JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Evaluating the performance of credit scoring models using data mining techniques. Case of the Ghanaian banking industry.

[1]Richmond Sarpong, [2]Kwame Simpe Ofori

[1]Instructional Technologist, [2]Lecturer,
[1]Centre for Online Learning and Teaching (COLT), [2]Schoolof Business and Social Sciences (BSS),
[1]Coventry University, Coventry, UK, [1]Ghana Communication Technology University, Accra, Ghana, [2]International University of Grand-Bassam, Côte d'Ivoire

Abstract: Banks and other financial institutions cannot do away with the outstanding credit balances with customers and the speed at which customers file for bankruptcy presents a risk to the credit industry sustainability. The demand for banks and financial institutions to break even in the credit industry and their quest to reduce the cost to the minimal in the future has attracted various research interests from industrial communities, research laboratories, and academics. To save losses in the future, it requires that a more accurate and robust model with a consistent predictive ability to evaluate credit either using existing models or through a proposal of new models. In this work, we develop a linear model using logistics model. We evaluated four algorithms based upon their generational history in terms of performance. Support vector machine, logistic regression, artificial neural networks and random forest as an alternative if all three models fail to predict accurately in the Ghanaian dataset and we benchmark the experiment with the German and Australian dataset. We developed a linear discriminant model for determining the probability of defaults. We adopted 10-fold cross-validation and data partition as two major data splitting technique to ensure efficiency of all the classifiers. In our findings, it was evident that Support vectors machine, in general, have higher accuracy in terms of classification accuracy compared to logistic regression and artificial neural networks using AUC, Type I and II errors and Risk charts. Random Forest was better than all three classifiers in the case of Ghanaian dataset but fail to give higher predictive accuracy in the German and Australian dataset. RF, SVM, and LR can be used as alternatives to each other from our study and proven by other related works. This study did not consider the default probability or creditworthiness of applicants who were rejected in the process of applying for loans. In other words, our sample dataset contains information of only applicants who were granted the loans and defaulted. This could lead to bias in our analysis even though it is acceptable in this field of study.

IndexTerms - Artificial Neural Networks, Support Vector Machine, Logistic Regression, Random Forest, Ensemble, Imbalance data set, Feature Selection, Area under the ROC Curve, Cost-sensitive, Classifier, Algorithms.

I. INTRODUCTION

The core of every country's economic power is how well vest their financial institutions are stable and how they operate. Every economy of a country that is characterized by financial crisis across board with dwindling of performance over time has a feature or is associated with substandard systems for bank use (Kabir et al., 2015). According to the December 2016 records of Bank of Ghana, they recognize 34 commercial banks and 4 representative banks in good standings and 10 FNGOS, 385 microfinance companies, 60 licensed money-lending companies are in good standing as at 31st July 2016 (BOG, 2016). Ghana bank lending rates have increased drastically within a space of eight (8) months from 38.3% in January to 42.84% in October 2016 (Trading Economics, 2016). By these indicators, financial institutions are the heartbeat of Ghana's economy as well as a driving force. Credit performance department within a financial institution and banks must have a good fit to assure profit and stability of the institution with the economy. Hence, this requires the need to investigate into the credit standings of the customer relating to his financial background and history, which is a key factor in granting credits as well as a determinant in reducing risk-associated credit (Hussain Ali Bekhet and Shorouq Fathi Kamel Eletter, 2012).

Banks' and financial institutions' main source of profit is from borrowers lending (Ala'Raj & Abbod, 2015). The credit industry is growing daily just as much as fraud cases arising each day, and this is causing financial institutions to incur losses within their operations. There are large numbers of credit or loan applicants daily applying for credit, which has led to rigorous screening of applicants' credit worthiness before loans or requested amount from the banks and other financial institutions are considered as worthy or bad. Credit scores are not for discrimination purposes only either to grant or reject applicant but it also looks at the

potential risks in giving out such loan or card or credit (Zhou et al., 2013). "Loan application evaluation would improve credit decision effectiveness and control loan office tasks, as well as save analysis time and cost" (Bekhet & Eletter, 2014:20).

1.1 Problem specification

The success of financial institutions is dependent on their ability to quantify and predict their risks accurately. Financial institutions are faced with credit risk issues in their line of operations of the business as daily activities (Dukiü et al., 2011). There is an increased number of financial institutions collapsing which is contributed by mass losses within the financial industry. This issue has resulted in strict bank regulations locally and internationally which lead to more accurate demand for models that can assess credit risk and also loan portfolios structuring within financial institutions (Hayashi, 2016).

