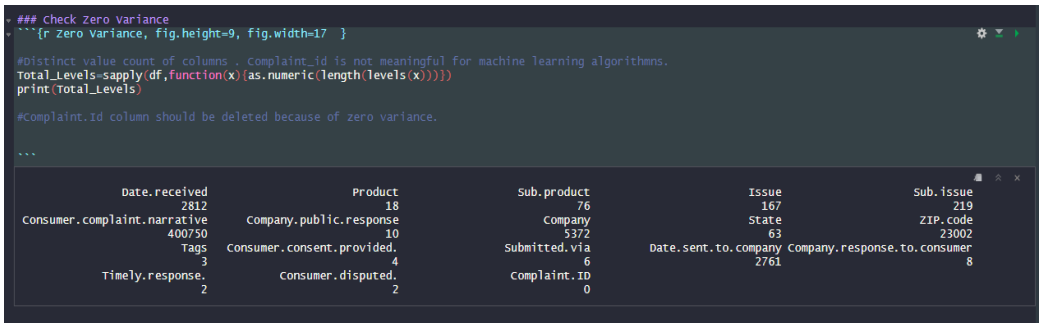

Figure 6: Zero and High Variance Detection

4.3. Feature Engineering

Date Columns:

2 Date columns are converted to the proper date format.

From date columns, 6 new date areas are created. And Original date columns **Date.received** and **Date.sent.to.company** are deleted. These 2 columns cannot be used for tree-based model in date format. Date columns can be used by this feature engineering. Mathematics formulas in R are below for new features. Date difference is found between complaint received and sent date. Year, month and day information was calculated separately for received date and sent date. Thus, date columns became number format.

difftime(df$Date.sent.to.company, df$Date.received , units = c("days"))

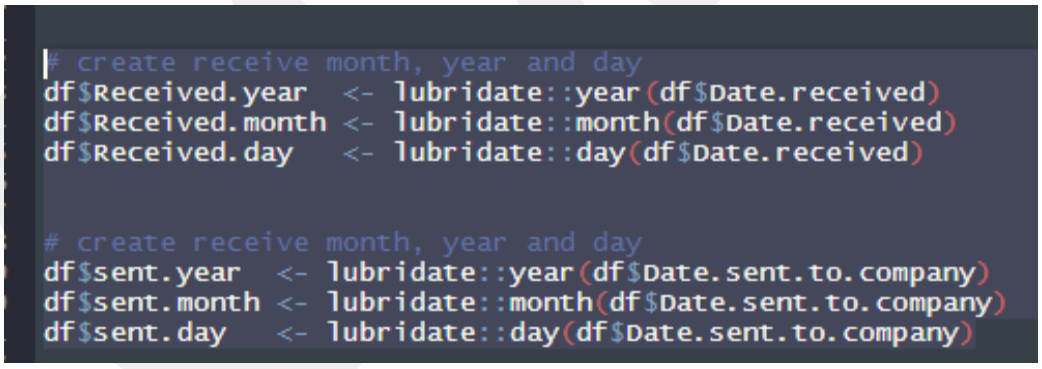

Figure 7: Creation of new features

Columns about sent date are deleted after EDA because they are highly correlated with received date columns.

Company_complaint_count:

A new column is created as **Company_complaint_count.**

The number of complaints received by each company was added as a new column.

4.4. Statistical Analysis

Complaint.count.by.company and Days.to.send.to.company are negatively correlated with 0.04567677. In other words, increasing the duration of the complaint to the company increases the number of complaints. This may be related to the delay in the solution.
We do not reject the null hypothesis that the consumer's disputed habit is independent from the company.

As we will see in the EDA analysis, the top 10 companies have a significant share of complaints

```
#for 2 categorical var correlation , look at chi-square .Pearson's Chi-squared test

chisq.test(df$Company,df$Consumer.disputed.)

#we do  not reject the null hypothesis that the consumer disputed  habit is  independent of company.
#p-value < 0.00000000000000022 p-value is less than   .05 significance level

#product and Consumer.disputed.


# Is there a connection between the company public response and the number of complaints received by that company?
tapply( df$Days.to.send.to.company, df$Company.response.to.consumer, length)
boxplot(df$Complaint.count.by.company ~ df$Company.response.to.consumer )
#yes there is strong correlation. The number of complaints is increasing if the company is in a bad way to respond to the customer.


# Is there a connection between the company public response and the number of complaints received by that company?
tapply( df$Complaint.count.by.company, df$Timely.response., length)   #TRUE 1206149  FALSE: 28475 There is no correlation between timely response and
complaint number.
boxplot(df$Complaint.count.by.company ~ df$Timely.response. )
#yes there is strong correlation. The number of complaints is increasing if the company is in a bad way to respond to the customer.
```

Figure 8: Statistical Analysis in R

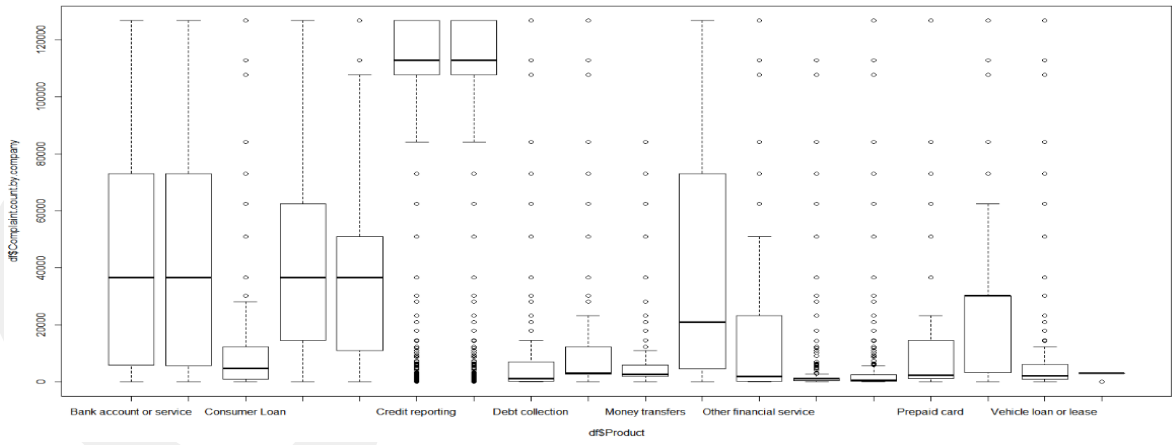

Figure 9: Statistical Analysis, Distribution of Company Complaint Count by product

