Optimisation of credit rating process

MSc Research Project

MSc. In FinTech

Amarnath Venkataramanan

Student ID: x18105149

School of Computing

National College of Ireland

Supervisor: Mr Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Amarnath Venkataramanan		
Student ID:	X18105149		
Programme:	MSc Financial Technology	**Year:**	2019
Module:	Research Project		
Supervisor:	Noel Cosgrave		
Submission Due Date:	12/08/2019		
Project Title:	Optimisation of credit rating process		
Word Count:	**10507**	**Page Count:**	31

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: 12/08/2019

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):	

Optimisation of Credit Rating Process

Amarnath Venkataramanan
MSc Research Project in Financial Technology
X18105149

Abstract

The credit rating process is considered to be one of the vital processes that defences the global economy. The majority of investments will be obtained based on these credit ratings which acts as the representation of the financial credibility of companies. As the current credit rating process found to be expensive, small and medium-sized enterprises(SMEs) which are considered to be the backbone of the global economy might find difficult to access the funds via investment for their development which in turn affects the global economy as well. This issue might be solved with the outcome of this research in terms of the optimised credit rating system with improved accuracy and continuous credit rating transition. Support Vector Machine(SVM) managed to achieve the highest accuracy of 92.0% whereas Random Forest(RF) and C5.0 decision tree also achieved greater accuracies with different formats of the dataset. With the help of dictionary-based sentiment analysis, this research proved that continuous credit rating transition system could track the changes in the financial status of the company which in turn helps to predict the crisis like bankruptcy and default in prior.

Keywords: Credit Rating Process, Support Vector Machine, Random Forest, C5.0 - decision tree

1 Introduction

When an individual requires money for personal use and approaches a bank or some lending companies with the required documents. A bank or lending companies analyse the information from the document to gauge the creditworthiness of the individual and create a report with a score that denotes the measure of creditworthiness of an individual. This report is called a credit report. Mostly, these credit rating process will be carried out by the third-party which are specialised in measuring the creditworthiness of an individuals and shares the credit report along with credit score to banks and lending companies. Then, based on the individual's credit report banks and lending companies will make a decision about lending money to the concerned person or not. Apart from the decision on lending, credit report also used to set the interest rate for the lending amount. Usually, the interest rate for the person with a low credit score will be high and interest rate will be low for the person with the high credit score (i.e. Interest rate is inversely proportional to the credit score). Not all banks and lending companies lend money for the person with the low credit score because of the high risk of default, only a few banks and lending companies lend the money to the person with the credit score with high-interest rates.

Similar to the lending process for an individual, companies and governments also have to prove their credibility to the banks or investors with their credit rating in order to access the funds for their development. Unlike credit scoring for an individual, the credit rating for companies will be more complex as they involve a huge amount of data. Credit ratings are considered as one of the most important measures that help to understand the creditworthiness of any organisations which in turn helps the investors to make decisions on the investments.

Such credit ratings are mostly provided by private parties called Credit Rating Agencies (CRAs). (White, 2018)

1.1 Motivation

In this credit rating sector, there are three big companies (S&P, Moody's and Fitch) holds the major part of the credit rating sector (>90%). As these big three CRAs hold the major portion of the sector so they consistently increase the price as skyrocket which makes in the near future only big companies could afford to be rated where small and medium-sized enterprises (SMEs) could not afford to be rated end up in the lack of investments. Singh et al. (2009) state that most of the fast-developing countries like China's economy depend on these SMEs if they continuously struggle to obtain investments to develop their business or sustain their business, most of the country's economy might be shaken. At the same time, CRAs are criticised for biased rating claimed by "2016 Annual Report on Nationally Recognized Statistical Rating Organizations," (2016) which also proven by the recent incident where S&P has been made to pay 58 million US dollar as a penalty. With the existing process, CRAs may fail to detect the economy collapse like the 2008 crisis. All these problems motivate this paper to address the issue and propose suitable solutions for the above-mentioned problems.

1.2 Context of research

This research addresses two major problems of CRAs namely 1) credit rating prediction with low accuracy and 2) delayed credit rating adjustments which can lead to disaster in most of the country's economy and proposes a solution by introducing the optimised credit rating process with high accuracy and continuity.

1.3 Early adopters (SMEs)

Also, the current credit rating process found to be expensive as they need to hire subject matter experts to rate the companies based on their credible values by assessing the financial parameters of the concerned companies (Hajek and Michalak, 2013). Because of the expensive current rating process, SMEs are skipping the rating process and fails to obtain a loan from a licenced bank. Even though SMEs are obtaining investments from secondary markets like P2P lending, they might fail to obtain enough investments due to the lack of credibility. So with this optimised credit rating process, SMEs may be advantaged by getting rated for their firms based on their credibility to repay the loan and with this rating, they could obtain loan/investments from primary as well as secondary capital markets.

1.4 Research question

Since 2008 financial crisis, CRAs have been criticised for their biased rating and there were no major changes have been introduced in their credit rating process which makes the existing credit rating process still to be vulnerable to repeat the disaster in the economy like 2008. Because of the above-mentioned issues, the current credit rating process has the opportunity to be improved significantly and this gives birth to the following research question,

"Can the optimisation of the credit rating process significantly improve the credit rating process by enabling improved accuracy and continuity in the rating adjustments".

1.5 Impacts and implications

a. **Impacts:** At the end of this research, the optimised credit rating process will be obtained which is believed to be adopted by SMEs and advantaged with this optimised credit rating process by proving their credibility to the backers and access more funds in the form of investments which in turn generates a healthy economy.

b. **Implications:** Since 2008 financial crisis CRAs credit rating process hasn't met any significant changes which still keeps it vulnerable to a disaster in the economy anytime in future. As I mentioned before SMEs are making a significant contribution to most of the countries' economy but if they consistently struggle to get access to funds for their development or sustain in the market there will not be an upgrade in the economy instead only downgrade will be faced in the future.

As SMEs are considered to be the backbone of most of the country's economy, it is important to help them to access required funds via investments in order to develop and safeguard them. Though there are many new funding opportunities have been created via FinTech, the trust issue stills exist between backers and SMEs similar to trust issue between traditional capital firms and SMEs. In order to obtain required funds from backers or banks, SMEs have to prove their credibility with the credit rating which is said to be an expensive process which may not be affordable for the SMEs. The optimised credit rating process will help SMEs to prove their credibility at an affordable price which in turn help them to obtain required funds from the backers without any hurdle which in turn enables free flow of the economy. This paper consists of the relevant literature review in the second section followed by research methodology, design specification, implementation, evaluation and finally conclusion and future work.

2 Related Work

This section discusses the existing literature works relevant to the research topic. This research can be broadly divided into two major divisions namely, 1) credit rating prediction with improved accuracy and 2) continuity in the credit rating transition. There are a substantial amount of academic research works are available on credit rating prediction using various machine learning algorithm, whereas an only limited number of literary works are available on the topic of continuous credit rating transition. Let us discuss in detail about the various research works available on these major divisions individually.

2.1 Achieving higher accuracy in the credit rating prediction

The credit rating is a more complex activity which involves entire financial information of the organisation and a subject matter expert hired by CRAs performs complex analysis with this information and rates the organization according to the outcome of the analysis. This entire process is considered to be more expensive and time-consuming but with the outcome of this research, this credit rating process is optimised in order to make less expensive as well as less time consuming without any compromise on the accuracy (Hajek and Michalak, 2013).

As previously said there are more studies available when it comes to credit rating prediction. There are two variants of studies available one that uses statistical methods and another variant of study using AI methods. Among different statistical methods ordered probit model (OPM) and ordered logistic regression (OLR) stands ahead (Hwang, 2013). It is said to be that OPM and OLR perform better because they consider the ordering of rating class.

There are numerous studies available for AI-based approach which learn the model from the data itself rather than making assumptions like traditional statistical methods (Huang et al., 2004). In many studies, credit rating prediction was performed using neural networks and it developed higher accuracy when compared to previous studies that use traditional statistical methods (Brennan and Brabazon, 2004). An extraordinary comparison among different neural networks like probabilistic neural network (PNN) and radial basis function neural network performance has been done in the study (Hájek, P., 2010) which has shown better results than other methods like logistic regression, multi discriminant analysis and multilayer perceptron. Other than neural network models, Huang et al. (2004) and (Lee, 2007) proved that support vector machines (SVM) also able to perform as good as neural networks on the credit rating prediction with high classification accuracy. Hájek and Olej (2011) showed that kernel-based approaches on the supervised learning methods didn't perform better than the semi-supervised learning methods. Hájek (2012) adopted classifiers based on the fuzzy logic like fuzzy decision trees (FDT), adaptive fuzzy rule-based systems (AFRB) and Wang Mendel algorithm for credit rating analysis.

There are also other studies that use unique AI models, Martens et al. (2010) used ant colony optimisation, Brabazon and O'Neill (2006) used evolutionary algorithms, Delahunty, A. (2004) used an artificial immune system method and Kim and Han (2001) used case-based reasoning method in their credit rating predictions. Apart from US dataset and credit ratings from three famous credit rating agencies (Moody's, Fitch and S&P), there are some studies that used different country's dataset and credit ratings of different credit rating agencies such as Lee (2007) which used Korean dataset to predict credit rating using the SVM method, Shin and Han (2001) used case-based reasoning model on the Korean dataset to predict credit rating and (Wijayatunga et al., 2006) used Bayesian networks on the Japanese dataset.

In addition to these studies, there are also furthermore recent studies available on the credit rating predictions such as a Yuan et al. (2018), Agrawal and Maheshwari (2019), Abdou et al. (2017) and Doumpos et al. (2019). Yuan et al. (2018) use a random forest method to predict the credit rating by using emotional features that are extracted from the social media websites in addition to the traditional financial variables. Yuan et al. (2018) state that updated information about any company can be obtained from social media websites and also states that social media website also contain pubic emotions on the company as well. The study Yuan et al. (2018) is proven that the accuracy of the credit rating prediction has been increased after the presence of the emotional variables along with the traditional variables. The study Agrawal and Maheshwari (2019) states the relationship of the probability of the default of companies and industry beta which is also called as the sensitivity of the financial variables of companies. The comparative study by Abdou et al. (2017) compares the accuracy of traditional statistical model and machine learning model on the prediction of the bank's credibility. The study Doumpos et al. (2019) uses utilities additives discriminates (UTADIS) method in order to predict the company's default situation.

Recently, Hsu et al. (2018), has performed credit rating prediction with the combination of support vector machine (SVM) and artificial bee colony (ABC) and proven that this hybrid model has generated higher accuracy than the earlier studies that have used the traditional statistical models to perform credit rating predictions.

