National College of Ireland

A classification-based approach for modelling disputed responses based on consumer complaint on financial products.

MSc Research Project
FinTech

Clara Onunkwo
Student ID: X18122558

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Clara Onunkwo
Student ID:	X18122558
Programme:	FinTech
Year:	2019
Module:	MSc Research Project
Supervisor:	Noel Cosgrave
Submission Due Date:	16/09/2019
Project Title:	A classification-based approach for modelling disputed responses based on consumer complaint on financial products.
Word Count:	9006
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	16th September 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A classification-based approach for modelling disputed responses based on consumer complaint on financial products.

Clara Onunkwo
X18122558

Abstract

This study is carried out with the purpose of creating and selecting the model with the highest performance rate among other models that can be used in predicting the likelihood of consumers disputing complaints responses made by financial service providers regarding products and services. This study purpose was as a result of the problem faced by financial service providers where responses provided based on complaints raised are disputed by the consumers. The financial service providers view this act as a problem to their services and which calls for solutions. This act could be said to have implications on the quality of financial service and thus affecting consumers satisfaction. For this reason, the study applies three classification model such as Naive Bayes, Random Forest and Logistic Regression to help achieve its purpose and this will be determined based on the accuracy rate of each models. Based on the various analysis carried out on each model, Random Forest presented itself with the highest accuracy rate compared to Naive Bayes and Logistic Regression which presented a good accuracy rate but not as high as that of the Random forest. Based on this, Random Forest was selected as the best fit model to be used.

Keywords: consumer complaint, disputed response, naive bayes, random forest, logistic regression, financial service providers, prediction, financial products and services.

1 Introduction

1.1 Background of the study

Recent studies have shown that the heart of every business organization should be centred on consumers satisfaction and the ability to reach their goals (James et al.; 2019). Such satisfaction and goals are said to account for product quality and thus, determines the success or failure of a business. Various business organizations, particularly the financial organizations are charged with the duties of carrying out financial services. These financial services could be regarded as those economic duties rendered by the finance industries to consumers for its products (Lui et al.; 2019). The economic duties as recorded today are said to cover a wide range of monetary activities (financial products) such as mortgages, loans, credit reporting, debt collection, insurance, stock brokerages, investment, bank account, credit card and many others (Alex et al.; 2016).

Chugani et al. (2018), identifies that various complaints are laid by consumers based on the various financial products carried out by the financial organizations. These complaints are said to have led to the new face of financial service and its product as presented by the Financial Technology (Fintech) circle for customers satisfaction (Kauffman et al.; 2018). Historically, a major financial crisis as recorded in the year 2007/2008 was said to have started as a result of various reasons and one of these reasons was attributed to the financial products particularly the mortgage (Fadzlan and Shah; 2014). Studies have shown that before the great recession, various complaints were laid by various individuals(consumers) on accessing the quality of financial products. February 2003, Buffett raised a complaint on the problems of the mortgage-backed securities (Gartenberg and Pierce; 2015). Ibid in June 2004, the federal reserve chairman Greenspan also raised a complaint on loan interest rate. August 2005, Dr Raghuram Rajan foresaw the financial crisis based on the then financial situation regarding the financial product(mortgage)and service and tried to address the need for an immediate response (Gilreath; 2018). It could be said that all complaints were ignored thereby leading to the financial crisis that resulted in the great recession.

The study of Kauffman et al(2018), also indicated that fintech came into light during the great recession and as such are taking advantage in providing better financial products and services based on consumer complaints to reach their satisfaction. This could also be said to help in establishing customers retention. These complaints as given by consumers could also be considered by the fintech circle due to its effective use of big data.

Following the financial crises of 2007/2008 and to reach at a better financial service for the sake of consumers, the Dodd-Frank wall street reform and financial reform was established in 2010 by the then-president Obama (Puiu and Puiu; 2016). Harvey (2019), stated that the act also led to the establishment of the Consumer Financial Protection Bureau (CFPB) in 2011 which was set out to protect consumer on the various financial product/services offered by banks. The CFPB serves as a base for receiving complaints from consumers on the various available financial products to help promote financial product thereby revealing financial products malpractice as was seen in the case of Wells Fargo in 2016 (Harvey,2019).

Basically, in most countries, it could be said that great priorities are being given to consumers complains regarding financial products and services they receive to help create value for financial products and promote consumer retention. Despite such effort, some responses made on some complaint are still said to be disputed by the consumers and this is seen to be a prevailing issue regarding financial products and services (Suomi and Jarvinen; 2018). Ibid also went further to indicate that about twenty European countries have also prioritized the need to secure their various financial services through the help of the consumers complaints, therefore, leading to the establishment of the Financial Service Pension Ombudsman (FSPO). This background thus leads the study to its research problem.

1.2 Problem of study

The research problem is centred in the finance sector where various complaints are made by consumers about different financial products. The problem also goes further to indicate that some responses as provided to the consumers about their complaints on certain products are still disputed by the consumers. Firstly, it is to an extent wrong for consumers to complain about any financial product or service and then, worse when such

consumers dispute the responses given to them by the financial provider regarding the product. Fonseka et al. (2016), also identified these issues as stated above as a matter that requires action to increase financial consumer service and enhance quality. It also went further to examine these issues using some analytical approach to consumers complaints.

1.3 Motivation of study

In todays economic environment, the financial sector is said to be one of the most challenging sectors among all others as it is squarely centred on consumers. Due to this fact, various issues regarding financial products like the loan, mortgage, credit card, interest rate and others are being recorded in the sector based on its services rendered to the consumers. Therefore, the need to carefully identify and analyse such problems, its channels of complaints, responses and especially the disputed responses by the consumers is very essential for the future purpose of good service, satisfaction and retention which is a necessity as it could be considered a driving force to help grow the quality of financial services and products. Secondly, this study is also being carried out with the notion of seeking how helpful will these complaints be to fintech particularly its start-ups in offering better financial services and products.

1.4 Research question and objectives

How well can classification methods based on machine learning algorithms in terms of accuracy predict the fact that consumers will dispute responses regarding financial products and services?

The specific objectives of this study include;

- To develop models that can be used for predicting the likelihood of consumers disputing responses for complaints made on financial products.
- To investigate and compare performance based on accuracy between models built and select the best fit model for future predictions.

This study applies three machine learning algorithms on a large customer complaint dataset of 18 variables and over 500,000 columns to help answer the research question and achieve its objectives.

1.5 Importance of study

There are various significant attributes linked to this study. One of the major importance of this study is to help increase business benefits/power either to the fintech circle or to the larger financial institutions. This study will enable both organizations to better understand the various financial issues faced by consumers. Such understanding will enable them to implement better solutions and complaint management procedures to help increase the quality of financial products and services thereby leading to consumers retention. Secondly, this study will serve as a valuable resource to academics who wish to carry out further studies regarding various major financial issues affecting consumers and the quality of financial services offered by organizations.