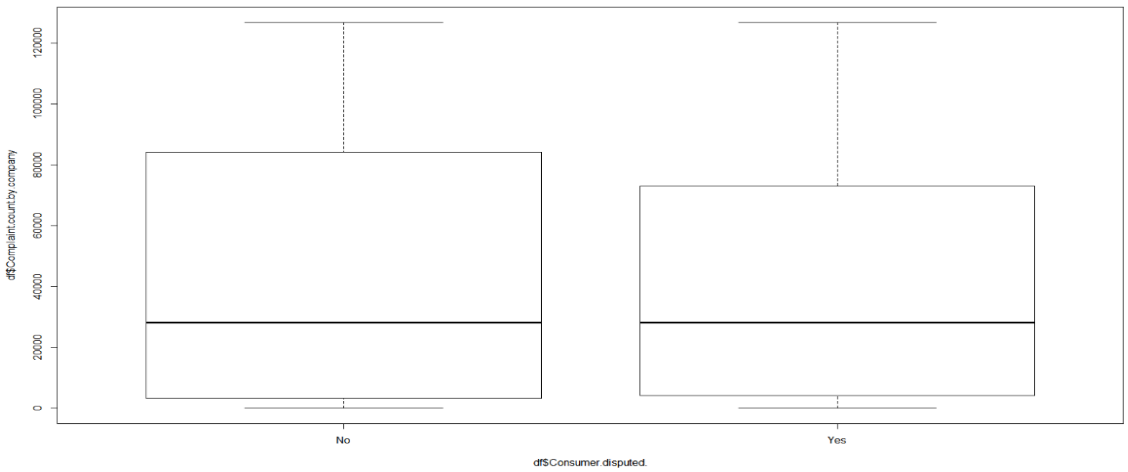

Figure 10: **Statistical Analysis, Distribution of Consumer Disputed and Company Count in R**

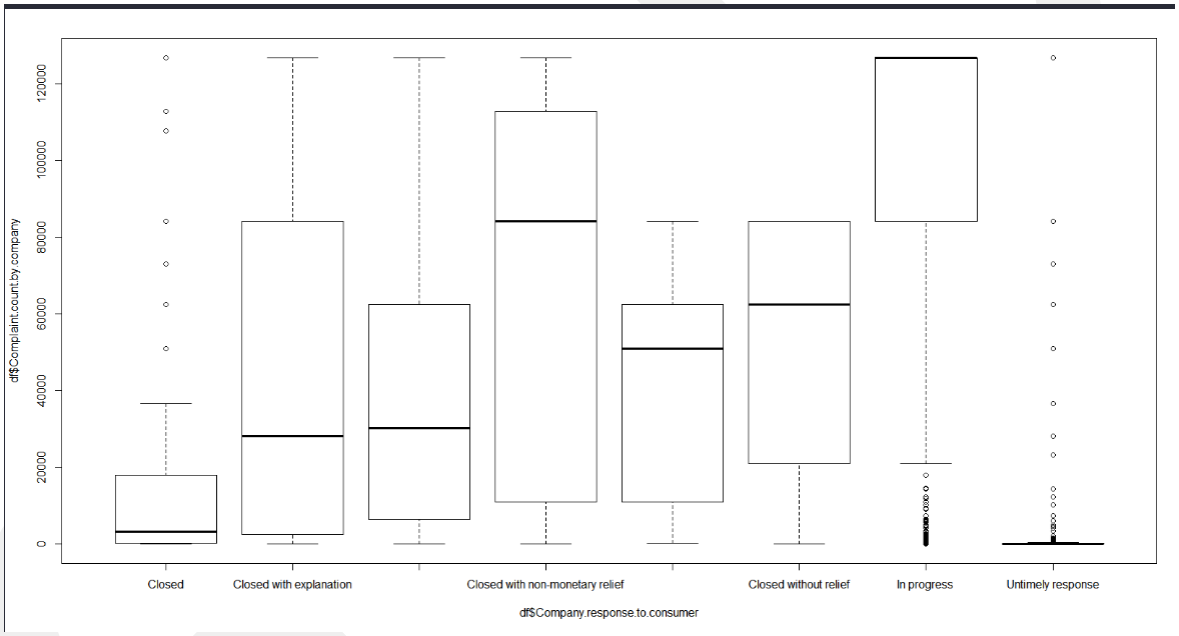

Figure 11: Statistical Analysis, Company Complaint Count and Company Public Response in R

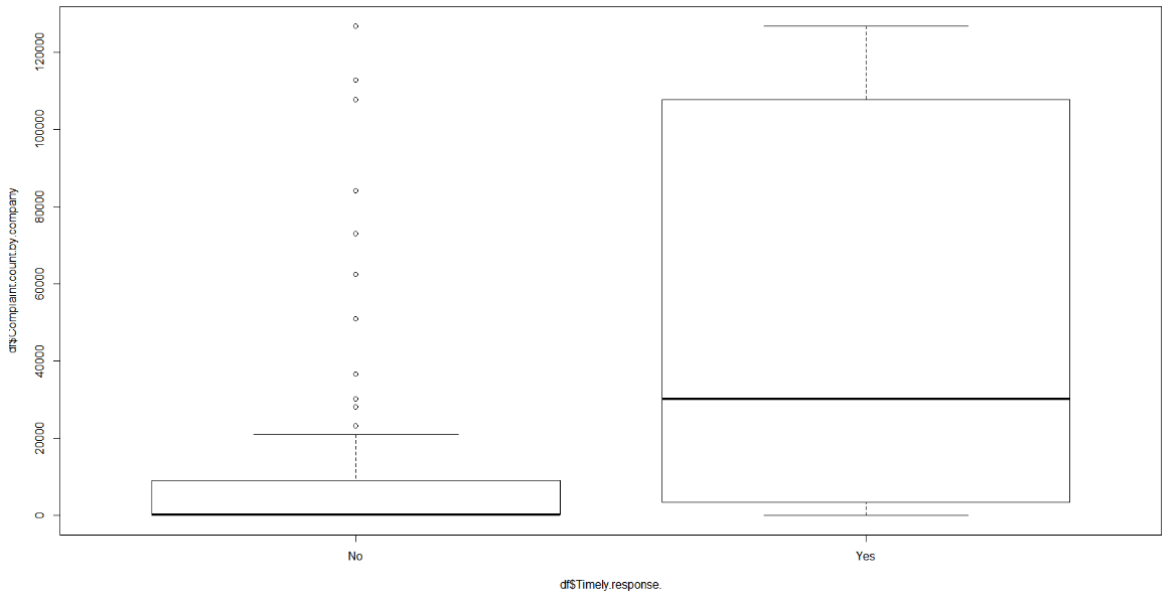

Figure 12: Statistical Analysis, Distribution of Consumer Disputed and Company Count in R

Is there a time seasonality in the consumer complaints dataset. Does this affect model performance

For time analysis, adf.test function is used.

```
boxplot(cc ~ cycle(cc))

plot(stl(cc, s.window = 'periodic', t.window = 15))


seasonplot(cc, year.labels = T, year.labels.left = T, col = 1:4, labelgap = 0.4,
           main = 'Comparing Seasons' )

adf.test(cc)  # -3.3502, Lag order = 3, p-value = 0.07491 alternative hypothesis: stationary .The test outputs a p-value greater than 0.05 therefore the
data are not stationary.
```

Figure 13: adf.test function and season plot in order to understand the time distribution of data