The other study Xia et al. (2017) developed a credit rating model using extreme gradient boosting (XGBoost) and obtained higher accuracy after tuning the hyper-parameters using Bayesian hyper-parameter optimisation. The result of the study also states that Bayesian hyper-parameter optimisation also overperformed grid search, manual search and random search.

Following are the previous studies that have produced a higher accuracy in terms of credit rating prediction by overperforming other traditional statistical models. The study Shin and Han (2001) has performed a credit rating prediction and produced a higher accuracy of 75.5% with genetic algorithm based CBR (case-based reasoning model) by overperforming normal case-based reasoning model (CBR) and multiple discriminant analysis (MDA).

The study Kim and Han (2001) has performed a credit rating prediction and produced a higher accuracy of 69.1% with the hybrid model of combining learning vector quantization (LVQ) model and case-based reasoning model (CBR) which overperformed other models like multiple discriminant analysis (MDA) and case-based reasoning model (CBR) whereas the research literature Brennan and Brabazon (2004) has performed a credit rating prediction and produced the accuracy of 84.0% with the multiple layer perceptron (MLP).

The study Delahunty, A. (2004) has performed a credit rating prediction and produced the accuracy of 72.5% with artificial immune system(AIS) model whereas the research study Huang et al. (2004) has performed a credit rating prediction and produced a higher accuracy of 80.0% with the support vector machine (SVM) model which has overperformed multiple-layer perceptron (MLP) and the research literature Kim (2005) has performed a credit rating prediction and produced the accuracy of 83.8% with adaptive learning network (ALN).

The study Brabazon and O'Neill (2006) has performed a credit rating prediction and produced a higher accuracy of 85.2% using grammatical evolution (GE) model which overperformed other models like multiple discriminant analysis (MDA) and multiple-layer perceptron (MLP) whereas the research literature Cao et al. (2006) has performed a credit rating prediction and produced a higher accuracy of 84.6% using SVM model which overperformed other models like feed-forward neural network (FFNN) and logistic regression (LR).

The study Lee (2007) has performed a credit rating prediction and produced a higher accuracy of 67.2% using SVM that overperformed other models like feed-forward neural network (FFNN), MDA and CBR and the research study Hwang et al. (2009) has performed a credit rating prediction and produced the accuracy of 72.8% with ordinal logistic regression (OLR).

The study from the author Hájek, Hájek, P. (2011) has performed a credit rating prediction and produced a higher accuracy of 87.4% using SVM by overperforming other models like PNN and RBF also the same author has performed a credit rating prediction in the study Hájek, P. (2010) and produced a higher accuracy of 58.5% using the probabilistic neural network (PNN) by overperforming other models like SVM and RBF whereas the literature research study Hwang et al. (2010) has performed a credit rating prediction and produced the accuracy of 81.1% using ordered semiparametric probit model.

The study Kim and Ahn (2012) has performed a credit rating prediction and produced a higher accuracy of 68.0% using the hybrid model that consists of the combination of SVM and organisation process performance (OPP) models and this hybrid model has overperformed other models like normal SVM and FFNN whereas the literature research study Hájek (2012) has performed a credit rating prediction and produced the accuracy of 59.6% using adaptive fuzzy rule-based system (AFRBS),

The study Hájek and Olej (2011) have a performed credit rating prediction using harmonic Gaussian model (HGM) and global consistency model (GCM) which produced the accuracy of 85.8% with HGM that overperformed GCM (83.7%) model whereas the study Yeh et al. (2012) has performed a credit rating prediction and produced a higher accuracy of 93.4% using rough sets (RS) that overperformed other models like SVM and decision tree (DT) and the literature Yuan et al. (2018) has performed a credit rating prediction and produced a higher accuracy of 77.6% using the random forest (RF) model which overperformed the SVM model.

Literature	Used method	No. of observations	No. of variables	Max accuracy achieved
Shin and Han (2001)	GA-CBR	3886	168	75.5%
Kim and Han (2001)	LVQ + CBR	2971	329	69.1%
Brennan and Brabazon (2004)	MLP	600	8	84%
Delahunty, A. (2004)	AIS	791	8	72.5%
Huang et al. (2004)	SVM	265	12	80%
Kim (2005)	ALN	1080	26	83.8%
Brabazon and O’Neill (2006)	GE	791	8	85.2%
Cao et al. (2006)	SVM	237	17	84.6%
Lee (2007)	SVM	3017	297	67.2%
Hwang et al. (2009)	OLR	736	24	72.8%
Hájek, P. (2011)	SVM	852	6	87.4%
Hájek, P. (2010)	PNN	852	11	58.5%
Hwang et al. (2010)	OSPM	779	4	81.1%
Kim and Ahn (2012)	SVM + OPP	1295	14	68%
Hájek (2012)	AFRB	852	11	59.6%
Hájek and Olej (2011)	HGM	1021	6	85.8%
Yeh et al. (2012)	RS	2470	18	93.4%
Yuan et al. (2018)	RF	2621	5	77.6%

Table 1: Results of the previous studies

Table1 shows the results of the credit rating prediction performed on previous studies in terms of classification accuracy along with the details like methods adopted, a number of rating classes used, a number of variables and objects taken into account. Most of these studies used the credit ratings of Moody's and S&P. As the most of the studies use different datasets that consists of different companies, different timeframes, different geographical locations and different economic condition, the comparison among their results will not be considered as the fair comparison.

2.2 A continuous credit rating transition system

Credit rating agencies (CRAs) are more concerned about the volatility of credit rating which in turn cause the volatility in the economy as credit ratings are indirectly controlling the investments on the companies. So, to control the volatility on the credit ratings CRAs mostly changes the credit ratings of the companies twice a year. Sometimes CRAs failed to make adjustments in the credit rating of the company even they obtained new information about the organisation. This action of CRAs may cause failure to detect severe distress in the country's economy in prior. So, to address this issue, this paper proposes a new system called continuous credit rating transition system (CCRT).

The main source of corporate funds is obtained from corporate bonds whose price totally depends on the company's credit rating (Yin et al., 2018). The bond price can also be considered as the measure of the organisation's financial strength. Usually, the studies on corporate credit ratings are considered as one of the most attractive topics in the financial field (Richelson and Richelson. 2011). Credit rating transition is abandoned for a very long time by CRAs while they considered default risk as a credit risk. But, after the 2008 financial crisis, CRAs and traders understand the importance of the credit rating transition. The investors and traders also realised that financial distress can be avoided to a greater extent if CRAs would have performed credit rating transition at the right time. The study Bessembinder et al. (2018) accused the credit rating transition as a reason for the drastic change in the bond pricing.

Credit rating transition or credit rating migration or transition matrix is the process of altering the settle credit rate of an organisation (D Hadad et al., 2009). In terms of credit rating transition, most of the studies used a Markov chain model in various forms. The number of studies that used Markov chain model is (Das and Tufano (1996); Jarrow et al. (1997); Lando (2000); Thomas et al. (2002); Duffie and Singleton (1999); Hurd and Kuznetsov (2007)). Even D Hadad et al. (2009) assumed that CRAs are using Markov chain model for the credit rating transition and explained that the usage of Markov chain model is the reason for periodical credit transition like once a year or twice a year.

The study Krüger et al. (2005) discusses the time-series properties of credit rating transition matrix and found that in the particular time duration the credit ratings of high rated companies were adjusted very less number of times whereas the credit ratings of low rated companies were adjusted more often. The study Krüger et al. (2005) illustrated the same with the following diagram which has been generated with the study's dataset (balance-sheet data from Deutsche-Bundesbank pool).

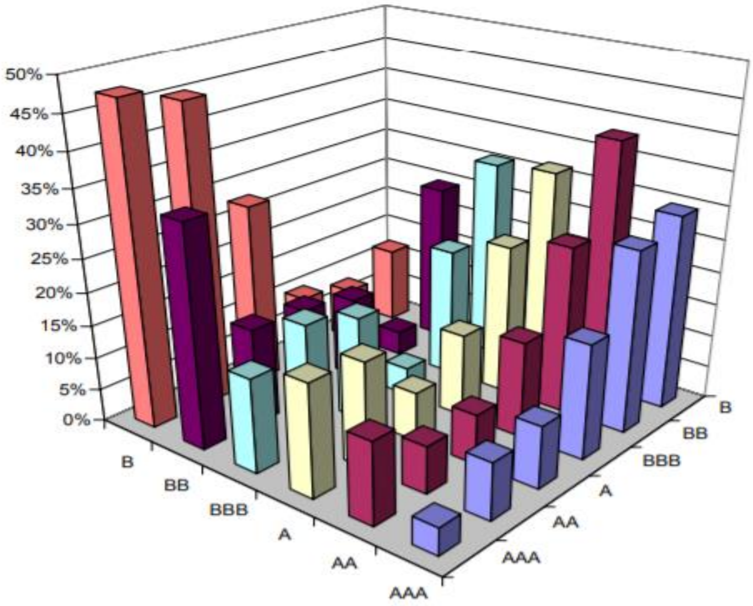


Figure 1: Coefficient of the transition of credit ratings from the study (Krüger et al., 2005)

From the diagram, it is clearly understandable that higher the rating lower the rate of transition and lower the rating higher the rate of transition. This diagram from the study Krüger et al. (2005) actually found this result from the data of the German organisation's credit rating between the period of 1988 and 2003. The study Krüger et al. (2005) also finds that the rating changes that occurred in a different period in the past have an impact on the present credit rating transition.

Finally, the studies Pfeuffer et al. (2019) and D'Amico et al. (2019) exposed the embedding problem in the Markov chain model which has been assumed as the model that is used by CRAs for the purpose of credit rating transition. All those previous studies stating the problems that exist in the Markov chain model and proposing a temporary fix for it but none of the studies have proposed an alternative solution but this paper does.

3 Research Methodology

As previously discussed, the aim of this research is to produce the optimised level of credit rating process in order to make sure that SMEs can have access to the funds that they require. But to achieve the optimised level of credit rating process this research project requires various methodologies and those methodologies will be discussed in detail in this section. As this research consists of two different sections namely, optimised credit rating prediction(OCRP) and continuous credit rating transition(CCRT). The methodologies used for those two sections will be discussed here.