For a better understanding and clarity of this study, a structured flow of information will be outlined and used to enable the reader to fully understand the whole concept

of the study without a doubt. Section 1 is the introduction which gives an overview of the research in question. Section 2 is centred on the existing state of the art that is the literature review. This section will point out related works carried out which indicates the novelty of the study. Section 3 outlines the methodology to be implemented followed by the design specification indicated in section 4. Section 5 and 6 covers the implementation and evaluation of the study respectively. This moves down to the discussion section and finally the conclusion and future work to help further future research of the problem study.

2 Related Work

The issue of consumers disputing responses regarding complaints made on certain financial products and services has being a trending factor in the financial sector for some years now (Jung et al.; 2017).For this reason, various studies have been carried out and are still being carried out due to its urgency in trying to analyse such complaints, establish solutions and educate consumers on such issues. Such disputes to certain responses could be said to have led to consumers awful behaviour. In 2010, the United States government saw the need to protect consumers against some unfair acts experienced through the service of the financial providers (Harvey, 2019). The study further noted that such led the United States government in establishing the Consumer Financial Protection Bureau (CFPB) on 21st July 2011. Since the establishment of this agency, a querulous tone of complaints is being registered and these complaints have kept the financial providers awake in providing quality service and giving a timely response to issues. Based on this, various analytical studies have tried to analyse tones of complaints using the United States CFPB data. This section will try to review various academic works which are related to this study, draw a comparison between their various findings and then, provide a reasonable conclusion on the general findings.For this purpose, the review will be divided into two sections;

- Studies based on customers complaints.
- Studies based on methodology applied.

2.1 Studies based on consumers complaints.

This review will cover various areas such as consumers satisfaction, behaviour, disputed responses but not limited to the CFPB. Basically, all these are said to be related to one another because the event of one occurring could lead to the other occurring. Bastani et al. (2019), referred to the concept of consumer complaints as a negative statement uttered by a consumer with the end goal of communicating vital information about a product or services that are not satisfactory enough. The study further examined the various implications associated with complaints ignored by the financial providers and hence called out for the need to improve consumers satisfaction. Panther and Farquhar (2004), also made known the fact that such statement (complaints) triggers dissatisfaction and hence, results to the concept of consumer behaviour which is said to be dependent on consumers loyalty and vice versa. Felix (2015), tries to establish the relationship that exists between the terms consumers behaviour and consumers complaint thereby analysing the effect of one on another. The terms as identified, are regarded as a polling strength

by which various financial providers could use in building up its service quality to prevent consumers disputing response on complaints (James et al.;2019). Yang et al. (2009), criticized this statement by examining consumer complaint as a threat to financial providers. Gartenberg and Pierce(2015), stated that **45**% of responses(feedback) produced by the financial providers are said to be disputed by the consumers and this has raised various challenges on the products in question, but the study failed to identify reasons why such responses were disputed by the consumers. Fierro et al. (2016), also identifies the need for proper handling of complaints, particularly in the finance industries. The study stated that **27**% of responses disputed was a result of late response while about **18**% was due to unsatisfactory products and services. These studies, in general, pointed out the benefits and threats associated with complaints regarding financial products, but a little emphasis was centred on the disputed responses.

James et al. (2019), identified that one of the first issues to be considered by financial providers is the concept of consumer dissatisfaction/ satisfaction. An understanding of this concept helps to make a decision on reaching customers satisfaction and creating out rules for proper management of complaints. The study explains the fact that dissatisfaction is an alternative way of erasing consumers retention. According to the study of Jung et al (2017), the study made it known that responses disputed by consumers are said to be one of the major issues affecting the activities of financial providers and these disputes are said to be far-fetched. Due to the various disputes arising from responses regarding complaints of certain products, the United State government sought the need to protect such consumer and as a result of this the CFPB was established (Harvey, 2019).

The CFPB is an agency established by the United States government charged with the duty and responsibility of handling consumers complaint raised against financial providers and in the long run protect such consumers McCoy (2019). The agency is also known for identifying potential problems that are tied to products. Consumer complaints are submitted to the CFPB for proper handling and these data(complaint) is said to be published to the public with the consent of the consumers (Gartenberg and Pierce,2015). The database created is seen to be the largest consumer database that is used for the various analytical task (McCoy, 2019).

2.2 Studies based on methodology applied.

The study of Fosenka et.al(2016), applied the data mining techniques on an analysis based on the prediction using the consumer complaints dataset. Ibid applied four techniques such as the Microsoft Decision Tree, Nave Bayes, Time Series and the Neural Network. Based on the study, it shows a relationship that exists between complaints and certain environmental forces which can be political or economic. The study also indicated the relationship of one variable on another. That is, how a financial product complaint is linked to a product and more. Bastani et al. (2019), also applied the Latent Dirichlet Allocation known as the LDA on various consumers complaints recorded in the CFPB database. The study applied a probabilistic approach to help analyse narrative to each of the given variables. Findings such as how timely response could affect disputed responses were identified. Both studies showed the relationship why an issue could affect the other leading to their relationship but failed to outline the prominent issues behind such complaints.

Ekinci et al. (2016), carried out two separate research based on this issue. First, the study tries to determine consumers personality with a consumers intention to complain

and secondly, products in relationship to complain. The study centred its analysis to determine if a product issue was as a result of a certain product. This analysis was carried out the use of logistic regression. The results showed that products price could influence consumers behaviour and, have also on consumers with new intentions. Manisa et al. (2016), developed a model for analysing performance based on consumer complaints. The study was centred on the United States airline industry. The study points out how complaints could be used for predicting performance. This was carried out with the use of the decision tree algorithm. The results showed that consumer complaints and performance are inversely related to each other and as such, the study calls out for the need for a better managerial structure to enhance performance. Also based on consumer complaints which are said to influence their loyalty, the study of Moedjiono et al. (2016), established a model that could predict consumers loyalty. The study further stated that consumers loyalty is dependent on the certain product offered. The study made its focus to the multi-media service and this could also be said to be related to the finance sector where consumers loyalty could be seen based on financial products and services.

In conclusion, having reviewed the various studies, certain conclusions can be drawn. First, each study reflects the importance of consumers and their complaints in every given sector as this could be used to determine consumers behaviour, satisfaction and sane time identify the quality of products and services offered to consumers. Based on this study, only but a few studies have tried to build a model that can predict the likelihood of consumers disputing complaints responses of financial service providers based on products and services.