Hence, these required for the development of models that can score loyal and potential customers of banks and financial institutions. According to Chen & Huang (2003), the development of credit scoring models is to segregate applicants of loans into either rejection group (bad credit) or acceptation group (good) including related features of the customers such as marital status, income, and age or criteria based on previous applicants data on accept and reject analysis. In credit scoring model there is the need for accuracy since potential and loyal customers are targets of banks and financial institutions.

Banks' and financial institutions' accuracy within the loans markets or credit markets through models of credit scoring have many prospects; a.) Financial institutions would have lower capital requirements if they can improve their credit scoring models accuracy, b.) Interest rate could also be lowered in the financial industry to customers around the globe if requirements for capital or expanding capital requirements could be reduced by lenders (Altman & Sabato, 2007).

According to Tsai & Wu (2008), a little advancement in the accuracy of credit ratings or scoring would significantly reduce risks associated with credit which would translate into an improvement of savings in the future. Ala'raj & Abbod (2015: 103) "Moreover, the better the accuracy of the classifier, the more impact it has on others." West (2000), cited by (Blanco et al. 2013: 362), states that just a slight improvement of 1% accuracy in the loan market would drastically result in the reduction of losses in banks and other financial institutions which would also save millions of dollars. West et al. (2005) suggested through their work that a 4% reduction of error annually could possibly save 1.2 billion dollars for the credit industry annually.

1.2 Objective of the study

This study seeks to attain the following feats:

a. To model credit default in the Ghanaian banking industry using logistic Regression.
b. To evaluate the performance of three different algorithms with a fourth algorithm as an alternative if all three algorithms or model fail to give the optimal solution.
c. To compare two data mining technique in terms of favorability to Ghanaian banking sector.

1.3 Scope of work / delimitations

This research attempted to solve issues and give recommendations within a period of three months. This research is not different from any other research and there were challenges within the time span to produce the report. Acquiring the needed dataset for analysing and comparing the systems selected in the methodology was a problem since most financial institutions found it very difficult releasing data deemed important, especially on matters bordering information on their customers even if it was for research purposes. Financial constraint was also a major challenge since data had to be garnered from all the ten (10) regions in order to provide a full representation.

Again, the reasons why customers default can only be based upon data availability at the bank, since there were no kind of survey conducted on the reasons or opinions of customers who default, this research would be biased towards customers.

1.4 Significance of study

a. The results of this research could be used by all commercial banks or financial institutions in developing a credit risk model in rationing credits to customers.
b. The models and algorithms used would determine the most significant variables, which are used as predicting customer's defaults risks.
c. The recommendations of this made from this work would add significantly to the body of knowledge in predicting risks associated with credit in Ghana, Africa and the world at large. It would also be a source of references for other researchers who are interested in credits risk and default predictions.

The study is organized in the following order: Sections 1 talks about introductions, problem specification and significance of the study whiles section 2 examines both the empirical and theoretical reviews of the study. Section three introduces the methodology used for this study. Experimental setups, data analysis and discussions is in section 4 and section 5 talks about limitations, findings and recommendations.

II. LITERATURE REVIEW

Over the years, there have been several approaches being adopted and developed in order to compute credit ratings or scores which includes Support Vector machine (Martens et. al, 2007, Van Gestel et al., 2003), Artificial Neural Networks (Information & Engineering, 2004), Logistic Regression (Patra et al., 2008), other algorithms genetic and ensemble classifiers (Abellán & Mantas, 2014). This chapter seeks to review the literature in details with relevance to the artificial neural networks, support vector machine, logistic regression, and evaluation for default risk accuracy. This chapter is categorized into three main parts, that are a theoretical framework, empirical framework, and justification for methodology adopted.

2.1 Empirical Literature Review

This section of the work deals with the definitions of our key words, problems, solutions and similar works under which this research is being conducted.

2.1.1 Credit Scoring

Credit scoring is a process of appraising risk of credits on applications of the loan. Through the usage of statistical processes and historical data, credit scoring helps to set apart the outcomes of several applicants based on their characteristics if they have any features of not complying to payments hence default. This statistical method is used as a "score" for banks to rate or rank applicants of loans or a borrows in terms of their risk (Mester, 1997). Credit scoring is the processes involved in classifying loan or credit applicants into two categories: those who would probably pay (good payers) and those who would default (bad payers) (Brown & Mues, 2012). According to Tang and Qiu (2012), Credit Scoring is described as classifying credit applicants into "good" and "bad" risk using formal statistical methods. "A credit score is a model-based estimate of the probability that a borrower will show some undesirable behavior in the future" (Lessmann et al., 2015:124). Credit scoring is a model used to predict new loan applicants into classification of good, the likeliness of applicants fulfilling his entire repayments obligation then accept request of the applicant or bad, highly probable of defaulting the payments agreements then reject the request of the applicant due to loss and costs incurred (Doori & Beyrouti, 2014).