3.1 Support vector machine (SVM) – For OCRP:

Support vector machine (SVM) is one of the familiar methods that has been introduced in the study Vapnik (2000) which was constructed using the principle of a computational theory called “Structural Risk Management”. The study Vapnik et al. (1996) states that SVM initially helped to solve the classification problems but later it is also being used for regression problems. SVM is being used in many financial applications in terms of default of person or organisation prediction, loan prediction, Stocks forecasting using time series, credit rating, prior detection of fraudulent claims of insurance, etc. (Viaene, Derrig, Baesens, & Dedene, 2002; Fan & Palaniswami, 2000; Huang et al., 2004; Gestel et al., 2001; Tay & Cao, 2001). These studies denoting that SVM has been performed well in terms of credit rating prediction by outperforming the various familiar machine learning algorithms like case-based reasoning model (CBR), multiple discriminant analysis (MDA) and artificial neural network (ANN).

SVM can implement non-linear class boundaries via non-linear mapping input vectors into a feature space of high dimension with the help of a linear model. The linear model developed in the new space able to characterise a non-linear decision border in the native space. An optimal separating hyperplane (OSH) is developed in the new space. Therefore SVM identifies the maximum margin hyperplane as a distinct variant of the linear model. The maximum parting between decision classes will be provided by the maximum margin hyperplane. Support vectors are the training samples that are positioned near to this maximum margin hyperplane which is appropriate for defining binary class boundaries (Hearst, Dumais, Osman, Platt, & Scholkopf, 1998; Gunn, 1998; Cristianini & Shawe-Taylor, 2000; Vapnik, 1998). As SVM is being more robust machine learning algorithm, it has been considered as the strong alternative to the traditional statistics models with its strengthen characteristics like easy to be used, more data-driven method, distribution-free method, etc.

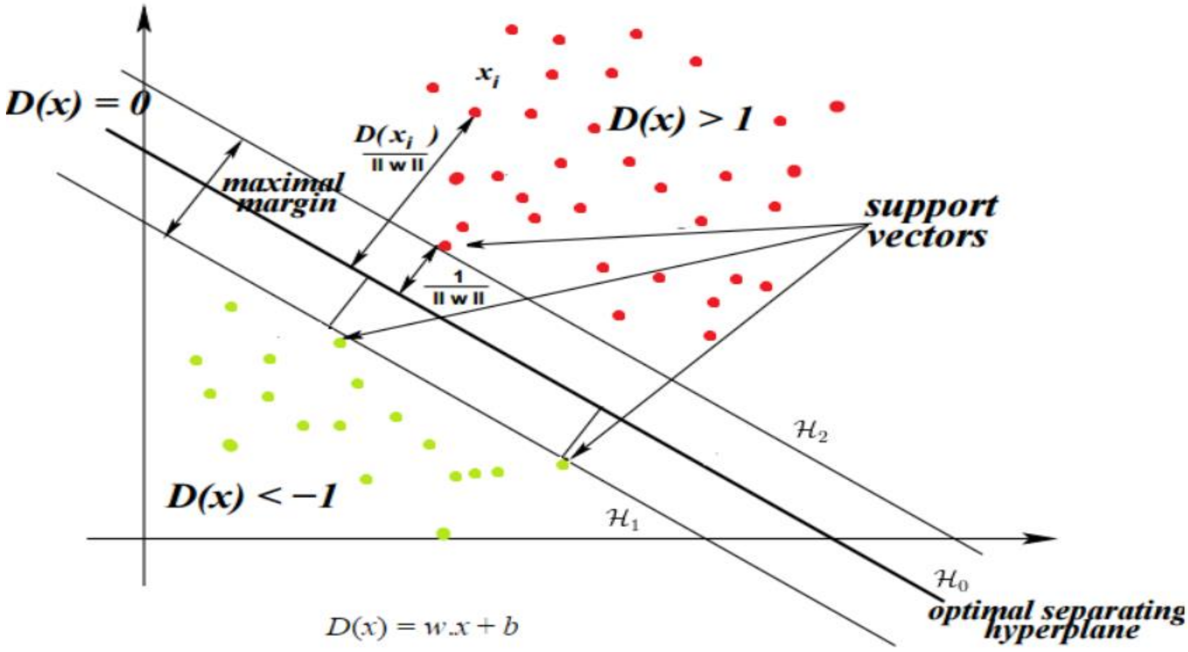


Figure 2: Different constraints of SVM[1]

[1] https://www.hackerearth.com/blog/developers/simple-tutorial-svm-parameter-tuning-python-r/

The normal SVM could solve only the binary classification problem but in order to solve multi-class classification problem either to construct multi-class classifier with many binary classifiers or making fundamental changes to the original form of SVM to consider all the available classes at the same time. As per the study Petropoulos et al. (2016), SVM has some serious disadvantages as other machine learning algorithms, SVM's computational complexity found to be higher which is n^3 (where n = number of training data points). So this makes SVM not suitable for bigger datasets where a number of training data points has to be increased in order to obtain a lower error.

3.2 Random Forest (RF) – For OCRP:

Random Forest (RF) is also one of the familiar machine learning algorithms which are now attracted by the academic researchers who are researching the credit rating analysis as per the study Petropoulos et al. (2016). Random forest model is a unique machine learning algorithm introduced in the literature study Breiman (2001) for the purpose of solving the classification problem. As per the study Breiman (1996) and Franklin (2005), random forest model makes use of an ensemble of classification trees in order to address the classification problems. The statistical framework of random forest considers the training set as $L = \{(X_1, Y_1)\ (X_2, Y_2), \ldots\ldots\ldots\ (X_n, Y_n)\}$ made up of "n" number of data points from a random vector (X, Υ). Vector is said to be $X = (X1, \ldots, Xp)$ holds the independent variables or predictors, say $X \in R^P$ and $Y \in \bar{Y}$ where $\bar{Y}$ is a response variable or class label. For the problems like a classification problem, mapping is a classifier t, $t = R^P \rightarrow \bar{Y}$ at the same time for regression problems, say that $\Upsilon = s(X) + \varepsilon$ with $[\varepsilon|X] = 0$ and "s" is called as regression function and more details on statistical framework of RF can be found in the study Franklin (2005). Random forest model is considered to be providing estimators of either Bayes classifier (minimising classification error, $P(\Upsilon \neq t(X))$), or regression function.

The principle of random forests is to merge more than one binary decision trees built with the help of many bootstrap samples generated from the learning sample L and also random forest will choose a subset of independent variables at each node in a random manner. When compared to the familiar classification and regression trees (CART) model, a random forest model has two differences. First is, the given number of input variables at each node is chosen in a random manner and also the calculation of the best split will occur only within this subset. other is, Second is, all the trees of the forest are considered to be maximal trees since there is no pruning step is performed.

Random forest algorithm posses a special characteristic which makes this algorithm a unique machine learning algorithm where random forest can rank the variables according to importance in terms of prediction and with this characteristic the accuracy of the prediction could be increased. The studies Breiman (2001) and Liaw and Wiener (2007) state that the random forest algorithm possesses many advantages when compared to other available statistical models. Among the various advantages of random forest algorithms, the major advantages of the RF discussed below.

1. Random forest algorithm is one of the few machine learning algorithms which can be applied to the categorical and continuous type of date.

2. There are only two extreme-parameters are possessed by the random forest algorithm. They are, a) number of variables from the random subset of each node, b) existing number of trees from the forest. As per the study Liaw and Wiener (2007), it is advisable, to begin these parameters with a default values even random forest algorithm is not dependent on these values
3. Random forest algorithm possesses an exclusive characteristic of ranking the variables based on their importance in terms of prediction of the response variable by analysing the predictive variable.
4. With the help of the exclusive characteristic of the random forest algorithm, it is possible to increase the accuracy on the prediction.

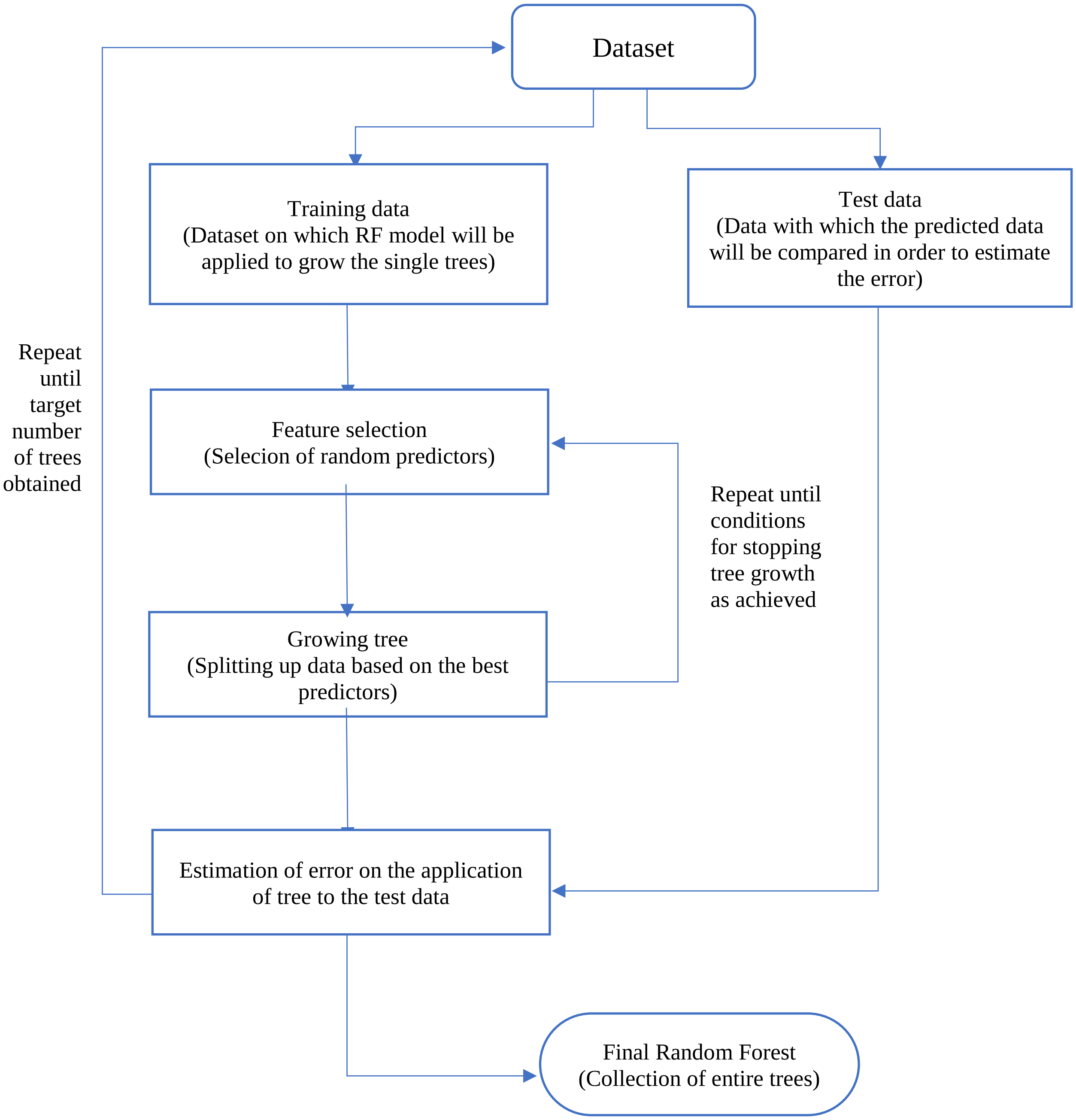

Dataset
Training data
(Dataset on which RF model will be applied to grow the single trees)
Test data
(Data with which the predicted data will be compared in order to estimate the error)
Repeat until target number of trees obtained
Feature selection
(Selecion of random predictors)
Repeat until conditions for stopping tree growth as achieved
Growing tree
(Splitting up data based on the best predictors)
Estimation of error on the application of tree to the test data
Final Random Forest
(Collection of entire trees)

Figure 3: Process flow of the random forest algorithm (Dalipi et al., 2016).