3 Methodology

3.1 Research procedure

This section illustrates the process to be carried out for this study from the understanding of the subject area (consumers complaints of financial products and services) to the deployment of the most appropriate predictive model. For this purpose, the methodology carefully selected to help in achieving the aim of this project is the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. The CRISP-DM involves six stages which are regarded to be hierarchical in nature and these stages will be applied in the course of carrying out a data mining project (Wirth and Hipp; 2000). This methodology was also selected because it has proven to help solve various challenges as faced by industries and as such, it is widely used by numerous industries, especially for predictive analysis. A detailed procedure of this study using the selected methodology will be explained below.

3.1.1 Business understanding

The first stage of the methodology chosen for this data mining project is the understanding of the business segment. For this study, the business understanding is solely on the financial products and services provided to consumers. Understanding of the business area shows that some of the products or services provided to the consumers could be regarded to be unsatisfactory and as such, various complaints have been raised regarding the products or service used by such consumers. As earlier stated, it is fair for a financial product provider to receive complaints about a product but worse when response provided

after such complaints raised regarding the products are disputed by the consumer. For this study, examining consumers complaint as an issue in the finance sector will help in understanding the study aim and requirements from a business point of view. At this stage, the problem of consumer complaints was converted into a data mining problem of predicting the likelihood of a consumer to complain of a financial product or service. The understanding of this phase then leads the study to the data understanding plan.

3.1.2 Data understanding

Based on the hierarchical stage of the selected methodology for this project, this stage deals with the understanding of the data to be used to help achieve the studys objectives. This study will begin with the collection of secondary data sourced from Kaggle. The secondary data will be obtained by downloading a published dataset of the U.S Consumers Financial Complaints which will be followed by the exploration of the selected data. The exploration will help give a better understanding of the data attributes, size and structure. The exploration will further help in selecting the variables to use while having in mind the problem and objective of the study in question. Further understanding of the data for this study will uncover data quality issue and insight. This will finally help to determine the type of data mining technique to be used for the study to help achieve a reasonable result.

3.1.3 Data preparation

The data for this study will finally be prepared after an understanding of the data. For this study, the data is planned to be prepared for its future modelling. Variable selection will be carried out to help select the needed variables. This is to say that some variables will be excluded and the rationale for this is because some variables in the data are seen to have less significant information while some will also be excluded because of its unique identification. Also, missing values will be checked to enable proper modelling. The raw data will be transformed to help deal with its categorical variables which could lead to derived attributes. In nutshell, the raw data will be cleaned to help achieve the best possible result during its modelling (Chu et al.; 2016).

3.1.4 Data modelling

After carrying out the data preparation to an attainable point, the data will thus present its Xs (independent variables) and Y (target variable) which will further be split into the testing and training data using a certain percentage for validation. The training will be allocated a certain percentage while the testing will be allocated its percentage though smaller. The training is allocated higher to enable the algorithm to learn from the higher part of the dataset. This splitting will enable the selected modelling technique to be applied. This study will be limited to three modelling techniques selected from other modelling techniques used for a data mining problem. The first technique is the Naive Bayes classifier also called the generative model which is known to use categorical input variables for developing a predictive model (Akinori and Ueda; 2016) while the other is the Random Forest also known as the discriminative or conditional model which uses numeric input variables(Silke et al.; 2016) and the third the Logistic regression. These models will, therefore, generate results to show its performance.

3.1.5 Model evaluation

For every data modelled for any analytical task, such model(s) are to be evaluated. The models to be deployed for this study will be evaluated because this will form a crucial part of the study process as required. The models to be used will be evaluated to ascertain its performance and this will be carried out using the confusion matrix based on its accuracy.

These metrics are to be used because the study solely deals with classification models.

3.1.6 Deployment

After carrying out a comprehensive modelling with its evaluation, the knowledge acquired from the results gotten will be documented and presented for its use. This could also mean that the models used will be compared and the one which is seen to have performed better will, therefore, be recommended to the finance sector particularly the financial service providers. This will enable them to use this work to identify the key issues and serve as a possible way of creating better solutions to deal with the identified issues thereby providing a better financial service and products to consumers which could eliminate complaints and disputation.

3.2 Software tool

For this study, the programming language or tool selected is R. The reason for this is that it deals with various libraries that could be used for any analytical task. Another is that it produces results that can be easily understood. This programming language is regarded as a well know programming software for data analysis and manipulation. R directly encourages two diverse object-oriented programmings (OOP) paradigms. Many operations in this programming language (R) could be vectorized and appreciating. The R will help in data manipulation, modelling and give help give a clear view of the data structures as to be seen in the course of this study. Nevertheless, R is itself a complete programming language, with its idioms much like varieties of other programming languages. In some ways, R is a functional programming language, although it is not fully functional (Zhifang and Yajing; 2018).

4 Design Specification Applied

For proper implementation of the models outlined for this study, the study will provide a comparative analysis between three machine learning models which will be regarded as the techniques to be used in this study. Also, for better understanding and clarity, the machine learning models can be referred to in the study as machine learning classifiers or classification methods since the study deals with a classification problem. This could also be said that the study deals with supervised learning because the dataset to be used have a given label for each instance and such label (output variable) are categorical in nature. These techniques are;

4.1 Naive Bayes Classifier Algorithm

The Naive Bayes Classifier is regarded as one of the machine learning algorithms which can be used for most predictive task and is known for its performance and reliability

(Chandrasekar and Qian; 2016). This machine-learning algorithm applies the use of the Bayes Theorem which takes into consideration of the fact that its attributes are independent of others (Jadhav and Channe; 2016). By independent it means that other variables cant be known easily even with an addition of another variable. In other terms, it could mean that a specific variable has no relationship with the rest variables even if a link exists among the variables. Nevertheless, history has it that this algorithm is known for its results and this explains one of the reasons for its use in this study.

It is important to remember that one of the objectives of this study is to help develop a model for the prediction of consumers disputing the response of financial companies to complaints on financial products. The Naive Bayes is, therefore, a technique or model selected in helping to achieve this objective. For this purpose, its performance based on this predictive task will be ascertained and compared with the other machine learning model to be able to conclude on what model is best regarding predicting consumers dispute to the response of financial companies to complaints on financial products.

The justification for the selection of Naive Bayes Classifier in this study is based on some reasons. First, Naive Bayes is a powerful algorithm which is known for its performance or accuracy in carrying out predictive modelling when dealing with categorical input variables (Jiang et al.; 2016). Jadhav and Channe(2016), also stated that this model design is one of the easiest to understand and comprehend especially when being described using either binary or categorical variables as to be seen regarding this study dataset. The Naive Bayes algorithm which is also known as a generative model is said to converge quicker compared to some discriminative model(Hanjun and Song; 2016). Jiang et al(2016), also stated that this technique acts best when dealing with classification problems as regards to this study. Lastly, the dataset to be used for this study is seen to be very large and this technique has been proven to work well with very large dataset (Wang et al.; 2015). Based on these justifications regarding scholarly articles used, the Naive Bayes algorithm was selected to help achieve the studys objectives and could act as a better model.