The Result of ADF test (Augmented Dickey-Fuller Test) is in the below figure.

```
        Augmented Dickey-Fuller Test

data:  cc
Dickey-Fuller = -3.3502, Lag order = 3, p-value = 0.07491
alternative hypothesis: stationary
```

Figure 14: adf.test result in R

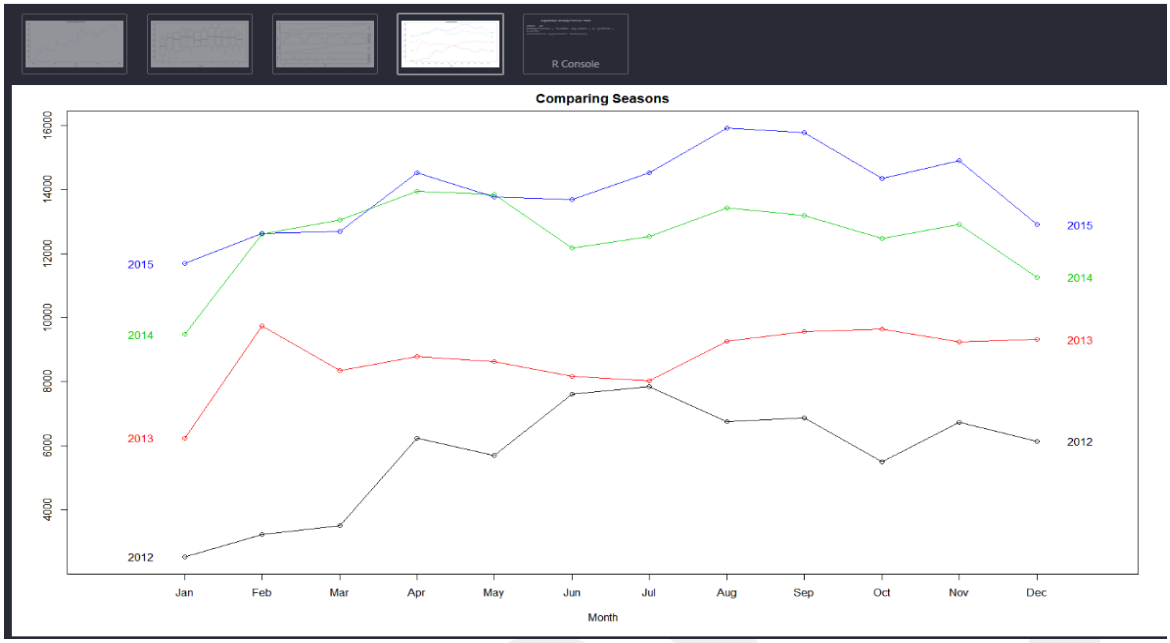

Figure 15: Comparing seasons of complaints number in R

Until the first half of 2013, the number of complaints had a seasonal effect and a fluctuating graph.

But after this point, especially after 2014, there is a rapid rise. The reason for this is the systemic problem caused by mortgagees in 2014. Most consumers complained about Mortgages that year

This data does not seem to have autocovariance but we will run an additional test to verify our intuition.

Box-plots appear a clear increment of complaints rising as months progress to mid-year at that point starts to drop towards the end of the year.

The graphs show that appears regularity a bit more clear but still no clear drift.

As a result, the newly created date columns usage in predictive analytics is considered in a skeptical way. Although there was a temporal effect on the distribution of complaints, these columns are used because the consumer dispute rate is high during periods of high complaints.

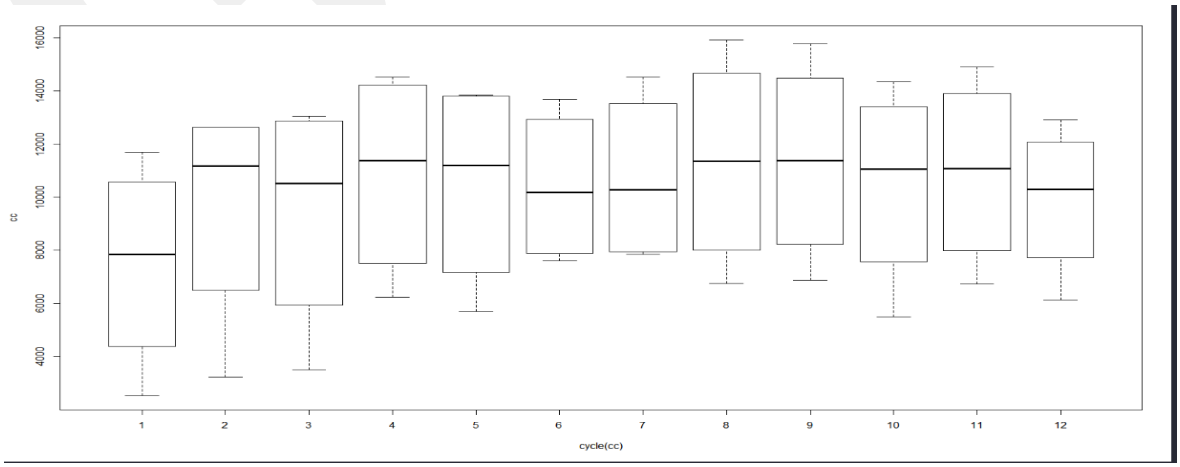

Figure 16: Seasons effect on complaints

Distribution of the number of days between Date Received and Date Sent to Company

Distribution is largely right-skewed, with many complaints taking few or no days to get to the company and some complaints taking an extraordinary amount of time to get to the company.

This time can affect the dispute of customer.

Because of this right-skewed distribution, sent and received dates correlation is high.