3.3 Bayesian hyperparameter optimisation – For OCRP

Most of the classification algorithm possesses parameters. The commonly used machine learning algorithms in the previous studies such as neural networks (NN), support vector machines (SVM), classification and regression tree (CART), random forest (RF), etc. are possessing many hyperparameters which have the capability to influence the accuracy in the greater manner. So, proper tuning of those hyperparameters (hyperparameter optimisation) leads to higher accuracy in terms of prediction. The study Bergstra et al. (2011) considering tuning as a "black art" as it majorly depends on practice, subjective judgements and trial-and-error method. Even though hyperparameter tuning for SVM and RF can be performed with the grid search, Bayesian optimisation overperformed grid search in terms of achieving higher accuracy in the prediction of credit rating in this study. Also, Bayesian optimisation achieved significant results in some of the optimisation problems and proven to be a better solution when it comes to hyperparameter tuning (Xia et al., 2017).

3.3.1 Hyperparameters and parameters

The ultimate aim of the machine learning algorithm is to identify a mapping y = G(x, ϴ), where x stands for input vector, y stands for output and G is the map that is parameterised by a vector ϴ. Apart from data, most of the machine learning algorithm itself possesses the parameters which cannot be learned straight away from the input and before model evaluation it has to be fixed. Such parameters are known as hyperparameters because they are found to be a higher level with respect to parameters obtained from data. Usually, parameters are used by the machine learning algorithms to fit the model to data well whereas hyperparameters primarily refine the model by controlling the complexity of the model (Xia et al., 2017).

3.3.2 Sequential model-based global optimisation (SMBO)

Bayesian hyperparameter optimisation consumes hyperparameters and relevant observations as inputs in order to discover the information on black-box (unknown) function. As per the study Hutter, Hoos, & Leyton-Brown (2011), sequential model-based global optimisation (SMBO) performs a model (S) building to map the settings of hyperparameter λ to function L and trails (H). The certain hyperparameter configuration can be estimated using loss function L. The inspection from the Bayesian hyperparameter optimisation is assisted by trials (H) and it also shows the relevant evaluation and the settings of the parameter. The SMBO iterates these steps in the loop until settings of global optimal hyperparameter with the minimum loss c is obtained.

3.4 Dataset used and data pre-processing – For OCRP

For this part of the study, to perform the credit rating prediction a dataset from the Kaggle website[2] has been used. This dataset consists of 1169 observation and 127 features. This dataset consists of financial information of the United States of America (USA) consumer good companies along with their credit ratings. Most of the variables in the dataset are holding numeric values except variables such as "Ticker" (stock name of the companies) and

[2] https://www.kaggle.com/peanutmochi/company-credit-rating-sp

"RATING" (credit rating of the companies). In this dataset, "RATING" variable is considering as a response variable all other variables are considering as predictive variables. Before proceeding with the model application dataset has to be preprocessed in order to make it fit for applying the model. From looking at the structure of the dataset using str() in R we could you see there are several missing values in various variables. The dataset usually contains missing in the four following forms.

1. Structurally missing data: Whenever data are missing for any logical reason (data which are not supposed to be present).
2. Missing completely at random (MCAR): Whenever data is missing without any pattern and if the data analyst assumes that data are missing completely at random(MCAR) then the missed data will be inputted based on the existing data (e.g. mean or median imputation).
3. Missing at random (MCR): Whenever data is missing without any pattern and if the data analyst assumes that data are missing at random then the missing data will be inputted after performing advanced imputation or any specific method designed for the missing data at random (e.g. multiple imputations).
4. Missing not at random: Whenever data is missing not at random then we cannot any available methods like imputation because any of them will provide incorrect value.

In the credit rating dataset from Kaggle, after the analysis, it is found to be that all the missing values are found to be structurally missing data (missing for the logical reason). For example, Coverage in the dataset stands for "Interest Coverage Ratio" and "Interest Coverage Ratio = EBIT / Interest Expense", where EBIT = Earnings before interest and taxes. The reason for the values missing from the "Coverage" variable is that Interest Expense = 0 (any value divided by zero will remain undefined). Re-checked with the calculation of the other values of the "Coverage" variable by performing the same calculation of EBIT / Interest Expense. As all the missing values in the dataset are found be structurally missing data all those missing are removed via R function na.omit(). After removing the missing values there are 125 observations are remaining in the dataset. The credit rating of the companies from the "RATING" variable is generated from S&P credit rating agency so the rating will be in the alphanumeric format such as "AAA", "A+", "BB-", etc. but model cannot be applied to the dataset with alphanumeric response variable so the values of the response variable "RATING" changed to numeric format in a manner, rating "AAA" to "A-" to "3" (low-risk companies and highly credible companies), "BBB+" to "B-" to "2" (medium risk and medium level of credible companies) and "CCC+" to "C-" to 3 (high risk and low credible companies). After the data pre-processing the processed data will be sent to the next stage of feature extraction which will be discussed in detail in the next section.

3.5 Model-based feature selection – For OCRP

In this research, model-based feature selection has been carried where the selection of features is carried out based on the importance rate of the features in terms of prediction to achieve higher accuracy. This importance rate is obtained after building the learning vector quantisation (LVQ) model on the dataset. But before training the LVQ model on the dataset, 10-fold cross-validation has been carried out to build the model on the dataset more precisely. Then, varImp() function in R has been used to obtain the importance of the feature which also plots the features based on their importance in terms of ROC curve.

3.6 Model building – For OCRP

After the selection of features based on the importance rate, the dataset has been subsetted with only features that are selected via model-based feature selection process. Then on the subsetted data, the models we have chosen as part of this project have been built. As part of this research three machine learning algorithms were chosen and all those models have been applied on the dataset.

3.6.1 SVM

First, the SVM model has been built on the dataset. Usually, SVM can be applied to the dataset with a binary classification label but our dataset consists of more than two classes in the response variable but "caret" package used to build SVM automatically converts SVM to fit with such dataset.

3.6.2 RF

The second model used in this research is random forest (RF), as RF can be applied to solve both classification and regression problems. RF has been used in this research to classify the companies based on their ratings. RF chooses the random subset of features to split a node in order to form a tree.

3.6.3 C5.0 decision tree

The third model used in this research is C5.0 decision tree, similar to RF, C5.0 can also be applied to solve both regression and classification problems and it has fitted well with the dataset and produced good accuracy. Unlike most of the decision tree models, C5.0 has a special characteristic of the pruning decision tree in order to avoid overfitting on data and it can also be optimized with the help of pre-pruning process which helps to stop decision tree to do work which is not required.

3.7 Hyperparameter tuning using Bayesian hyperparameter optimisation – For OCRP

The ultimate aim of this research is to optimise the credit rating process to achieve higher accuracy so this research adopted Bayesian hyperparameter optimisation in order to more accuracy. This optimisation via hyperparameter tuning makes this research work unique where none of the aforementioned studies is not using this method to obtain higher accuracy. Using the library "MlBayesOpt" in R, hyperparameter tuning can be performed on RF, SVM and XGBoost. So, we are performing hyperparameter tuning for only SVM and RF. After performing Bayesian hyperparameter tuning this research could obtain a higher accuracy of 92.0% when compared to the aforementioned studies on the credit rating prediction. The details of these results will be discussed in the implementation and evaluation section of this report.

3.8 Sentiment analysis – For CCRT

Sentiment analysis is the process of discovering the emotions of an author from their text computationally. Usually, the sentiment is constructed as a binary point (positive/negative), but it is possible to tune further to discover the exact emotion of a writer expressing while drafting the text like anger, sadness, joy, fear, etc. There are many applications are available for sentiment analysis such as estimating the survey response, identifying public response on movies or product from their reviews, etc. Even sentiment analysis contains some error, it cannot be so perfect in identifying the specific feeling of the author but when there are lots of text which is not possible for a human to read then it may create some value in identifying the overall response of the crowd. As the current period is being addressed as the internet age, people are mostly sharing their emotions on any matter in social networking websites like Twitter, Facebook, etc. The data from the Twitter platform are considered to be most attractive to the researchers who work on the financial domain. In the past, most of the opinions were extracted from the way of mouth of relatives, friend, neighbour, etc. But, after the world introduced to the internet the way of extracting opinions from the set of people has been completely changed. Most of the organisations are no more conducting a survey to obtain opinions instead they are obtaining opinions from the publicly available information from the social networking websites. Apart from financial applications, there are many other applications that use sentiment analysis like, prediction of election results, tracking employee feedback, etc. In this research, the sentiment() in R will be used to obtain the sentiment scoring for each tweet with which the credit rating of Apple company will be adjusted.

3.9 Dataset, it's pre-processing and feature selection– For CCRT

The dataset for this part of the research has been obtained from the website data.world[3] This dataset consists of different tweets about Apple company from the period of December 2014. This dataset consists of 12 variables and 3886 observations but among those 12 variables, only one variable "text" that consists of tweets will be considered for this part of the research. Most of the data pre-processing of this part of the research like removal of hyphens, valence shifters, etc. will be carried out via sentiment() function in R along with the estimation of sentiment score. Apart from that, this dataset consists of 82 missing values which fall under the structural missing values because those tweets are not at all relevant to Apple's company financial values. So, being the structural missing values those 82 missing values are removed from the dataset. None of the model used in this part of the research so we are skipping the model building section for this part of the research.