According to Finlay (2006) gross income of applicants, years of employments, marital status, property status, etc. were criteria used for scoring loans of individual applicants. In order to develop credit scoring models then, one must understand the criteria selected and statistical functions used in computing the credit scores of an individual, organization or business entity. Also, in order to effectively model credit scoring then one must consider the objectives for credit scoring, the source of data sets and application of algorithms. The reason why we model credit score is to be able to assess customer worthiness to credit. Several types of research have been conducted in an attempt to define credit worthiness based upon the area of concern. Credit scoring models can be categorized into the following:

2.1.1.1 Profitability Scoring

Thomas (2000) defines profit scoring as lenders' ability to minimize risks associated with consumer defaults through maximization of profits yield on consumers lending. This is the situation where banks or financial institutions are more interested in calculating or forecasting the profit margins on the loans given to customers than considering the risk involved is giving out the loan (Crook, Edelman, & Thomas, 2006). According to Marron (2007), there are no risks in consumer competitive markets driven by profits, instead of considering risks they are rather actively engaged not to identify and divide but to be able to define the risk involved and estimate that risk as price. However, there is limited literature on profit scoring which is able to analyze and discuss results relating to factors of loan profitability. This is as results of absence of data needed to compute and calculate the profitability of customers (L. C. Thomas, Edelman, & Crook, 2002). Inaccurate capturing of customers' information or data such as properties owned by the customer, usage of credit facilities and mortgages, and several banks options available to customers all account to lack of adequate centralized data for analyzing of profit scoring models by researchers which have been a big challenge (Lessmann et al., 2015). According to Thomas (2000: 165), profit scoring has been faced with several challenges from data warehousing issues which are to enforce and give accurate accounts of all indicators assuring profits. For profit scoring models to be efficient then there must be fully integrated information systems in the industry to give all information on transactions of all customers to be accessed by the industry.

Profit scoring could be a promising future area that needs more attentions only if researchers and organizations would develop interest and invest in building a centralized data warehousing which would capture all transactions either within or outside the country in other to have more accurate information on customers' financial status in dealing with profit scoring. Researchers did not identify the relationship between the ability of customers being able to meet the repayments terms and the profit on the amount of loan given to borrowers. In addition, accurate information on all transactions of customers would clearly show those who would be able to meet their loan agreements including profits by their transactions with other banks while those with high banks obligations with other institutions.

2.1.1.2 Bankruptcy Scoring

This form of scoring is to predict the possibility of an individual or business entity declaring him/herself before the repayments terms agreed by both parties as bankrupt (Abellán & Mantas, 2014; Andrés, Lorca, Sánchez-lasheras, & Cos-juez, 2012; NANNI & LUMINI, 2009; Smaranda, 2014; Sun & Shenoy, 2007; C. F. Tsai & Wu, 2008).

2.1.1.3 Behavioural scoring

So much attention has been given to the behavioral scoring since the patterns of customers repayments could determine the survival of a bank it either sustain or collapses in the past. According to Kennedy et al. (2013), behavioral scoring is failing due to recent attention is given to application scoring due to arbitrarily fixed period performance to select from a wide range of data period which causes instability to predict accurately. In the works of Thomas et al. (2001); Thomas (2000); Marqués et al. (2012); Lee et al. (2002), Behavioural scoring is to be able to accurately predict customers odds in such that banks can foretell if the customer would default or not and to group them into good and bad.

According to Anderson (2007:310), "In general, most behavioral scoring is done using bespoke, internally-hosted scores, and if other scores are used, they are usually kept separate." According to Thomas et al. (2002), behavioral scoring and credit scoring is basic tool considered for financial institutions operations as means for managers decisions, to evaluate and minimize possible risks, and also to improve marginally the cash flow of the financial institution. In other to evaluate potential customers' worthiness of a bank or financial institutions credit scores must be considered whereas behavioral scoring has to do with monitoring and being able to forecast the behavior of repayments standings of borrowers.