Yang and Webb (2002), identified that this modelling technique (Naive Bayes) applies the use of discretization to help achieve better performance. By discretization, it implies that the variables which are regarded to be numeric are changed to be categorical in nature (categorical variables). This Naive Bayes model is seen to be compared in this study with Random forest which is known to use a numeric response variable when dealing with classification problems.

As earlier stated, the Naive Bayes algorithm makes use of the Bayes Theorem as shown in equation one below. This Bayes Theorem deals majorly on the probability of an event occurring and such an event could be as a result of the previous occurrence which could be related to the event (Chandrasekar and Qian, 2016). It predicts the probability of a sample to belong to a class. Jadhav and Channe(2016), states one of the assumptions of the Naive Bayes is that it assumes conditional independence of the input variable. Based on this study, the Bayes formula will be applied during the implementation of the model in R which usually happens in the back-end. The equation is thus;

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \tag{1}$$

Where;

- P(A/B) is the possibility of A to occur when B has occurred.

- P(B/A) is the possibility of B to occur when A has occurred.
- P(A) is the possibility of A.
- P(B) is the possibility of B.

A and B are regarded as events that occurred or yet to occur and each are said to be independent of each other following the assumption of Naive Bayes explained previously.

4.1.1 The Architecture of Naive Bayes

Based on Naive Bayes architecture, the data is split into the training and testing where the data target variable is of two-class classification problem as in the case of the data for this study. Both are computed using the log of each probability and the maximum performance is reproduced as the expected output.

As stated earlier, the Naive Bayes assumes the conditional probability of an event given that another has happened based on its assumption thereby given rise to calculate the probability of the initial event. As a result of this, sequential events are mostly the focus when dealing with classification probability thus leading to the use of prior and posterior probability calculation which shown in equation two below;

$$Posterior = \frac{likelihood * Prior}{Evidence} \tag{2}$$

The use of applying the conditional independence which is a Naive Bayes assumption automatically assumes that its variables(input) are independent. Stating the formula as;

$$\begin{aligned} &P(x_1, x_2, , x_n\ c) = P(x_1\ x_2, , x_n, c)P(x_2, , x_n\ c) \\ &\quad = \mathrm{P(x_{(}1)}\ c)P(x_2, , x_n\ c) \\ &\quad = \mathrm{P(x_1}\ c)P(x_2\ c)P(x_n\ c) \end{aligned} \tag{3}$$

4.2 Random Forest

Random Forest is also a machine-learning algorithm to be applied in the course of this study as it could be regarded to be a valuable technique for most business analytical task. This model is also regarded as a powerful algorithm used for predictive purposes (Jose and Gopakumar; 2019). This algorithm is also referred to as Random Decision Forest and can also be described as a discriminative model (Fang et al., 2019). By being a discriminative model, which could also be referred to as a conditional model, it means that the model is solely built on the observed data(Fang et al.; 2019). This model (Random Forest) is a supervised learning that falls under the classification model as in the case of this study which deals with a classification problem. The target variable for this study is seen to be categorical that is, being identified as Yes and No. At the training stage of modelling, it is planned that this algorithm will build various trees which will be chosen at random by the algorithm itself as its desired decision (Jose and Gopakumar, 2019). This model takes n to be the number of observation and p as the number of features in the training phase while k¡¡p is used as a variable selector in determining the node.

The motivation behind the selection of this algorithm is based on various reason as related to the study being carried out. One of the reasons is that Random Forest is said

to work effectively with large dataset especially in todays finance world where large data is mostly used (Angshuman et al.; 2018). This study applies the use of large data which deals with a good number of variables for its analysis. Ibid also identifies the fact that this model deals better with input variables that are large enough. This model also tries to estimate data that are seen to be missing (Fang et al.;2019). By estimation, it means that its performance or accuracy can be retained despite the presence of missing values. The use of this model can also be justified because it does not overfit while modelling (Angshuman et al.;2018). Ibid also identified that the model can deal with a dataset that is seen and proven to be unbalanced. Research has shown that when it comes to the training phase, Random Forest doesnt deal with much training time but gives less training time.

4.2.1 The Architecture of Random Forest

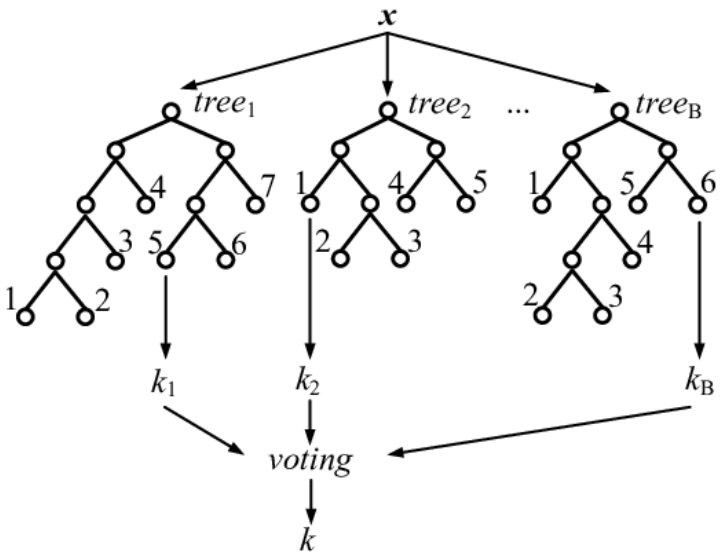

Figure 1: Architecture of Random Forest (Jose and Gopakumar, 2019)

Figure 2 above shows the framework or architecture upon which the Random Forest algorithm works. This architecture involves 5 phases or stages to give the desired output.

The letter x is seen as the first phase. The x signifies the test sample which is also regarded as the test sample input or data. The data is imputed to build the model. These test samples are then created into various trees such as tree1, tree 2 till the last tree it can create which is the second phase. These trees created also builds various trees which can be called the forest trees. On the third phase, decisions are made based on each group of trees created as represented in figure 2 above as k1, k2 and kb. These decisions are also referred to as the predictions which are carried out from each class of tree created. Based on the predictions made from the multiple classes of trees, majority rule is applied. This is the next stage which requires voting for selection. At this stage, the average on all predictions are taken and finally moves to the last stage which is the decision stage represented by the letter k. At this point also, the random forest decision takes place. This decision is made from multiple predictions carried in on different classes. The description of the structural architecture of the random forest is simply an overview

given to note how it will be implemented to help achieve the prediction task and other objectives.