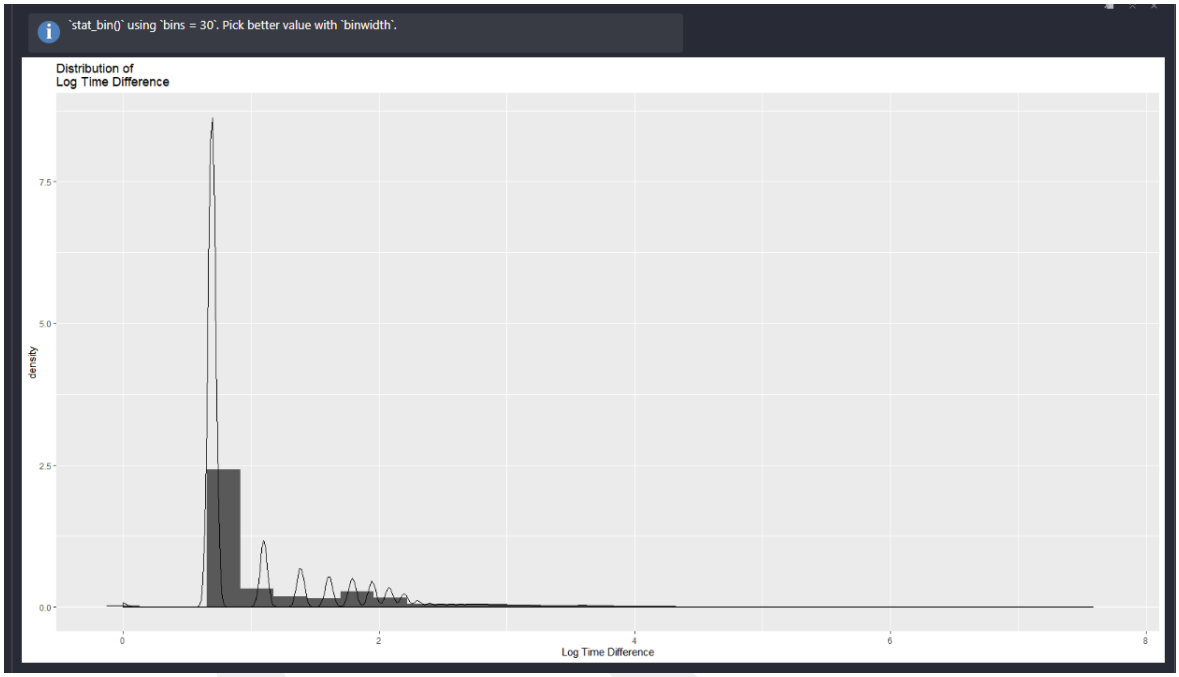

Figure 17: Log time difference graph days between Date Received and Date Sent in R

4.5. Understand Data Deeply

Most frequently complained product

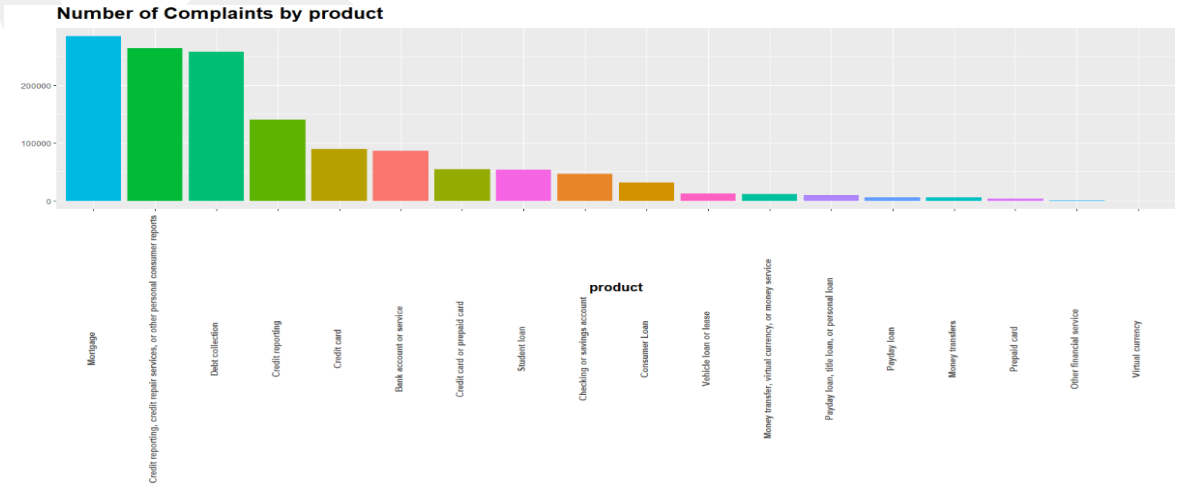

Figure 18: Most complaint product

Figure 19: Most complaint product

Mortgages, customer reports, and debt collection are respectively most complaint products.

Consumer Dispute Distribution

NA's are not included in the modelling. These rows are deleted. Model train and test are done with "No" and "Yes" rows.

```
  No       Yes      <NA>
0.4537569 0.1086794 0.4375637
```

Figure 20: Consumer Dispute Percentage

Consumer Dispute Versus Product Correlation

Mortgage customers dispute at a higher rate.

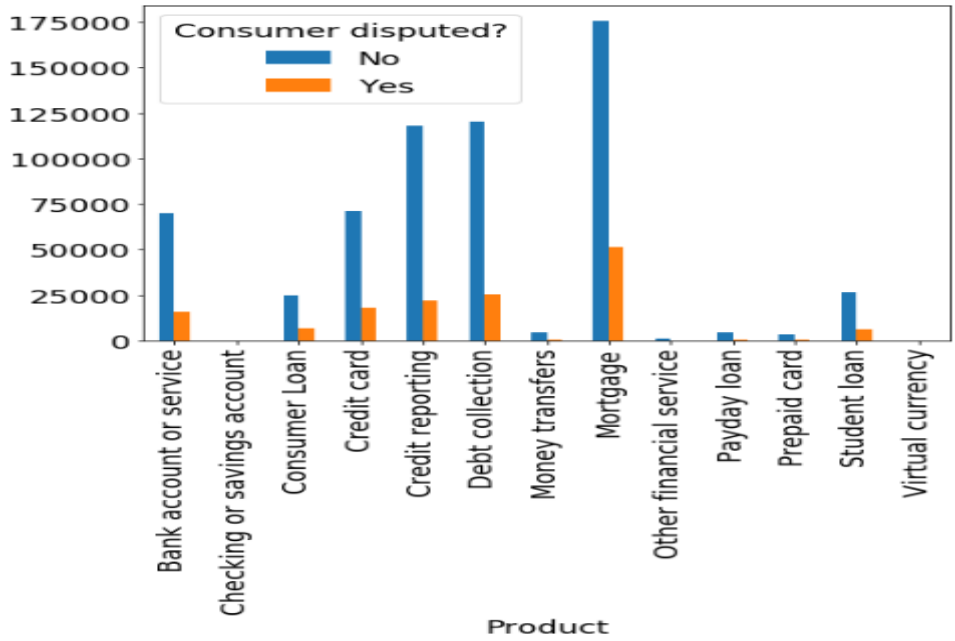

Figure 21: Consumer Dispute vs Product.

Consumer Dispute Versus Company Response to Customer

There is a correlation between closed with relief and monetary relief. Customers get closed with explanation more dispute to companies.