2.1.1.4 Classification of credit default and overdue periods

According to Giesecke (2012:1) "Credit risk is the distribution of financial losses due to unexpected changes in the credit quality of a counterparty in a financial agreement". Niklis et al. (2014), define credit risk as the probability of a client's inability to meet the obligations of a debt (default). It is the possibility that counterparty or bank borrower would fail to meet the entire obligation enlisted in the agreement terms (Basel Committee on Banking Supervision, 1999). Zhang et al. (2016), also defines "Credit risk is the possibility of loss that the bank will suffer after offering loan to the borrowers". Spuchľáková et al., (2015) Also stated that credit risk or default risk is the unwillingness or inability on the part of the customer to fulfill his commitments stated in relation to

hedging, trading, lending, settlements and other financial transactions. Jorion (2003) It is the counterparty fails to meet its contractual agreements, which in effect would result in economic loss. From these definitions, none of the authors addressed reasons why the default or intentions of the borrower meeting that demands through revision of agreement with time. Hence, Credit default or risk can is the unexplainable inability of the borrower not being able to meet the initial contractual agreements between the lender or banks or other financial institutions and the borrower or customer, which would lead to financial gap over time to the lender or bank. In the work of Avery et al. (1996) stated that two occurrences happen at the same time once a borrower refuses to make payments agreed by the lender and the borrower thus, delinquency and defaults. In addition, he defines loan defaulter as borrower missing payments schedules and delinquency occurring when the borrower does not honor the payments schedules on a loan acquired.

Loan or credit can be classified based on day's overdue or late payments as either defaulted or paid off. Avery et al. (1996: 621) stated that usually, loans are due every month. They added that within the lending industry loans on delinquents are categorize into 30 days, 60 days, 90 days or 120 days or more as late depending on the amount of outstanding as overdue. According to Thomas et al. (2002:213) either full payments have been recovered or outstanding balance with a customer with a regular number of times payments has been made 30 days, 60days, or 90 days overdue. Sousa et al. (2016: 343) also stated that borrowers would be term as defaulters when they still have outstanding balance payments over 90days after 12 months would be considered as bad otherwise good. In addition, He proposed the third approach as an indeterminate group, which is 15days to 90days overdue, which he stated that people within this group are usually unclear to classify them into a group of bad or good debtor. Tong et al. (2012), classify defaulters into two groups based on the outstanding amount with the borrower. Thus, borrowers with arrears with a minimum of days of 90 days as overdue (Bad) and those within 90 days has non-defaulters (Good). West (2000) was the first to use (LDA) Linear Discriminant Analysis algorithms as a method of simple parametric statistics in credit scoring and benched marked his methods against other traditional methods such as decision trees, K nearest neighbor, logistic regression, and kernel density estimation. In addition, he explains that complex algorithms and models are being developed to give accuracy that is more precise due to the deficiencies of LDA. The deficiencies of LDA has received more criticism because of its poor classification of good and bad credit classes' covariance matrices, which are not possible to be equal based on the nature of data category available.

More attention has been drawn to data mining techniques in credit scoring such as; Classification tree or Decision tree, Neural Networks (Kiruthika & Dilsha 2015; Wang et al. 2008; C. F. Tsai & Wu 2008; Qin et al. 2010; Basheer & Hajmeer 2000; West et al. 2005; West 2000; Schmidhuber 2015), k-Nearest Neighbour, and Logistic Regression, Bayesian Network (Hsieh & Hung, 2010; Sun & Shenoy, 2007; Zhuang, Xu, & Tang, 2015), and Naïve Bayes (Antonakis & Sfakianakis, 2009; Baesens et al., 2003; Y. Jiang & Wu, 2009; Vedala & Kumar, 2012), Support vector machine (SVM) (Crook et al., 2006; Harris, 2013, 2015; Van Gestel et al., 2003), Genetic Programming (Aliehyaei & Khan, 2014; Ri et al., 2008; D. Z. D. Zhang, Hifi, Chen, & Ye, 2008). However, there more complex or hybrid models used in credit scoring today problem today. Even though majority of the researchers concentrate on Australian and German data sets. Credit scoring datasets sources are available from several countries. Most used datasets source is UCI Machine Learning Repository (http://mlearn.ics.uci.edu/MLRepository.html or http://archive.ics.uci.edu/ml/) which has German credit of 20 attributes on 1000 observations which are group into class attributes of 300 (Bad) and 700 (Good) observations. This attributes consist of years of employment, marital status, properties, job, age, gender, number of dependencies etc. In the case of Australian dataset contains 690 observations categorized into 383 (Bad) observation and 307 (Good) observations. Brown & Mues (2012: 3447) used data from Benelux region (Belgium, Netherlands, and Luxembourg) which is also available on (http://kdd.ics.uci.edu/). Ince & Aktan (2009:236) used data set on credit cards from Turkish bank with nine attributes on 1260 observations. In the works of He et al. (2010: 833) also used that from the United Kingdom and the United States. Abdou et al. (2008: 1281)Also used data from personal loans which was acquired from commercial banks in Egypt containing good loans of 433 and bad loans of 148.