4.3 Logistic Regression

The model or algorithm is commonly used for an analytical task. Statistically, this model increases the log-likelihood to enable it to perform well with the selected data (Neslin et al.; 2006). It shows that it works just as the linear regression just that it deals on variables that said to be categorical such as Yes or No, 1 and 0 and others as can be seen in this study. This model makes use of the receiver operating characteristics curve which gives the function as;

Y = f(x)

This simply means that y is the function of x which is unknown. Y represents the binary which shows only two possible cases(Peng et al.; 2002). These two states can be referred to (1, 0) data. Logistic model has a logit function;

$$g()i.e. : g(t) = 1/(1 + e^{(} - t)) \tag{4}$$

5 Implementation

For this study, the Consumer Financial Protection Bureau dataset was sourced from Kaggle. This dataset comprises of consumers complaints on various financial products which range from 2011 to 2016. The dataset comprises of over 500,000 rows and 18 columns [1].

After selecting the dataset to achieve the stated objectives, the dataset selected was read into the R studio using the read.csv function. Before reading in the data file into R, several libraries such as psych, caret, nnet, rocr and others were installed. Just after reading the data, data exploration was carried out to have a better understanding of the data to be used. The data read was then viewed alongside with its structure. This is usually done to have a clear knowledge of what the data consists of to identify its rows, columns, variables both dependent and independent and other features(Shichao et al.; 2003). After a clear view of the data structure with its dimension, irrelevant variables such as data sent, date received, company, zip code and five others were removed. These variables were removed based on the assumption that they do not affect the target (output) variable. Variables such as zip code and complaint identity were regarded to be a unique identity. Based on the dataset, it was clear that some columns contained no information and as such were also removed.

Categorical variables having more than 53 levels were also removed to make the dataset executable by the random forest classifier algorithm. These include variables such as issue, sub-issues and others. Rows with missing values were also removed from our dataset and as a result, the dataset was reduced to 2,422 observation and 9 categories. The study proceeded with omission, in contrast to imputation because the derived dataset was relatively large and hence still appropriate for the analysis. Data frame subsetting was used to eliminate some variables, and this is regarded as a vital aspect of data management. The na.omit function was used to remove missing values, leaving the study with a consistent and executable dataset.

[1] https://www.kaggle.com/cfpb/us-consumer-finance-complaints

Based on the summary statistics of the pruned dataset, it revealed that there were more missing values in some columns than present bearing in mind that any form of prediction algorithm comes with its error. Due to this fact, predicting missing values, in this case, would but increase error and this would, in turn, accumulate or spill over to the prediction modelling phase. Actions carried out such as the omission of missing value and removal of variables are regarded as the data cleaning and transformation. Also, since the study is dealing with a classification problem and considering the size of the data, dummy encoding would be superfluous as the same result is bound to be achieved despite using both methods. This explains why dummy encoding was not done in this analysis, and the levels were used as characters (nominal).

Also, just before the cleaning and transformation of the selected data, the data was explored using the pairs panel. This was done to give a better overview of the data because the dataset to be used is very large. Data exploration is a procedure seen to be like any initial data analysis. The exploration was performed on the dataset to give both the numerical and visual summary of our variables contained in the dataset. The summary function was used to give the numeric summary of our dataset. The use of plots was also implemented to give the visual summary of the variables like products, sub-products and others. This plot was used because it is known to work better with categorical data bearing in mind that the ggplot function was used for this.

Data partitioning was implemented for the study purpose. The data selected was therefore partitioned into the training and testing data. This was done because according to Shichao et al(2003), data is split to help evaluate the model using a sample set and when dealing with a large dataset. For this study, the training data was apportioned a higher percentage of 70 while the test data was apportioned 30 percent. The reason for training the data at 70 percent is because the model learns from this data and it contains outputs which are known while the testing data was selected at 30 percentage because it requires only but a little set to help test or validate the data(Shichao et al.;2003). In other words, it will be used to make a prediction.

This study applied the concept of replicability. This simply implies that the study during its data partitioning had to set the seed using the set seed function. The purpose for this was to ensure that the same result is obtained every time the code is run thereby making it possible for anyone with the analysed dataset to verify the results of this work. In summary, set seed affords uniformity.

After preparing the data to enable it to kick off its modelling, three machine algorithms were used. These algorithms are Naive Bayes, Random Forest and Logistic Regression. These algorithms were used to evaluate or determine the best performing model. This implies that its selection was done based on the algorithm with the highest accuracy rate using its confusion matric.

Naive Bayes also called the probabilistic classifier which employs Bayes theorem was used in developing a model to correctly classify a consumers disputed response as either a yes or a no. Also, while using this model, the summary of the model was carried out showing the call. This call displayed showed an overview of the iterations such as the Laplace, classes, samples, features and prior probability. The model was run with the training dataset. The generated model was validated with the test dataset by the creation of a confusion matrix to illustrate the proportion of misclassification from which the accuracy was computed. The higher the misclassification error, the more inaccurate the model is incorrectly classifying a response as either a yes or a no. In general, an initial plot of the model was included to show the extent to which misclassification was present

in the model which will help determine its accuracy. Along with the plot, a mathematical equation to numerically capture misclassification was added and this indicated the error ratio while its accuracy will simply be 1 minus error rate. A similar procedure was conducted on the test dataset and the accuracy was compared.

The random forest classifier, as the name implies, is another classification algorithm applied in this study. In determining the best model for the intended analysis, this model was generated to provide a basis for comparison. Before using this model, a validation set was carried out at random using 70:30 and a set seed function was used. Two random forest models were compared to select the best model based on estimated minimum error. Predictions with both train and test datasets revealed the extent to which error was present in the classification model. A confusion matrix for accuracy was developed and a plot of the variables of importance was made to give an insight into the level of importance of each predictor in the model.

The third algorithm applied to this study was the logistic regression. This algorithm was implemented with the use of multinom. This was applied to help identify any relationship existing between an output or outcome and a variable to enhance proper prediction of an experiment. For this model. The predict function was applied to this model.

In general, in evaluating model performance, a prediction probability was used. A ROC curve was applied for logistic regression. This was used with the so aim of checking model performance using visualization. To evaluate the performance of the model, a probability type prediction method was used. A Receiver Operating Characteristics curve was plotted to visualize the performance of the model.

Model comparison was carried out after the modelling and validation of models. This act is a purposive search for the best-suited model for classifying responses disputed by consumers regarding financial products such as the one under consideration. The choice was based on the accuracy rate which can be calculated as 1 minus error rate. Though the misclassification error rate can also be used.