4 Design Specification

The three-phase design architecture has been adopted to implement this research. In the first phase collection of data in the raw format will be carried out. The second phase consists of five steps to implement the first part of this research project. The second phase of the project includes pre-processing of the US companies credit rating dataset, performing feature selection, model building, hyperparameter tuning using Bayesian optimisation and finally evaluation of results. The third phase this project include, pre-processing the tweets on Apple dataset, selection of variables based on logical reason, application of sentiment analysis on the data to extract the emotion of tweets and finally evaluation of results.

[3] https://data.world/crowdflower/apple-twitter-sentiment

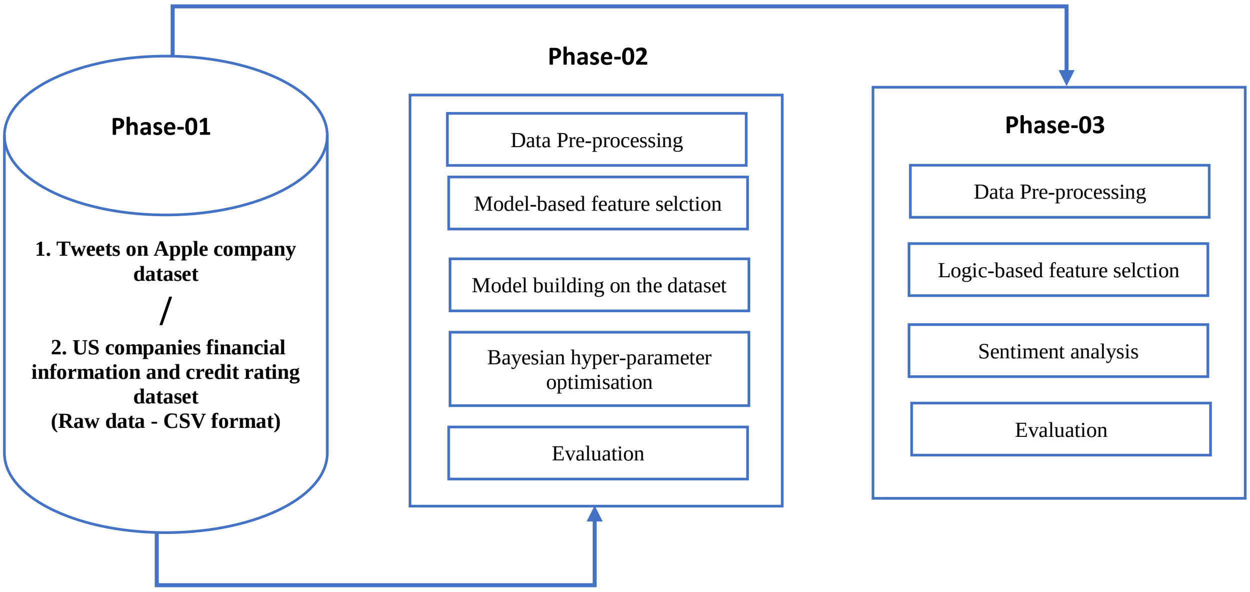

Phase-01
1. Tweets on Apple company dataset
/
2. US companies financial information and credit rating dataset
(Raw data - CSV format)
Phase-02
Data Pre-processing
Model-based feature selction
Model building on the dataset
Bayesian hyper-parameter optimisation
Evaluation
Phase-03
Data Pre-processing
Logic-based feature selction
Sentiment analysis
Evaluation

Figure 4: Design architecture of this research project

5 Implementation

5.1 Implementation of OCRP

5.1.1 The outcome of data preprocessing

Once the data pre-processing is completed, the outcome of the process will be the processed data on which the model can be built without any issues. In this part of the project, the adopted data have many missing values in different features. But all those missing values are found to be structural missing values(missing for the logical reason) so all the rows with the missing values can be eliminated from the dataset as they are considered to hold incomplete information. Once the missing values are eliminated, the variable "RATING" will be converted to the numerical form so that it will be convenient to build the model on it. After the data pre-processing data with no missing values and numeric credit ratings will be allowed to enter the next stage where the selection of features will be carried out.

5.1.2 The outcome of feature selection

In this research, the model-based feature selection is adopted to obtain improved accuracy. The top 20 features have been chosen based on the importance of ranking in terms of credit rating prediction. The following are the list of variable chosen on the basis of top 20 important features in terms of credit rating prediction. These are the features considered for this research. ROC curve variable importance, only 20 most important variables shown (out of 123 independent variables)

Feature name	Importance
COVRAGE	0.8038
TCE_RATIO	0.7976
RETURN_ON_ASSET	0.7953
NONOP_INCOME_LOSS	0.7794
IS_NET_INTEREST_EXPENSE	0.7757
SALES_ON_TOT_ASSET	0.7311
RETURN_ON_CAP	0.7311
RETURN_ON_INV_CAPITAL	0.7247
BS_OTHER_ASSETS_DEF_CHRG_OTHER	0.717
IS_INT_EXPENSE	0.7052
BS_MKT_SEC_OTHER_ST_INVEST	0.6852
DEBT_RATIO	0.6828
NET_DEBT_TO_SHRHLDR_EQTY	0.6828
BS_LT_BORROW	0.6808
NON_CUR_LIAB	0.6656
BS_TOT_NON_CUR_ASSET	0.6644
GROSS_MARGIN	0.6597
CF_INCR_CAP_STOCK	0.6595
IS_INC_TAX_EXP	0.6583
NET_DEBT	0.6525

Table 2: Top 20 important variables selected.

5.1.3 Model building

As part of this research, three models(SVM, RF, C5.0) were chosen to build on the dataset. The model building of this research has been divided into five divisions as follows,

1. Model built on the sub-setted data with no missing values at all.
2. Optimised model(obtained by tuning the hyperparameters) built on the sub-setted data with no missing values at all.
3. Model built on the sub-setted data with no missing values only in the 20 selected features.
4. Optimised model(obtained by tuning the hyperparameters) built on the sub-setted data with missing values except in the 20 selected features.
5. Optimised model built on the dataset with all available classes of the response variable.

In this section, the output of the above-mentioned four divisions will be discussed in detail.

5.1.3(a) The output of first division

	Accuracy	Kappa	Sensitivity	Specificity	P-value	Balanced Accuracy
SVM	0.898	0.7967	0.9565	0.8462	3.843e-08	0.9013
RF	0.8571	0.7168	0.9524	0.7857	1.786e-05	0.869
C5.0	0.7755	0.5527	0.8261	0.7308	0.0003551	0.7784

Table 3: Output of the first division of implementation

The output of SVM is an impressive 89.8% accuracy and it has been significantly well fitted on data with p-value less than 0.05 and RF alse fitted well on the dataset with p-value less than 0.05 and also produced the accuracy of 85.71% and it also been checked for the overfitting by fitting the model on the test data and it has produced the accuracy of 82.89% which is close to 85.71%. Whereas the C5.0 decision tree is found to underfitting(77.55% for training data and 80.26% for test data). So, from the output of this division, RF found to be overperformed SVM and C5.0. But, SVM found to be produced higher values of sensitivity, specificity and kappa value than the other two models.

5.1.3(b) The output of the second division

SVM:

```
Best Parameters Found:
Round = 8       degree_opt = 3.0000    coef0_opt = 3.5472     Value = 0.9200
```

RF:

```
Best Parameters Found:
Round = 9       mtry_opt = 6.4125      min_node_size = 1.0000 Value = 0.9184
```

C5.0:

```
winnow  trials  Accuracy   Kappa
FALSE   15      0.9038753  0.8076241
```

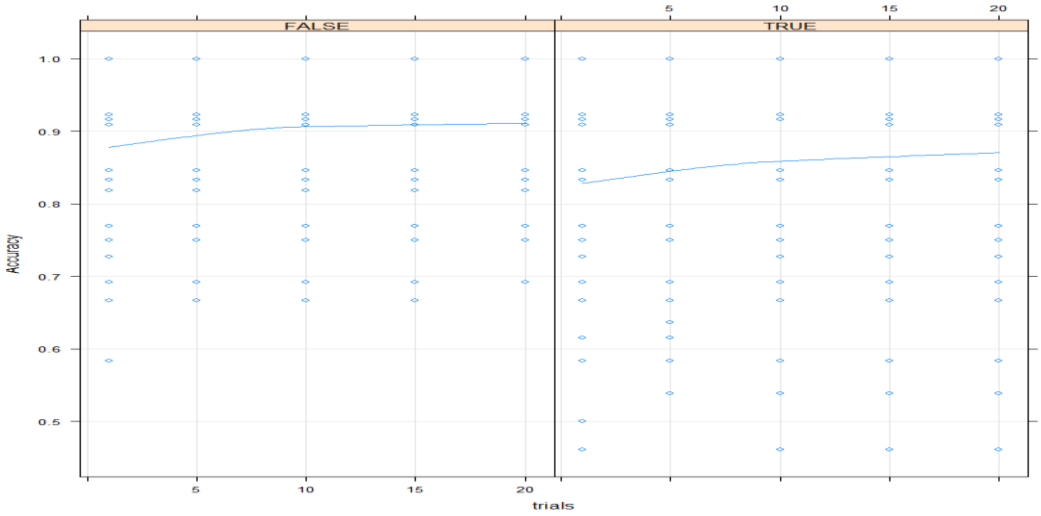


Figure 5: C5.0 decision tree's output after parameter tuning

From the output of the second division, it is found to be SVM has overperformed both RF and C5.0 in terms of accuracy after tuning their parameters. As per the goal of this research, this study has produced higher accuracy than the aforementioned studies on the credit rating prediction. From the above figure, it is clearly known that C5.0 has performed well without winnowing process.

5.1.3(c) The output of the third division

	Accuracy	Kappa	Sensitivity	Specificity	p-value	Balanced Accuracy
SVM	0.7639	0.5242	0.7738	0.7518	8.958e-15	0.7628
RF	0.823	0.6421	0.8161	0.8321	<2e-16	0.8241
C5.0	0.7902	0.5833	0.8623	0.7305	<2.2e-16	0.7964

Table 4: Output of the third division of implementation

The output of the third division clearly states that RF has overperformed the other two models with an accuracy of 82.3% accuracy and a good amount of sensitivity, specificity and kappa value. And also, all the models are found to be fitted well on the dataset with p-value less than 0.05.