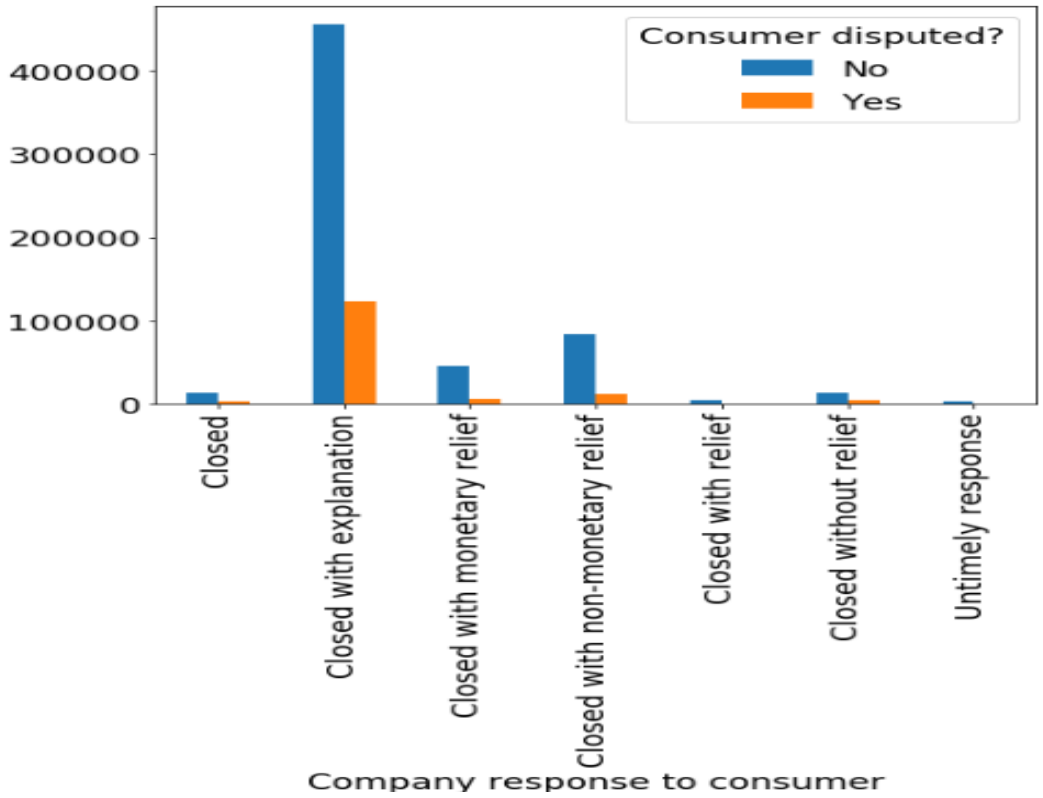

Figure 22: Consumer Dispute vs Company Public Response

4.6. Correlation Analysis and Remove High or Low Correlated Columns

Because of the low correlation rate, all features are kept for the modeling phase.

Table 3: Correlation with Dependent Variable

Independent Variables	Dependent Variable	Cor Value
Consumer.disputed.	Consumer.disputed.	1.00000
Consumer.consent.provided.	Consumer.disputed.	0.03320
Sub.product	Consumer.disputed.	0.01436
Received_month	Consumer.disputed.	0.01032
company_complaint_counts	Consumer.disputed.	0.00853
Received_day	Consumer.disputed.	-0.00046
State	Consumer.disputed.	-0.00087
Issue	Consumer.disputed.	-0.00531
Product	Consumer.disputed.	-0.00752
Company.public.response	Consumer.disputed.	-0.00977
Company.response.to.consumer	Consumer.disputed.	-0.02856
Sub.issue	Consumer.disputed.	-0.02937
Timely.response.	Consumer.disputed.	-0.03169
Received_year	Consumer.disputed.	-0.04780

4.7. Handling NA's

When string columns are encoded, too many columns appear. this causes a performance problem in R. And also, model Since the model has too many columns and rows, the model is too complicated to learn. In confusion matrix of this model, learning rate for 1's is near to zero. As a result, without encoding, NA's are handled with the below methodology. Encoding is not used, instead of numeric and logical categorical variables are created.

Subproduct NA values replaced with 'Not provided'

Issue NA values replaced with 'OTHER'

Submitted via NA values replaced with 'Other'

Consumer consent provided? d with 'Consent not provided'

New data frame after all data prepossessing is below:

Check new data frame for model

```
In [39]: print(df_model.dtypes)
         df_model.head().T
```

```
Consumer consent provided?          bool
company_complaint_counts           int64
Product                         category
Sub-product                     category
Issue                           category
Sub-issue                       category
Company public response         category
Tags                            category
Company response to consumer    category
State                           category
Timely response?                category
Received_year                      int64
Received_month                     int64
Received_day                       int64
Consumer disputed?              category
dtype: object
```

Figure 23: Model data frame

	587820	587821	587822	587823	587824
Consumer consent provided?	True	True	True	False	False
company_complaint_counts	20314	12347	7843	40006	2098
Product	0	1	2	2	1
Sub-product	0	1	0	0	2
Issue	0	1	2	2	3
Sub-issue	0	0	1	1	0
Company public response	0	1	0	1	0
Tags	0	0	0	0	1
Company response to consumer	0	0	0	1	0
State	0	1	0	1	2
Timely response?	False	False	False	False	False
Received_year	2017	2017	2017	2017	2017
Received_month	4	4	4	4	4
Received_day	22	22	22	22	22
Consumer disputed?	0	0	0	0	0

Figure 24: New data frame that is used in the modelling

4.8. Creating Labels

Train and set datasets are created. Split ration is 60 percent.
Dependent variable is Consumer Disputed.

Dependent train and test variable labels are created. Methodology is given below:

TRAIN TEST SPLIT
y_train <- train[,"Consumer.disputed."]
x_train <- train
x_train$"Consumer.disputed."<- NULL

```
x_test <- test
```

Attaching new variable to the data frame

```
x_test=data.frame(x_test,"Consumer.disputed.")
y_test <- x_test[ , "Consumer.disputed."]
x_test <- x_test
x_test$Consumer.disputed.<- NULL
```

Data frame shape that is ready for modelling:

```
dim(x_train)  #461095       13
dim(y_train)  #461095     1
dim(x_test)   # 307395    13
dim(y_test)  #307395      1
```

5. MODELLING

We obtained train and test data sets separately; hence, it is not required to make any further split on data. Two types of boosting algorithms are adopted in this study, which are Gradient Boosting Machine (GBM) and Extreme Gradient Boosting (XGB). Since boosting algorithms are tree-based ensemble methods, train and test data are not normalized. With the adoption of two popular boosting algorithms, we wanted to observe the performance of boosting algorithms.

Random Forest and Logistic regression are other models used in order to predict data.

5.1. XGBoost

We now have a binary classification problem, and two boosting algorithms are created for classification purposes. “caret” package is used again to obtain the best parameters for GBM and XGB classifiers.

XGBOOST ensemble algorithm is performed in 2 different ways. In each method R “CARET” package is used. “caret” package used again for cross-validation in order to maximize accuracy score.

The First data is prepared for XGBOOST.
The XGBoost model only works in matrices data format and all data types must be numeric.

```
#make dependent variables to dataframe
y_train<-as.data.frame(y_train)
y_test<-as.data.frame(y_test)
dim(x_train)  #461095         13
dim(y_train) #461095       1
dim(x_test) # 307395     13
dim(y_test) #307395       1


#whether there is a factor or not
factor.names <- names(x_train[, sapply(x_train, is.factor)])
factor <- x_train[, factor.names]
colnames(factor) #  9 factor name



#make all dataframe matrix
x_train_m <- as.matrix(x_train)
x_test_m <- as.matrix(x_test)
y_train_m<- as.matrix(y_train)
y_test_m <- as.matrix(y_test)



#convert regular data to xgboost data
dtrain.c <- xgb.DMatrix(data = x_train_m,label = y_train_m)

dtest.c <- xgb.DMatrix(data =x_test_m,label=y_test_m)
```

Figure 25: XGBoost Algorithm, Prepare XGBoost data.