6 Evaluation

Based on the application of the data mining approach used for this study, three algorithms as earlier mentioned are used systematically to help identify patterns from the selected dataset. These algorithms used for this study makes use of various iterations to deliver results or output. At this phase of the study, the results gotten from the use of the algorithms will be displayed and evaluated though not fully discussed as this will be done fully in the following section. Classification models are evaluated in various ways. This means that various matrix can be used to evaluate results derived from the use of classification models. For this study, the results will be evaluated based on the confusion matrix to calculate the misclassification error for each algorithm or model from which the accuracy of each model was computed. Accuracy is used for this study because its result is on par (equal value) with the standard that is acceptable. It is important to note that this model was split initially into training and testing data. The evaluation and results of the three model will thus be described below.

6.1 Naive bayes

For this model, a comparison will be drawn between the testing and training data having in mind that prediction was carried out with the test dataset. Table 1 below

shows the confusion matrix for the training dataset of this algorithm.

Training the data using train dataset:

Table 1: confusion matrix for the training dataset

		Actual NO	Actual YES
Predicted	NO	1386	335
	YES	1	2

Based on table 1 above, there are two possible predicted classes known as Yes and No. The classifier considered a total of 1724 observation in the training dataset, out of which the classifier predicted a No response 1721 times and a Yes response 3 times whereas in the actual sense 1387 were No and 337 were Yes. Based on the model, the misclassification rate or error rate is 0.1948956. This implies that the accuracy using the Naive Bayes classifier is 0.8051044. This is calculated as 1 minus error as earlier stated. This means that **80.51**% of the actual set was correctly predicted whereas **19.49**% were incorrectly predicted.

Predicting model using the test dataset:

Table 2: confusion matrix for the test dataset

		Actual NO	Actual YES
Predicted	NO	538	158
	YES	2	0

Table 2 below shows the Prediction of the consumer disputed responses using the model generated by Naive Bayes classifier for the test dataset. Based on this model using the test dataset, the accuracy is seen to be **77.08**% while the misclassification rate or error rate is **22.92**%. In summary, out of 698 cases, the model correctly predicted 538 No, and 0 Yes. It incorrectly predicted 2 No as Yes and 158 Yes as No.

6.2 Random Forest

For this model, the call was also achieved. The call simply gives an overview which shows the number of iterations. But in the case, we will be using the confusion matrix because the produced same results.

Training the model using the training dataset:

Table 3: confusion matrix for the training dataset

		Actual NO	Actual YES
Predicted	NO	1369	326
	YES	0	0

There are two possible predicted classes as earlier stated which are Yes and No. The classifier made a total of 1724 predictions for the training dataset out of which the classifier predicted No 1369 times and Yes zero times whereas, the initial set was No 1369 and Yes 326. Misclassification or error rate is 0.1923304 which implies that the accuracy of the classification using Random Forest classifier is 0.8076696. It explains that **80.77**% of the actual set was correctly predicted whereas **19.23**% was incorrectly predicted.

Predicting the model using the test dataset

Table 4: confusion matrix for the test dataset

		Actual NO	Actual YES
Predicted	NO	558	169
	YES	0	0

Table 4 shows the prediction of the consumers disputed responses using the model. Out of 698 cases, the classifier predicted No 558 times and Yes 169 times while in the initial sense, the No is 558 and Yes 169. Based on this, the accuracy of the test data is **79.94**% and the misclassification rate or error rate is **20.06**%.

6.3 Logistic Regression

Unlike the Random Forest and Nave Bayes where the confusion matrix was created for the test data to validate the result, the Receiver Operating Characteristics Curve (ROC) was used to validate the outcome.

Table 5: confusion matrix

		Actual NO	Actual YES
Predicted	NO	1926	493
	YES	1	2

The misclassification error or error rate from the matrix above is 0.2039637 which is **20.40**%. This algorithm also gives an accuracy rate of **79.64**%. This is also illustrated in Table 6 below.

Table 6: Accuracy rate

Accuracy	cutoff. 217507
0.7964492	0.4044142

Below is the ROC curve of the Logistic Regression which is used to validate the outcome. This is represented in figure 3

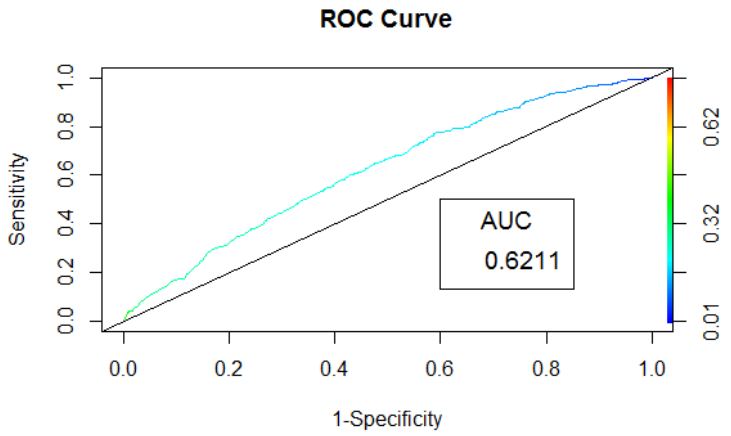

Figure 2: Receiver operating characteristics (ROC) curve

Based on the curve above, it shows that the curve is above the line as found on the y-axis. This can also be seen on the sensitivity axis which means true positive. With this curve, it can be said that the model is valid for prediction.

Model comparison The figure below shows the model comparison based on their accuracy rate.

Table 7: Model comparison based on accuracy

Algorithms	Accuracy rate %
Naive Bayes	80.51
Random Forest	80.77
Logistic Regression	79.64

Table 7 above, shows the accuracy rate using the training dataset on which the model was built. Considering the accuracy rate of the three models under study, the Random Forest Classifier has a higher accuracy rate of **79.94**% in the testing data and **80.77**% in the training data compared to the Naive Bayes classifier and the logistic model. This accuracy rate shows the rate at which the classifier correctly predicted the outcome variable. Thus, the study can say based on the produced results that the Random Forest Classifier is a preferred classification algorithm to the Naive Bayes and Logistic Regression in classifying and predicting the consumers disputing responses.

6.4 Discussion

This study began with writing out a research question and its stated objectives as a guide to carry out research. Based on this, a data miming methodology was selected and followed with the application of three algorithms to help achieve the stated objectives, answer the question and achieve a purposeful study. Before this, research was carried out to help discover areas in the financial sectors that can be improved or areas that require further studies. For this reason, the study was carried out. Various research has been carried out regarding consumers complaint as this important because it deals with consumers. As earlier stated in the introduction, this issue led the United State government in establishing an agency known as the Consumer Financial Protection Bureau (CFPB) to help protect consumers against ill act carried out by financial service providers. This shows how important this is and as such studies have been done and will continue to be carried out on this issue either for one reason or the other. The data published by the CFPB happens to be one of the biggest consumer data available for research.