5.1.3(d) The output of the fourth division

SVM:

```
Best Parameters Found:
Round = 1       degree_opt = 2.0000    coef0_opt = 9.6244     Value = 0.7911
```

RF:

```
Best Parameters Found:
Round = 2       mtry_opt = 5.6766      min_node_size = 1.0000 Value = 0.8721
```

C5.0:

```
  winnow  trials  Accuracy   Kappa
  FALSE   20      0.9089638  0.8164180
```

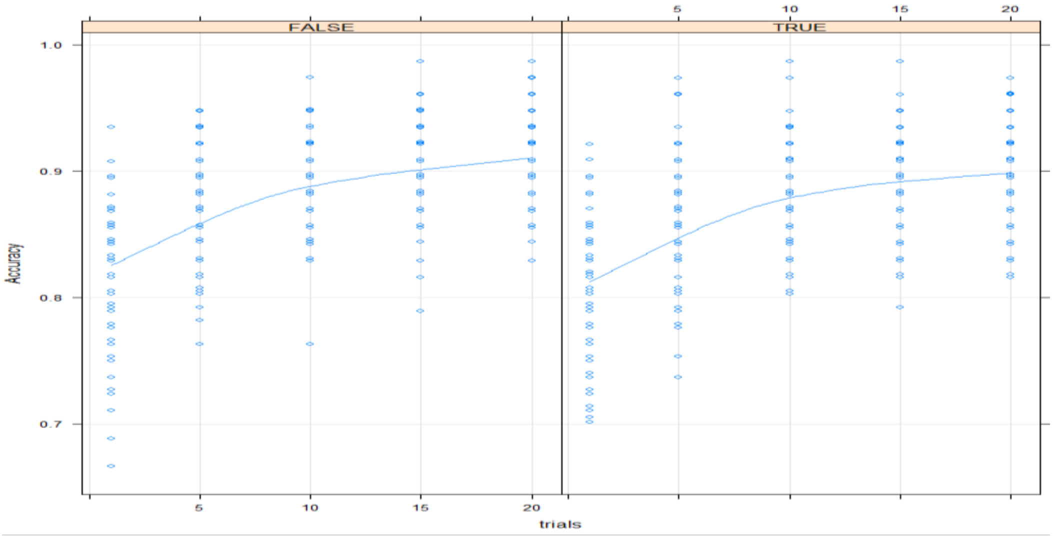


Figure 6: The output of parameter tuned C5.0 decision tree

The output of the fourth division states that the C5.0 decision tree overperformed the other two models with a maximum accuracy of 90.89% after the process of parameter tuning.

5.1.3(e) The output of the fifth division (C5.0 decision tree)

```
C5.0

766 samples
 20 predictor
 14 classes: 'A', 'A-', 'A+', 'AA', 'AA-', 'AA+', 'AAA', 'B+', 'BB', 'BB-', 'BB+', 'BBB', 'B
BB-', 'BBB+'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 686, 692, 689, 691, 688, 686, ...
Resampling results across tuning parameters:

  winnow  trials  Accuracy   Kappa
  FALSE    1      0.4957020  0.4300328
  FALSE    5      0.5997514  0.5457768
  FALSE   10      0.6442474  0.5963990
  FALSE   15      0.6566781  0.6106932
  FALSE   20      0.6650593  0.6200531
   TRUE    1      0.4918457  0.4257735
   TRUE    5      0.5987445  0.5448276
   TRUE   10      0.6382706  0.5896836
   TRUE   15      0.6526602  0.6059762
   TRUE   20      0.6620481  0.6166390

Tuning parameter 'model' was held constant at a value of tree
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were trials = 20, model = tree and winnow = FALSE.
```

When all three models built on the dataset with all 14 classes of the response variable, C5.0 managed to achieve higher accuracy of 66.5% after the tuning of parameters. It managed to achieve higher accuracy than all other optimised models and un-optimised models.

5.2 Implementation of CCRT

This part of the research involves the implementation of a continuous credit rating system (CCRT) with the help of sentiment analysis. This section involves data preprocessing and the application of sentiment analysis.

5.2.1 Data pre-processing and feature selection

As this research using the dictionary-based sentiment analysis, the outcome of this section is the subsetted data with no missing values and a feature with tweets about the apple company. The output data from this process will be suitable to apply sentiment analysis using sentiment() function of "sentimentr" package in R.

5.2.2 Implementation of dictionary-based sentiment analysis

Before applying sentiment analysis, this research explored data to understand the data in detail. The following charts are generated from R which can help to understand the data clearly in terms of analysis perspective.