In order to make grid search and CV, PC core separated 4 cores.socket cluster with 4 nodes on host 'localhost' so that R performance problem is handled.

c <- makeCluster(4)

XGBoost with Grid Search and 10K Fold CV with Expand Grid Method

XGBoost Gradient Boosting is done with the below values. XGBoost can be easily over-fit compare to Bagging algorithms because of that reason, Cross Validation should certainly be done.

And also in order to maximize accuracy score, best parameters are found with grid search method. Grid search method can be done manually. This method is given below:

```
#grid search and 10 K fold used together to find best parameters.
xgbGrid <-  expand.grid(nrounds = c(10,15,25),
                        max_depth = c(5,8),
                        eta = c(0.3),
                        gamma = 0, colsample_bytree=1,
                        min_child_weight=(1), subsample=1)

fitControl <- trainControl(## 2-fold CV
  method = "repeatedcv",
  number = 10,
  ## repeated ten times
  repeats = 1)

#FIT XBG BOOST CLASSIFICATION
gbmFit <- caret::train(x_train_m, as.factor(y_train_m), method = "xgbTree",
                trControl = fitControl, verbose = T,
                tuneGrid = xgbGrid)
```

Figure 26: Caret Package XGBoost CV and Grid Search

According to the below results,

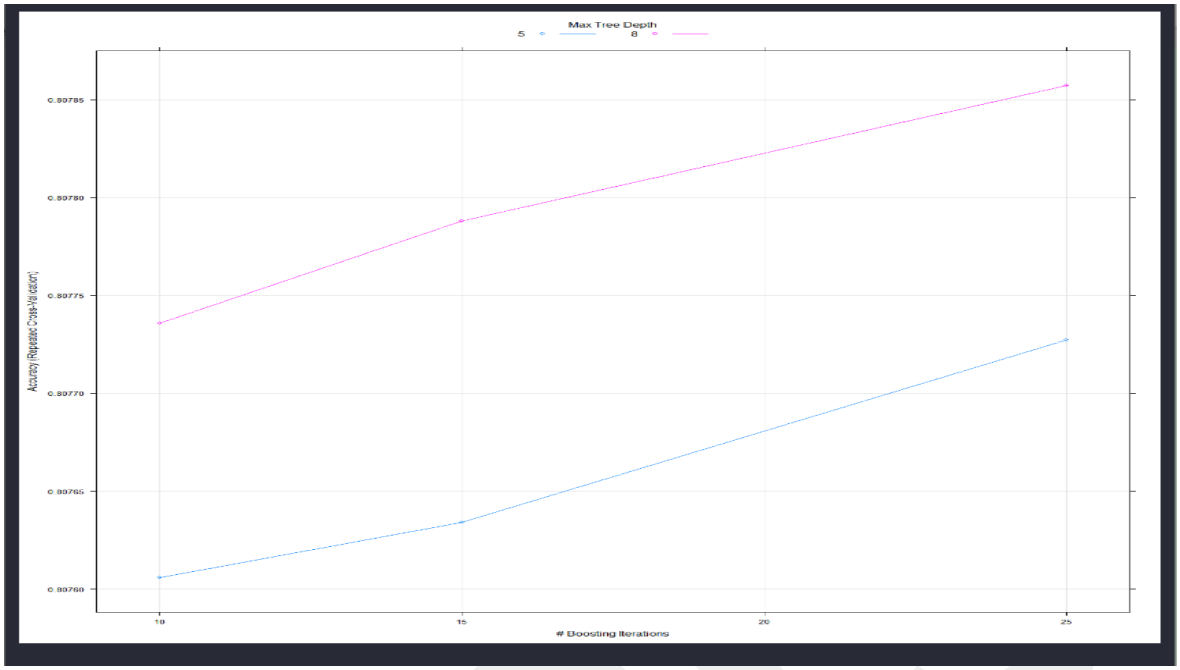

Figure 27: Accuracy score is higher 8 max tree depth with higher iteration

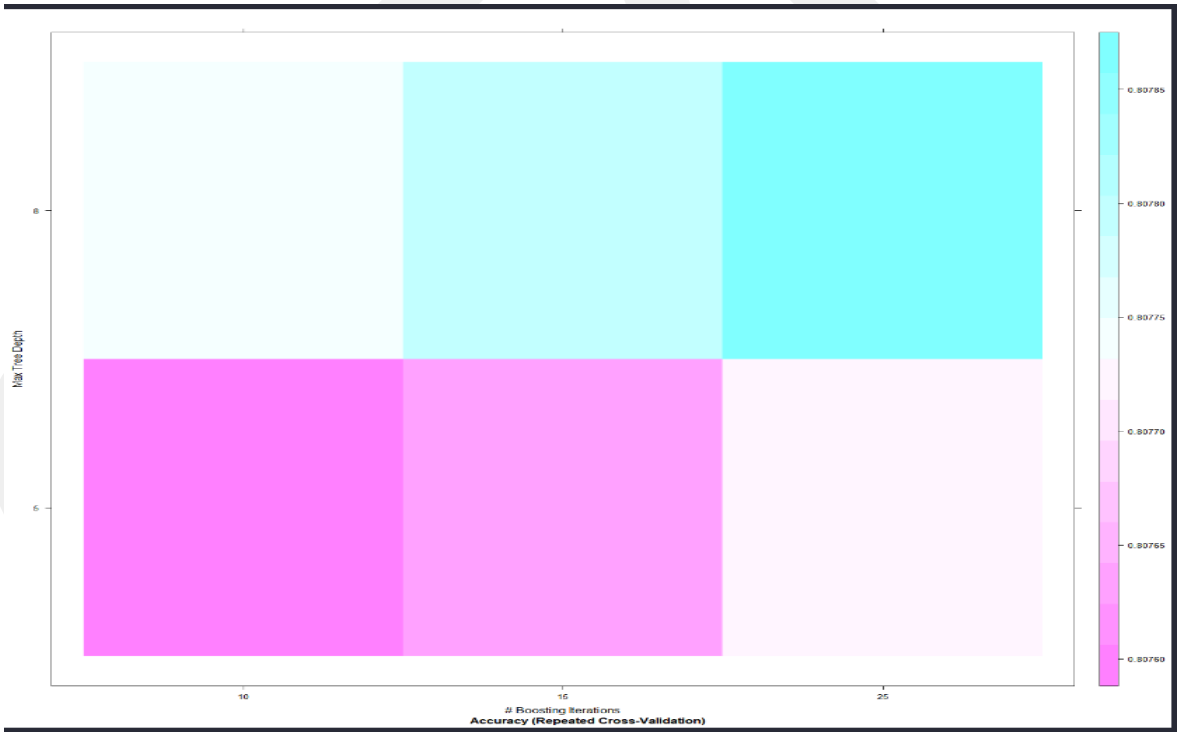

Figure 28: Accuracy Change with Repeated Cross Validation with Tree Depth and Boosting Iteration