This study applied three classification models to achieve its stated objectives based on previous research carried out and, on this note, various outcome or results were produced using the three selected algorithms which are Naive Bayes, Random Forest and Logistic Regression applied in this study. The results derived from the analysis carried out to help build a suitable model that can be used for predicting if consumers will dispute responses based on their complaints on financial products and services or not was due to some valuable factors. First, the scientific, systematic and efficient way such models were built for easy use and recommendation. Also, was as a result of the features selected from the dataset used for the study but one of the limitations to the analysis was the fact some columns in the selected data contained little or no information and this could be said to be valuable for the analysis and thereby having an impact of the results produced. Based on the design models implemented, the accuracy rate was high for all models especially for Random forest having a higher accuracy rate of 79.94 on the testing data and 80.76 on the training data. This shows that the Random Forest will be regarded as a model good enough to carry out a prediction on if consumers will dispute responses provided by financial service providers based on complaints made against financial services and products. Though having said that all models are good especially Random Forest which stands out as the best model to use for the stated prediction, improvement should be done by trying other classification models to see how best the models can perform better than Random Forest. Previous work carried out showed that Naive Bayes performed well with an accuracy rate of **78**% but in this case, Random Forest showed an accuracy rate of **80.76**% using the training dataset. Therefore, the likelihood that another model could perform better for the stated prediction still stands out.

As a suggestion, financial service providers should set condition that will help manage and possibly prevent factors that could arise and lead to consumers disputing responses based on complaints made as this could have implication on the quality of their services and on consumers retention as in the case of Wells Fargo where complaints made by consumers revealed their acts. The quality of a financial product or service reflects strongly on consumer and their behavioural act toward the products. It can also be said to be determined by their level of satisfaction using such service.

This study can be used by the financial service providers and even to the fintech circle whose aim is to provide better financial services to consumers. This will help both make an informed decision on service management to avoid future issues of their products or

services. Also, this study could be used to determine the possibility or likelihood that their respective consumers will dispute their responses based on the current features surrounding their issue under prediction.

7 Conclusion and Future Work

This study is carried out with the objective of building models that can be used for prediction on consumer disputing responses. The analysis carried out in this study gave several results based on the three algorithms used and based on such analysis, Random Forest gave a better accuracy implying that it stands to be a better model to use for the prediction.

This study can clearly state that its outlined objectives are met. First, the study was able to build several models using classification models which are seen in the results to be of high accuracy and the second objective which is to select the best fit based on its accuracy, this study presented Random Forest as the best among the models used. In trying to achieve the stated objective of which the study did, it also answers the research question. Yes, classification models can perform well when carrying out a prediction on consumers disputing complaint responses. Each model had a high accuracy rate which shows how effective classification models are for the prediction.

One of the key findings of this study is that timely responses given by the financial service providers to the consumers based on their complaint on financial products and services have little or no effect on their behaviour. This shows that the quality of financial products and services can be said to be dependent on consumer satisfaction. In other words, consumers complaints are a means of measuring financial service quality and this can be seen in the way various countries especially the United States handles issues relating to financial products in relation to consumers and their satisfaction.

One of the implications of this study is that this work based on the models built can be used in the financial sector for future predictions regarding disputed responses and this will enable the financial providers to create out an efficient and effective way of handling complaints and also create better financial services and complaint responses if need be.

Further research can be carried out by trying to find out if the responses disputed by consumers based on issues arising from a financial product are traceable or related to a particular product issue having in mind that disputed responses could either be negative or positive also known as yes or no. Also, further model can be built to help ascertain one with the highest rate of performance.

Acknowledgement

My sincere gratitude goes to my supervisor Mr Noel Cosgrave who gave his time, supported and extended his knowledge to me while carrying out this study. During this study, he was always willing to accept questions and give valuable feedback to me and went to the extend of making available another day to meet with his supervisees.

Also, I will like to appreciate the Fintech career advisor of National College of Ireland in person of Kate Honan who encouraged me to accept the challenge of doing the thesis. I am very grateful for your words of advice and I am happy to carry out this study.

References

Akinori, F. and Ueda, N. (2016). 'Semi-supervised auc optimization method with generative models, *in 2016 IEEE 16th International Conference on Data Mining (ICDM).* , Barcelona, Spain, 12-15 dec 2016, pp. 105-115, IEEE Xplore. doi: 10.1109/ICDM.2016.0107.

Alex, K., Ruud, V., Marrije, W., Bart, L. and Ronald, J. (2016). Quality quandaries: Improving a customer value stream at a financial service provider', *Journal of Quality Engineering* **28**(1): 155–163.

Angshuman, P., Dipti, P. M., Das, P., Abhinanda, G., Appa, R. C. and Saurabh, K. (2018). Improved random forest for classification, *Journal of IEEE Transaction on Image Processing* **27**(8): 4012–4022.

Bastani, K., Namavari, H. and Shaffer, J. (2019). Latent dirichlet allocation(lda) for modelling of the cfpb consumer complaints., *Journal of Expert System with Application* **27**: 256–271.

Chandrasekar, P. and Qian, K. (2016). 'The impact of data processing on the performance of nave bayes classifier, *in 2016 IEEE 40th Annual Computer Software and Application Software (COMPSAC)* , Atlanta, United states, 10-14 june 2016, pp. 35-48, IEEE Xplore. doi: 10.1109/COMPSAC.2016.205.

Chu, X., Ilyas, I. F., Krishnan, S. and Wang, J. (2016). 'Data cleaning: Overview and emerging challenges, *in Proceedings of the 2016 International Conference on Management of Data (ICMD).* , San francisco, United states, 26 june, pp. 2201-2206., IEEE Xplore. doi: 10.1145/2882903.2912574.

Chugani, S., Govinda, K. and Ramas, S. (2018). 'Data analysis of consumer complaints in banking industry using hybrid clustering', *in 2018 2nd IEEE International Conference on Computing Methodologies and Communications (ICCMC.*, Eroda, India, 15-16 february 2018, pp. 74-78, IEEE Xplore. doi: 10.1109/ICCMC.2018.8487638.

Ekinci, Y., Calderon, J. and Siala, H. (2016). Do personality traits predicts complaining consumers, *Journal of business environment* **8**(1): 32–39.

Fadzlan, S. and Shah, M. (2014). Accessing the impact of financial crisis on bank performance, *Journal of Economic Science,* **27**(3): 245–262.