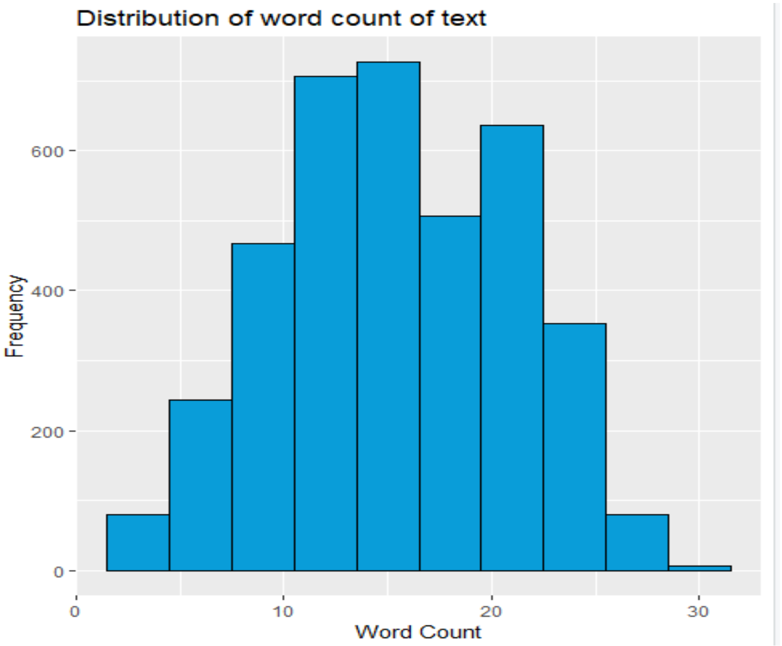


Figure 7: Distribution of word count of "text" variable which consists of tweets

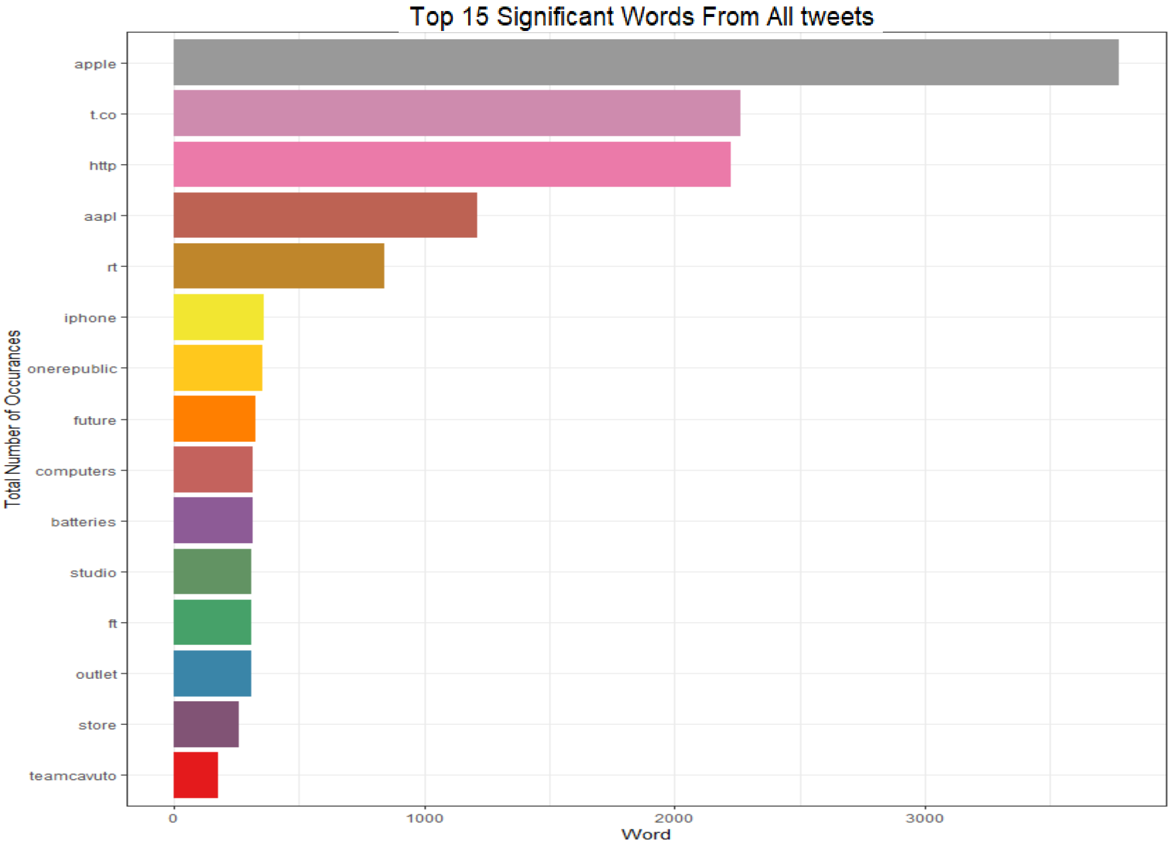


Figure 8: Top 15 significant words from entire tweets

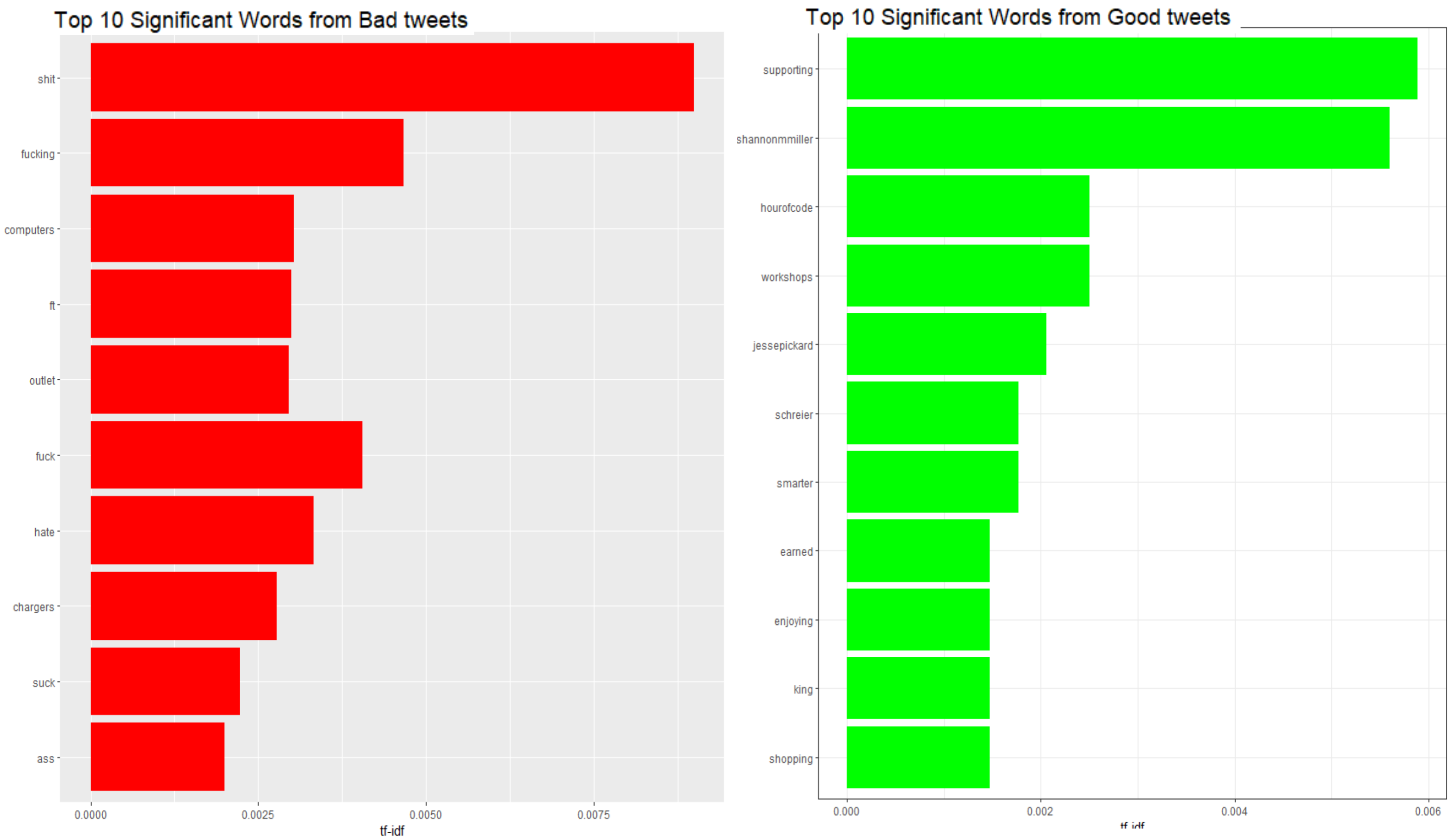
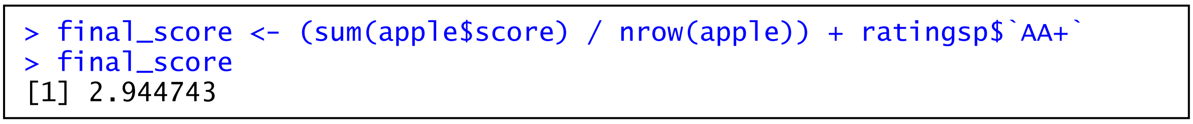


Figure 9: Top 10 significant words from the tweets rated as good and bad

The above-shown diagrams consist of the details on the exploration of the dataset such as the distribution of words and the graphs that consist of various good and bad words that has a significant impact on the analysis as y-axis and frequency of those words(tf-idf) in the tweets.

The output of the sentiment analysis is given below has been calculated based on the numerical representation of credit ratings provided in the table. So, as per the entries in the below table and the equation ((sentiment_score / no. of rows) + numerical credit rating), Apple's credit rating will be decreased(changed) as per the CCRT system but not changed as S&P system, if it continuously decreases or increases at one point Apple will reach very-high(3.0) credible company or medium credible company(NCR = 2.4). In this research new form of credit rating representation called as numerical credit ratings(NCR) is introduced which is derived from credit rating representation from three familiar CRAs.

The final output of CCRT:



```
> final_score <- (sum(apple$score) / nrow(apple)) + ratingsp$`AA+`
> final_score
[1] 2.944743
```

Moody’s	S&P	Fitch	NCR
Aaa	AAA	AAA	3
Aa1	AA+	AA+	2.95
Aa2	AA	AA	2.75
Aa3	AA	AA	2.5
A1	A+	A+	2.4
A2	A	A	2.2
A3	A-	A-	2.1
Baa1	BBB+	BBB+	2
Baa2	BBB	BBB	1.9
Baa3	BBB-	BBB-	1.75
Ba1	BB+	BB+	1.5
Ba2	BB	BB	1.4
Ba3	BB-	BB-	1.2
B1	B+	B+	1
B2	B	B	0.9
B3	B-	B-	0.75

Table 5: Numerical consideration of credit ratings (Amarnath Venkataramanan, 2019)

6 Evaluation

This section will discuss in detail about the final results of this research and show how it addresses the research question and objectives. The main objective of this research is to produce the optimised credit rating process with the improved accuracy and continuously changing characteristics which can be used by both big companies and small and medium-sized enterprises at an affordable price. The following results show that this research achieved the objectives successfully.

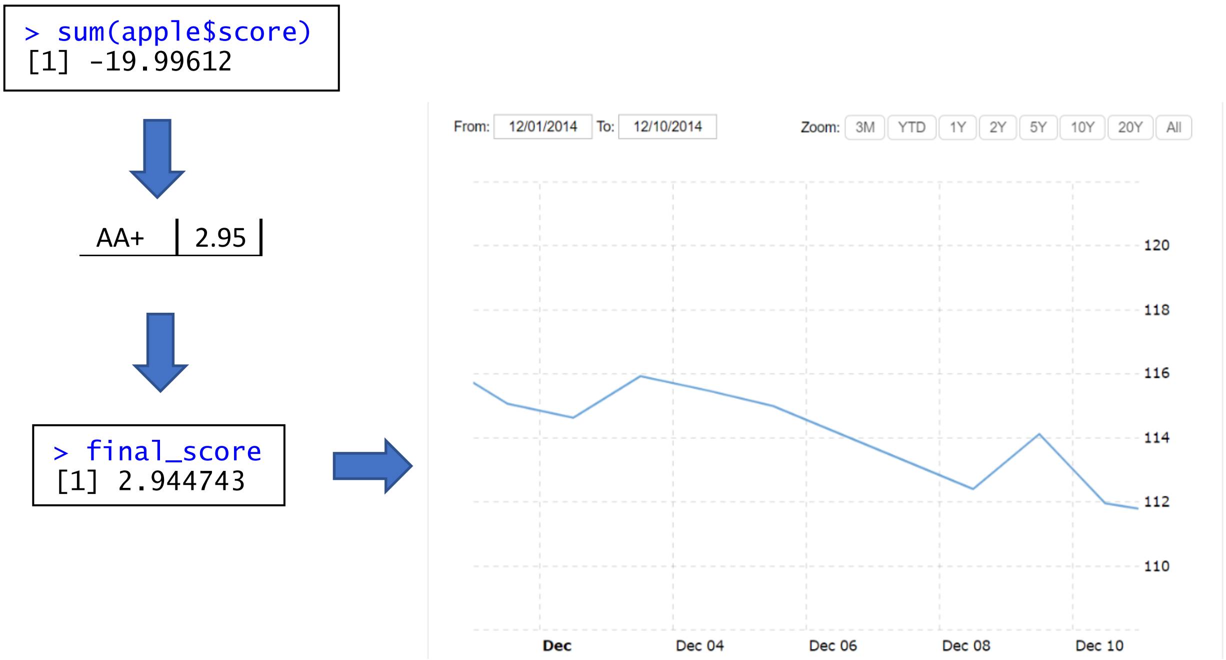
6.1 Optimised credit rating prediction (OCRP)

Implementation type	Model	No. of observations	No. of classes in Y	Accuracy achieved
Non-optimised	SVM	125	3	89.80%
Non-optimised	RF	125	3	85.71%
Non-optimised	C5.0	125	3	77.55%
Optimised	**SVM**	**125**	**3**	**92.00%**
Optimised	RF	125	3	91.84%
Optimised	C5.0	125	3	90.38%
Non-optimised	SVM	766	3	76.39%
Non-optimised	RF	766	3	82.30%
Non-optimised	C5.0	766	3	79.02%
Optimised	SVM	766	3	79.11%
Optimised	RF	766	3	87.21%
Optimised	C5.0	766	3	90.89%
Optimised	C5.0	766	14	66.50%

Table 6: Final output of OCRP

The ultimate aim of the investor is to know whether the company to be invested is highly credible, medium credible or low credible based on these the investor decides to invest. So, in this perspective, this research has reduced the 14 different classes of the response variable into 3 classes to achieve greater accuracy(92%) than most of the previous studies on credit rating prediction. With this result, one of the objectives of this research has been addressed successfully.

6.2 Continuous credit rating transition (CCRT)



```
> sum(apple$score)
[1] -19.99612
```

```
> final_score
[1] 2.944743
```

Figure 10: The final output of CCRT

The above diagram illustrates that CCRT of this research successfully tracked the downfall of Apple company financial value in terms of the stock price. The above diagram consists of Apple company's stock chart[4] from 01/12/2014 to 10/12/2014 as per our dataset which consists of tweets about apple company from 01/12/2014 to 10/12/2014. So, from this result, this research proves that CCRT will definitely add value to the optimized credit rating system.

6.3 Discussion

The objective of this research to produce an optimised credit rating system with increased accuracy and continuous transition in credit ratings. Among these two objectives, increased accuracy has been achieved via parameter tuning but the continuous transition in credit rating can be achieved by including readily available online automated sentiment analysis tool like "brand24" available for just $49 for a year subscription. By combining the output of automated sentiment analysis and the equation introduced in this research, the CCRT can be made possible.

After making this code to be open-sourced, this research could provide credit rating process in a significantly better way than the current state of credit rating process to SMEs and even big

[4] https://www.macrotrends.net/stocks/charts/AAPL/apple/stock-price-history

companies at an affordable price(for CCRT, $49) with significantly proven greater accuracy than most of the previous studies in terms of prediction with p-value of less than 0.05 for all the models used in this research and new characteristic of continuity in the credit rating transition

7 Conclusion and Future Work

The final outcome of this research is an affordable credit rating process from which SMEs can be benefited. As previously mentioned SMEs are considered as the backbone of most of the country's economy so the development of SMEs by providing them access to required funds in terms of investments will contribute to the country's economy. The outcome of this research considered to make this true. At the same time, CRAs are criticized for their intransparent credit rating process and this can be solved by the blockchain technology which will enable the transparency in the existing process and gains the confidence of investors that means more investments can be made possible which in turn generates nurtured economy.