XGBoost model is fitted with the best result of manual grid search with 10K fold. Parameters are:

Table 4: Grid Search Best Parameters Result with Expand Grid

XGBoost Expand Grid Search	
Parameter	**Valuer**
nrounds	25
max_depth	8
eta	0.3
min_child_weight	1
colsample_bytree	1
subsample	1

Train Accuracy score: 80.7885 %

After fitting XGBoost model with best parameters which are listed below, results are below:

Table 5: Confusion Matrix of XGB that boosted with grid search

Confusion Matrix for XGB with **Expand Grid Search**		**Actual Label**	
		"Low Probability to Dispute" (0)	"High Probability to Dispute" (1)
Predicted Label	"Actual No Dispute" (0)	247127	59170
	"Actual Dispute" (1)	601	497

0.8055564: Test accuracy rate

XGBoost with Grid Search and 10K Fold CV with Tune Length Method

Tune Length is a function in Caret. It finds best parameters based on the data. Best parameters of tune length grid search are:

Table 6: Grid Search Best Parameters Result with Tune Length

XGBoost with Tune Length	
Parameter	**Valuer**
nrounds	150
max_depth	4
eta	0.3
min_child_weight	1
colsample_bytree	0.8
subsample	1

Train Accuracy score: 80.8085 %

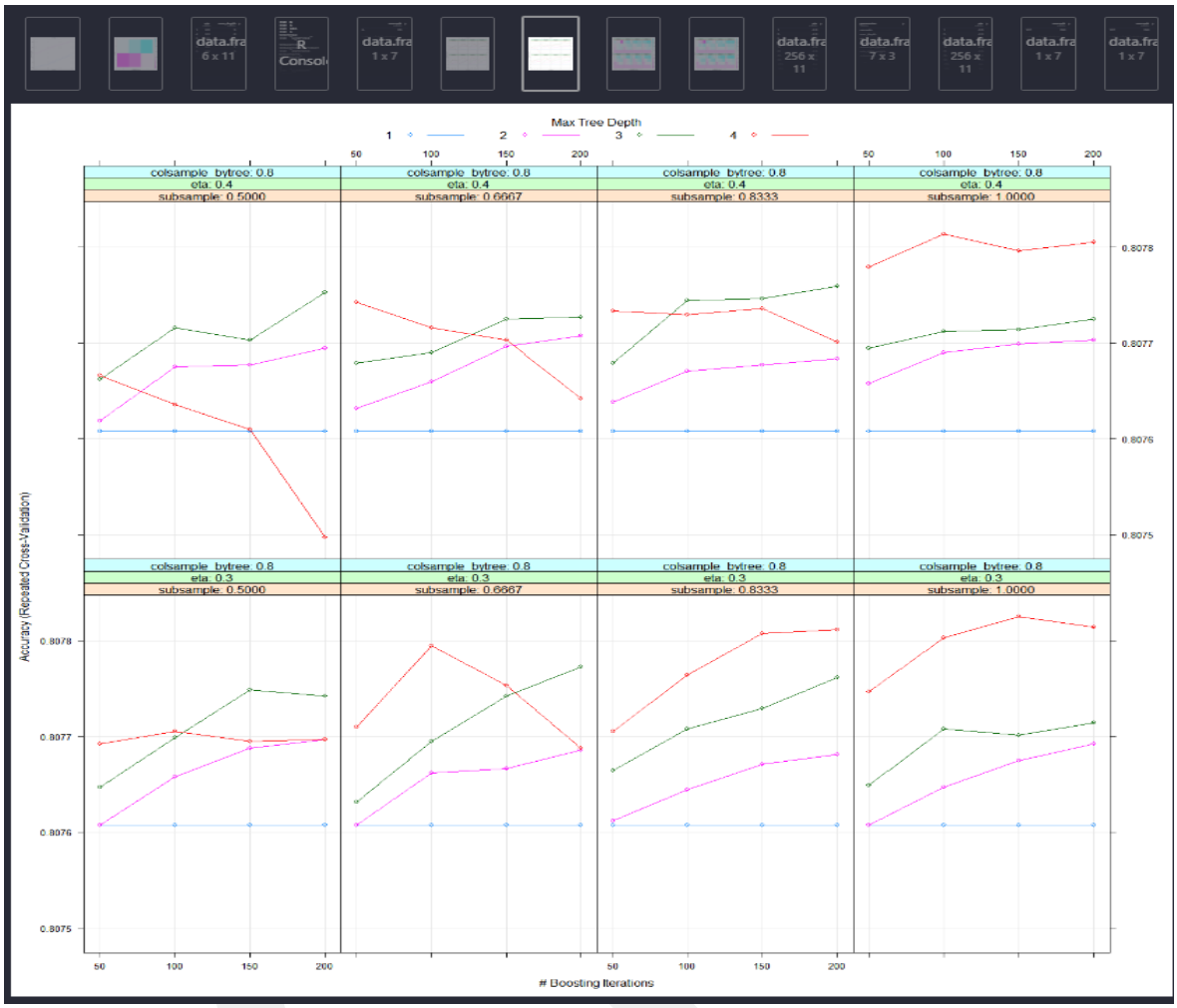

Figure 29: Accuracy Change with Repeated Cross Validation with Tree Depth and Boosting Iteration with Tune Length

0.8050514: Test accuracy rate.

Table 7: Confusion Matrix of XGB that is boosted with Tune Length

Confusion Matrix for XGB with **Expand Grid Search**		**Actual Label**	
		"Low Probability to Dispute" (0)	"High Probability to Dispute" (1)
Predicted Label	"Actual No Dispute" (0)	246803	58974
	"Actual Dispute" (1)	925	693

Feature Importance:

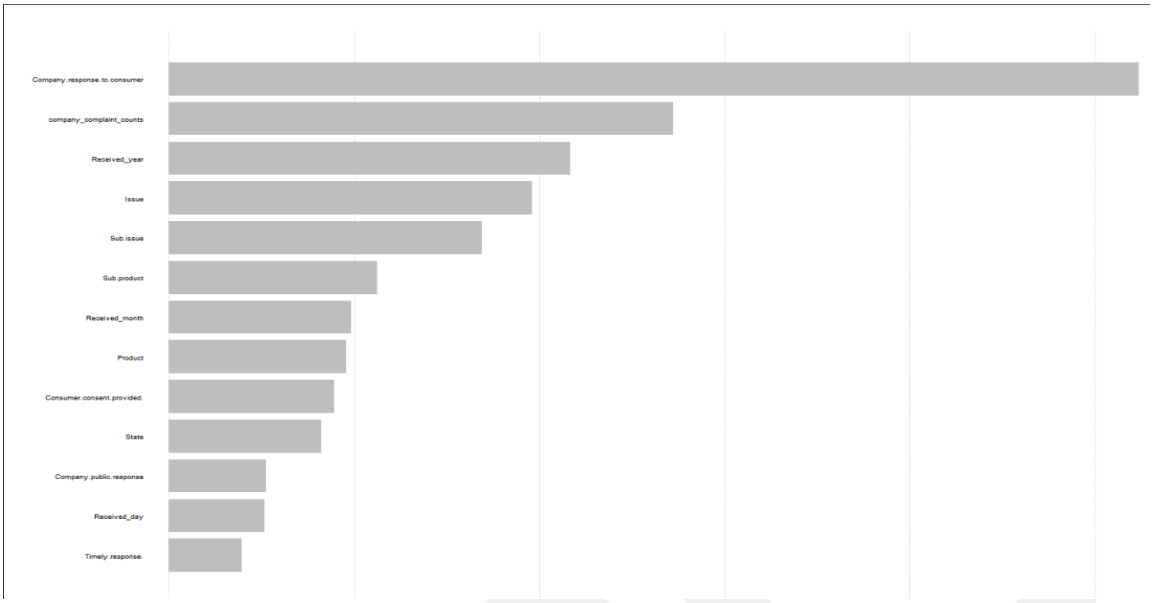

Figure 30: Feature Importance are is the same both Tune Length and Expand Grid in XGBOOST

As a result, for complaint dataset, there is no big difference in train error and test error when tune length and basic grid search methods are used. However, it may be difficult to estimate parameter values in the expand grid search technique. Therefore, Tune length method is preferred although it takes longer.

5.2. Random Forest

Random Forest is a bagging method. This algorithm is applied with CV and grid search.

Parameters are: n_estimators=300, random_state=0

CV is applied as 5

Train accuracy Rate:0.79833115

After prediction test results and confusion matrix are below:

Table 8 Confusion Matrix of Random Forest that is boosted with Tune Length

Confusion Matrix for Random Forest Tune Length		**Actual Label**	
		"Low Probability to Dispute" (0)	"High Probability to Dispute" (1)
Predicted Label	"Actual No Dispute" (0)	118852	5260

	"Actual Dispute" (1)	27031	2555

Classification Report is below:

Table 9: Classification Report

	precision	recall	f1-score	support
False	0.81	0.96	0.88	124112
True	0.33	0.09	0.14	29586
micro avg	0.79	0.79	0.79	153698
macro avg	0.57	0.52	0.51	153698
weighted avg	0.72	0.79	0.74	153698

Based on these results Random Forest has a higher precision rate but a lower accuracy score.

Aim of this projects is to predict whether a customer disputed or not. Because of higher precision rate Random Forest is preferred.

5.3. Logistic Regression

Basic logistic regression is fitted to the Complaints Dataset.

Accuracy score is very low compared to other advanced machine learning algorithms.

Logistic Regression

```
lr = LogisticRegression(class_weight='balanced')
lr.fit(X_train, y_train)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:433: FutureWarning: Default solver will be changed
to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
```

```
LogisticRegression(C=1.0, class_weight='balanced', dual=False,
          fit_intercept=True, intercept_scaling=1, max_iter=100,
          multi_class='warn', n_jobs=None, penalty='l2', random_state=None,
          solver='warn', tol=0.0001, verbose=0, warm_start=False)
```

```
#accuracy score
lr.score(X_test,y_test)
```

```
0.5248474280732345
```

Figure 31: Low accuracy score in Logistic Regression

6. CONCLUSION

In this study XGBoost, Random Forest and Logistic Regression algorithms are used to predict consumer dispute rate. Software development is done on both R and Python. The first development is done on R. Then because of some performance problems Python is used as an additional tool.

For both Random Forest and XGBOOST, the CV method is used to increase the accuracy score and avoid over-fit problems. With usage of cross-validation, over-fitting is not observed in classification. Bagging algorithms have a lower probability of over-fit compared to Boosting algorithms. Therefore, two different tune methods are used together with CV method in XGBOOST modeling. Grid Search Method with Tune length has higher train and test accuracy scores. XGBoost has a higher accuracy score compared to Random Forest (Random Forest: 0.79833115 train accuracy, XGBoost Tune Length: 0.8085). On the other hand, in tree-based models, the accuracy rate alone does not measure the success of the model. Precision and Recall Rate are also significant. Random Forest has a better precision rate. Thus, it is used in the modelling of the Consumer Complaint Dataset.

5. REFERENCES

Ayres, I., Lingwall, J., & Steinway, S. (2013). Skeletons in the Database: An Early Analysis of the CFPB's Consumer Complaints. *Fordham J. Corp. & Fin. L.*, *19*, 343.

Au, W. H., Chan, K. C., & Yao, X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE transactions on evolutionary computation*, *7*(6), 532-545.

Bloemer, J. M., Brijs, T., Vanhoof, K., & Swinnen, G. (2003). Comparing complete and partial classification for identifying customers at risk. *International journal of research in marketing*, *20*(2), 117-131.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, *34*(1), 313-327.

Chen, Y., Wang, J., & Cai, Z. (2018, July). Study on the Application of Machine Learning in Government Service: Take Consumer Protection Service as an Example. In *2018 15th International Conference on Service Systems and Service Management (ICSSSM)* (pp. 1-5). IEEE.

Chagas, B. N. R., Viana, J. A. N., Reinhold, O., Lobato, F., Jacob, A. F., & Alt, R. (2018, December). Current Applications of Machine Learning Techniques in CRM: A Literature Review and Practical Implications. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 452-458). IEEE.

Datta, P., Masand, B., Mani, D. R., & Li, B. (2000). Automated cellular modeling and prediction on a large scale. *Artificial Intelligence Review*, *14*(6), 485-502.

Fornell, C., & Didow, N. M. (1980). Economic constraints on consumer complaining behavior. *ACR North American Advances*.

Ghazizadeh, M., McDonald, A. D., & Lee, J. D. (2014). Text mining to decipher free-response consumer complaints: Insights from the NHTSA vehicle owner's complaint database. *Human factors*, *56*(6), 1189-1203.

Ha, S. H., Bae, S. M., & Park, S. C. (2002). Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case. *Computers & Industrial Engineering*, *43*(4), 801-820.

Kandasamy, P., Raji, D., & Arun, S. (2018, December). Data Science Techniques to Improve Accuracy of Provider Network Directory. In *2018 IEEE 25th International Conference on High Performance Computing Workshops (HiPCW)* (pp. 119-128). IEEE.

McCoy, P. A. (2012). Public Engagement in Rulemaking: The Consumer Financial Protection Bureau's New Approach. *Brook. J. Corp. Fin. & Com. L.*, *7*, 1.

Morel, K. P., Poiesz, T. B., & Wilke, H. A. (1997). ≫ Motivation, Capacity and Opportunity to Complain: Towards a Comprehensive Model of Consumer Complaint Behavior. *ACR North American Advances*.

Vladislava, R. (2017). Machine learning methods application for real estate price prediction in the Russian market.