Fang, D., Jiangshe, Z., Junying, H. and Rongrong, F. (2019). Discriminative multi-modal deep generative model, *Journal of Knowledge Based System* **173**(5): 74–82.

Felix, K. (2015). Factors for customer satisfaction and customer dissatisfaction in commercial banks, *Journal of Social Science* **6**(4): 251–257.

Fierro, J. C., Melero, I. and Sese, F. J. (2016). Can complaint handling effort promote customer engagement, *Journal of Service Business* **10**(4): 847–866.

Fonseka, W. R., Nadeesha, D. G., Thakshila, P. M., Jeewandara, D. M. and Wijesinghe (2016). 'Use of data warehousing to analyse customer complaint data of consumer financial protection bureau of united states of america', *in 2016 IEEE International Conference on Information and Automation for Sustainability (ICIAFS)*, Galle, Sri lanka, 16-19 dec 2016, pp. 152-158, IEEE Xplore. doi: 10.1109/ICIAFS.2016.7946520.

Gartenberg, C. and Pierce, L. (2015). Sub-prime governance: Agency cost in vertically integrated banks and the 2008 mortgage crises, *Journal of Strategic Management,* **38**(2): 300–321.

Gilreath, Z. (2018). The culprit of the great recession: A detailed explanation of mortgage backed security, their impact on the 2008 financial crises and its aftermath on the banking service, *Journal of Business and Technology,* **13**(2): 318–336.

Hanjun, B. D. and Song, L. (2016). 'Discriminative embeddings of latent variable models for structured data, *in Proceeding of the 33rd International Conference on Machine learning (ICML).* , New york, United states, 2-4 may 2016, pp. 6-11, IEEE Xplore. doi: 10.1109/COMPSAC.2016.205.

Harvey, H. (2019). Constitutionalizing consumer financial protection: The case of the consumer financial protection bureau, *Journal of Management Information System* **103**(6): 2429–2476.

Jadhav, S. C. and Channe, H. P. (2016). Comparative study of knn, nave bayes and decision tree classification techniques, *International Journal of Science and Research* **5**(1): 1842–1845.

James, W., James, H., Babin, J. and Parker, M. (2019). Is customer satisfaction really a catch-all? the discrepancy between financial performance and survey results., *Journal of Managerial Issues* **31**(2): 137–150.

Jiang, L., Wang, S. and Zhang, L. (2016). Deep feature weighting for nave bayes and its application to text classification, *Journal of Engineering Applications and Artificial Intelligence* **52**(3): 26–39.

Jose, C. and Gopakumar, G. (2019). 'An improved random forest algorithm for classification in an imbalanced dataset, *in 2019 International Conference on Asia Pacific Radio Science Conference (AP-RASC)* , New delhi, India, 9-15 march 2019, pp. 115-125. IEEE Xplore. doi:10.11.9/AP-RASC.2019.007.

Jung, K., Garbarino, E., Briley, D. and Wynhausen, J. (2017). Blue and red voices: Effects of political ideology on consumers complaining and disputing behaviour, *Journal of Consumer Research* **44**(3): 477–499.

Kauffman, R. J., Parker, C. and Weber, B. (2018). On the fintech revolution: Interpreting the forces of innovation, disruption and transformation in financial services, *Journal of Management Information System,* **35**(1): 220–265.

Lui, M., Zhang, L. and Keh, H. (2019). Consumer responses to high service attentiveness: A cross-cultural examination, *Journal of International Marketing* **27**(1): 56–73.

Manisa, S. U., Anitsal, M. and Anitsal, I. A. (2016). A model of business performance in the us airline industry: how customers complaints predict the performance, *Journal of Business Studies* **8**(2): 96–111.

McCoy, P. A. (2019). The assault on the structure of the consumer financial protection bureau, *Journal of Marketing Research* **103**(6): 2543–2615.

Moedjiono, S., Isak, Y. and Kusdaryono (2016). 'Customer loyalty prediction using k-means segmentation and c4.5 algorithm, *in 2016 International Conference on Informatics and Computing(ICIC)* , Mataram, Indonesia, 28-29 oct 2016, pp. 210-215. IEEE Xplore. doi:10.1109/IAC.2016.7905717.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J. and Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models, *Journal of Marketing Research,* **43**(7): 204–211.

Panther, T. and Farquhar, J. D. (2004). Consumer responses to dissatisfaction with financial service providers: An exploration on why some stay and while others switch, *Journal of Financial Service Marketing* **8**(4): 343–353.

Peng, C. Y., Lee, K. L. and Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting, *Journal of Educational Research* **96**(1): 3–14.

Puiu, C. and Puiu, V. (2016). Estimates on measuring the consumers satisfaction, *Journal of Management Information System,* **9**(3): 209–216.

Shichao, Z., Chengqi, Z. and Qiang, Y. (2003). Data preparation for data mining, *International Journal of Applied Artificial Intelligence* **17**(5): 114–125.

Silke, J., Gerhard, T. and Boulesteix, A. (2016). Random forest for ordinal responses: Prediction and variable selection, *Journal of Computational Statistics and Data Analytic* **96**(3): 57–73.

Suomi, K. and Jarvinen, R. (2018). Consumer complaint in the financial sector, *Journal of Research for Consumers* **33**(4): 38–82.

Wang, S., Li, C. and Jiang, L. (2015). Adapting nave bayes free text classification, *Journal of Knowledge and Information System* **44**(1): 77–89.

Wirth, R. and Hipp, J. (2000). 'CRISP-DM: Towards a standard process model for data mining, *in 2000 Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (ICPAKDDM).* , Berlin, Germany, 18-20 june, pp. 29-39., IEEE Xplore. doi: 10.1024/ICPAKDDM.2000.7942220.

Yang, S. C., Chiayu, T. and Suechin, Y. (2009). Exploring the solution-the contextual effect on consumer dissatisfaction and innovativeness in financial service companies, *Journal of Service Industries* **29**(4): 557–568.

Yang, Y. and Webb, G. (2002). 'A comparative study of discretization methods of nave bayes classifier, *in 2002 Proceeding of the 2nd International Conference on the Pacific Rim Knowledge Acquisition Workshop (PKAW)* , Tokyo, Japan, 15-16 august, pp. 159-173, IEEE Xplore. doi:10.459/COMPSAC.2002.1.

Zhifang, H. and Yajing, D. (2018). 'Application of r programming for bayesian discriminant method in effective teaching, *in 2018 9th International Conference on Information Technology in Medicine and Education (ITME)* , Hangzhou, China, 19-21 oct 2018, pp. 728-732, IEEE Xplore. doi: 10.1109/ITME.2018.00165.