8. Acknowledgements

I would like to thank and extend my sincere gratitude to the Supervisor, Noel Cosgrave for his continuous support on this research with his valuable insights, supervision, motivation and technical expertise. I would say that I am lucky enough to have such an intellectual supervisor. I would like to thank my parents who backed me to make this happen. I would also like to thank the National College of Ireland and its School of Computing for their resources for this research project.

References

2016 Annual Report on Nationally Recognized Statistical Rating Organizations, 2016. 40.

Abdou, H., Abd Allah, W., Mulkeen, J., Ntim, C.G., Wang, Y., 2017. Prediction of Financial Strength Ratings Using Machine Learning and Conventional Techniques (SSRN Scholarly Paper No. ID 3100986). Social Science Research Network, Rochester, NY.

Agrawal, K., Maheshwari, Y., 2019. Efficacy of industry factors for corporate default prediction. IIMB Management Review 31, 71–77. https://doi.org/10.1016/j.iimb.2018.08.007

Amarnath Venkataramanan, 2019. Contemporary topics in FinTech.

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Proceedings of advances in neural information processing systems (pp. 2546–2554).

Brabazon, A., O'Neill, M., 2006. Credit classification using grammatical evolution. Informatica (Ljubljana) 30, 325–335.

Breiman, L., 1996. Bagging predictors. Mach Learn 24, 123–140. https://doi.org/10.1007/BF00058655

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brennan, D., Brabazon, A., 2004. Corporate bond rating using neural networks. Presented at the Proceedings of the International Conference on Artificial Intelligence, IC-AI'04, pp. 161–167.

Cao, L., Guan, L.K., Jingqing, Z., 2006. Bond rating using support vector machine. Intelligent Data Analysis 10, 285–296. https://doi.org/10.3233/IDA-2006-10307

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines. Cambridge, England: Cambridge University Press.

D Hadad, M., Santoso, W., Santoso, B., Besar, D., Rulina, I., 2009. Rating migration matrices: empirical evidence in Indonesia.

D'Amico, G., Dharmaraja, S., Manca, R., Pasricha, P., 2019. A review of non-Markovian models for the dynamics of credit ratings. Reports on Economics and Finance 5, 15–33. https://doi.org/10.12988/ref.2019.81224

Das, Sanjiv, and Peter Tufano. 1996. Pricing credit-sensitive debt when interest rates, credit ratings, and credit spreads are stochastic. Journal of Financial Engineering 5: 161–98.

Delahunty, A., 2004. Artificial immune systems for the prediction of corporate failure and classification of corporate bond ratings (Doctoral dissertation, University College Dublin, Graduate School of Business).

Doumpos, M., Lemonakis, C., Niklis, D., Zopounidis, C., 2019. Analytical Techniques in the Assessment of Credit Risk: An Overview of Methodologies and Applications, EURO Advanced Tutorials on Operational Research. Springer International Publishing, Cham, pp. 77–98. https://doi.org/10.1007/978-3-319-99411-6_4

Duffie, Darrell, and Kenneth J. Singleton. 1999. Modelling Term Structures of Defaultable Bonds. The Review of Financial Studies 12: 687–720.
Fan, A., & Palaniswami, M. (2000). Selecting bankruptcy predictors using a support vector machine approach. In Proceedings of the international joint conference on neural networks.

Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27, 83–85. https://doi.org/10.1007/BF02985802

Gestel, T. V., Suykens, J. A. K., Baestaens, D.-E., Lambrechts, A., Lanckriet, G., Vandaele, B., et al. (2001). Financial time series prediction using least squares support vector machines within the evidence framework. 12(4), 809–821.

Gunn, S. R. (1998). Support vector machines for classification and regression. Technical Report, University of Southampton

Hájek, P., 2010. Probabilistic Neural Networks for Credit Rating Modelling. In IJCCI (ICFC-ICNC) (pp. 289-294).

Hájek, P., 2011. Municipal credit rating modelling by neural networks. Decision Support Systems, 51(1), pp.108-118.

Hájek, P., 2012. Credit rating analysis using adaptive fuzzy rule-based systems: an industry-specific approach. Cent Eur J Oper Res 20, 421–434. https://doi.org/10.1007/s10100-011-0229-0

Hajek, P., Michalak, K., 2013. Feature selection in corporate credit rating prediction. Knowledge-Based Systems 51, 72–84. https://doi.org/10.1016/j.knosys.2013.07.008

Hájek, P., Olej, V., 2011. Credit rating modelling by kernel-based approaches with supervised and semi-supervised learning. Neural Comput & Applic 20, 761–773. https://doi.org/10.1007/s00521-010-0495-0

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent System, 13(4), 18–28

Hsu, F.-J., Chen, M.-Y., Chen, Y.-C., 2018. The human-like intelligence with bio-inspired computing approach for credit ratings prediction. Neurocomputing, Advances in Human-like Intelligence towards Next-Generation Web 279, 11–18. https://doi.org/10.1016/j.neucom.2016.11.102

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., Wu, S., 2004. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision Support Systems, Data mining for financial decision making 37, 543–558. https://doi.org/10.1016/S0167-9236(03)00086-1

Hurd, Tom, and Alexey Kuznetsov. 2007. Affine Markov chain models of multi-firm credit migration. Journal of Credit Risk 3: 3–29.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In International conference on learning and intelligent optimization (pp. 507–523).

Hwang, R.-C., 2013. Forecasting credit ratings with the varying-coefficient model. Quantitative Finance 13, 1947–1965. https://doi.org/10.1080/14697688.2012.738935

Hwang, R.-C., Cheng, K.F., Lee, C.-F., 2009. On multiple-class prediction of issuer credit ratings. Applied Stochastic Models in Business and Industry 25, 535–550. https://doi.org/10.1002/asmb.735

Hwang, R.-C., Chung, H., Chu, C.K., 2010. Predicting issuer credit ratings using a semiparametric method. Journal of Empirical Finance 17, 120–137. https://doi.org/10.1016/j.jempfin.2009.07.007

Jarrow, Robert A., David Lando, and Stuart M. Turnbull. 1997. A Markov model for the term structure of credit risk spreads. Review of Financial Studies 10: 481–523.

Kim, K., Ahn, H., 2012. A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. Computers & Operations Research, Special Issue: Advances of Operations Research in Service Industry 39, 1800–1811. https://doi.org/10.1016/j.cor.2011.06.023

Kim, K.S., 2005. Predicting bond ratings using publicly available information. Expert Systems with Applications 29, 75–81. https://doi.org/10.1016/j.eswa.2005.01.007

Kim, K.-S., Han, I., 2001. The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases. Expert Systems with Applications 21, 147–156. https://doi.org/10.1016/S0957-4174(01)00036-7

Krüger, U., Stötzel, M., Trück, S., 2005. Time series properties of a rating system based on financial ratios, Discussion paper / Deutsche Bundesbank Series 2, Banking and financial studies. Deutsche Bundesbank, Frankfurt am Main.

Lando, David. 2000. Some elements of rating based credit risk modelling. In Advanced Fixed-Income Valuation Tools. New York: John Wiley & Sons, Inc., pp. 193–215.

Lee, Y.-C., 2007. Application of support vector machines to corporate credit rating prediction. Expert Systems with Applications 33, 67–74. https://doi.org/10.1016/j.eswa.2006.04.018

Liaw, A., Wiener, M.C., 2007. Classification and regression by randomForest.

Martens, D., Van Gestel, T., De Backer, M., Haesen, R., Vanthienen, J., Baesens, B., 2010. Credit rating prediction using Ant Colony Optimization. Journal of the Operational Research Society 61, 561–573. https://doi.org/10.1057/jors.2008.164

Petropoulos, A., Chatzis, S.P., Xanthopoulos, S., 2016. A novel corporate credit rating system based on Student's-t hidden Markov models. Expert Systems with Applications 53, 87–105. https://doi.org/10.1016/j.eswa.2016.01.015

Pfeuffer, M., Möstel, L., Fischer, M., 2019. An extended likelihood framework for modelling discretely observed credit rating transitions. Quantitative Finance 19, 93–104. https://doi.org/10.1080/14697688.2018.1465196

Shin, K., Han, I., 2001. A case-based approach using inductive indexing for corporate bond rating. Decision Support Systems, Decision-making and E-Commerce Systems 32, 41–52.

Singh, R.K., Garg, S.K., Deshmukh, S.G., 2009. The competitiveness of SMEs in a globalized economy: Observations from China and India. Management Research Review 33, 54–65. https://doi.org/10.1108/01409171011011562

Tay, F. E. H., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. Omega, 29, 309–317.

Thomas, Lyn, David Allen, and Nigel Morkel-Kingsbury. 2002. A hidden Markov chain model for the term structure of bond credit risk spreads. International Review of Financial Analysis 11: 311–29.

Vapnik, V. (1998). Statistical learning theory. New York: Springer

Vapnik, V., 2000. The Nature of Statistical Learning Theory, 2nd ed, Information Science and Statistics. Springer-Verlag, New York.

Vapnik, V., Golowich, S.E., Smola, A.J., 1996. Support Vector Method for Function Approximation, Regression Estimation and Signal Processing, in: NIPS.

Viaene, S., Derrig, R. A., Baesens, B., & Dedene, G. (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. Journal of Risk & Insurance, 69(3), 373–421.

White, L.J., 2018. The Credit Rating Agencies and Their Role in the Financial System (SSRN Scholarly Paper No. ID 3192475). Social Science Research Network, Rochester, NY.

Wijayatunga, P., Mase, S., Nakamura, M., 2006. Appraisal of Companies with Bayesian Networks. Int. J. Bus. Intell. Data Min. 1, 329–346. https://doi.org/10.1504/IJBIDM.2006.009138

Xia, Y., Liu, C., Li, Y., Liu, N., 2017. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications 78, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

Yeh, C.-C., Lin, F., Hsu, C.-Y., 2012. A hybrid KMV model, random forests and rough set theory approach for credit rating. Knowledge-Based Systems 33, 166–172. https://doi.org/10.1016/j.knosys.2012.04.004

Yin, H.-M., Liang, J., Wu, Y., 2018. On a New Corporate Bond Pricing Model with Potential Credit Rating Change and Stochastic Interest Rate. JRFM 11, 87. https://doi.org/10.3390/jrfm11040087

Yuan, H., Lau, R.Y.K., Wong, M.C.S., Li, C., 2018. Mining Emotions of the Public from Social Media for Enhancing Corporate Credit Rating, in: 2018 IEEE 15th International Conference on E-Business Engineering (ICEBE). Presented at the 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), pp. 25–30. https://doi.org/10.1109/ICEBE.2018.00015