2.1.2 Credit scoring problem

Despite vast research conducted to solve predicting accuracy and to have a more efficient scoring model and some criticism raise of some approaches. We cannot look at credit scoring in isolation without discussing some of the problems in credit scoring. According to Akko (2012: 176), he stated that it always good to predict accurately a bad credit applicant than a good applicant due to misclassification issues. Doumpos et al. (2015) Even though the results from their work was positive and restricted all misclassification into two notches. He recorded 94% errors in only misclassification of one-notch. According to Luo et al. (2016:5) recorded very positive results for accuracy with SVM 87.4%, MLR 77.31%, and MLP 87.75%. They added that there were average errors present in their work, which was as a result of ten independent datasets partitions they used in the cross-validation methodology. According to Jiang & Wu (2009), the issues of selecting a technique to use for classification in credit scoring have become a problem of challenge and very difficult. In the work of Jiang & Wu (2009) recorded 24% has an overall rate of error which was higher than 20.1% for C4.5. He adopted Simulated Annealing Algorithm (SAA) to minimize discrimination issue of misidentifying bad clients has good. However, statistics depicted that conditions under this are 5-20 times highly probably of misclassifying a "bad" applicant as "good" applicant. In this work, When Beta and gamma is shifted towards 0 or greater than two (2), SAA identifies all the credits applicants as either good or bad due to the discrimination limitation function imposed. In addition, He also highlighted that in building model of credit scoring then one has to solve the problem of variable selection and the type of model to apply. Even though there are several methods and solutions, accurate and flexible techniques are limited.

Also, Han et al. (2013:859) stated that Support vector machine which is to help in the classification problem cannot work on numerical variables except is normalized. In concluding their experiments, they stated that no model could be superior to other models in any way. Experience from Siddiqi (2006:11) depicts that development of scorecards or credit scoring models cannot be in isolation else the results would be undesirable with issues. Issues such as the addition of variables that are no longer collected or operationally difficult to collect and using strategies that would result in unimplementable models or surprise in your findings. Any significant changes in the procedures methods for credit scoring would demand further analysis. Variables with large ranges of scores present challenges since a slight shift in the distribution population would result in a significant drift in the scores (Siddiqi, 2006). Anderson (2007:250) raised some disadvantages with credit scoring software: inflexibility: rigidity in the use and limitation of options; Opacity: Resulting from some of these software or models are not easily understandable; Cost: Models or software can be very expensive in the case of customization and maintenance of the application to serve its purpose intended.

Roman & Stefano (2016:97) also identified the listed issues that must be dealt with using the current state of the art technology in the world to make credit scoring effective; i.) Delay Data: Increase demands and speed at which financial transactions across the globe with its complex nature are being conducted demands effective and efficient data storing systems to help access relevant data within the shortest possible time when needed. Unwanted bureaucratic delays from an official data source or public administration contribute to the ineffectiveness of credit scoring and unrepresentative of the real world. ii.) Trust issues of proprietary of data sharing: People have little or no understanding of the benefits of sharing data for credit scoring as well as an unwillingness to lose full control over their credit scoring data. Algorithm opacity for Credit scoring poses a challenge for individuals and companies to release their data except forced by law. iii.) Lack of Data: availability of data has become a big challenge in credit scoring model especially in small and micro-enterprises. Enough official economic data are readily not available and most of the best results or data available are unofficial or used based on statistical dimension approximated.

Wendel & Harvey (2003:4) identified that credit scoring would work best in an environment where there is sufficient information to demonstrate the performance of a borrower within a multi-year period and sufficient availability of credit information which also demonstrates the performance of similar borrower over a multi-year period. In addition, they highlighted issues affecting credit scoring reliability which are; (1) unfair treatment by minority, (2) reports from the credit bureau which is used for credit scoring are most at times inaccurate, (3) biases of customer data due to the use of limited period of time usually two years, (4) Special consultation of companies involve manipulating true finances and scores of consumers. Kellison & Brockett (2003:5) identified credit scoring is associated with a common criticism in the insurance industry in the use of underwriting decisions to discriminate against minority application or/and low-income earners which in effect results to "Red lining".

2.1.3 Benefits of credit scoring

According to Ghana Borrowers and Lenders Act (2008: 68), a person cannot be denied access to lending or borrowing by the lender based on the grounds of political affiliation, sex, race, ethnic, or religion. In addition, Equal Credit Opportunity Act (ECOA) (2013:1-2) of USA states that, the purpose of the statutes is to help banks and other lending institutions to provide credit and give equal opportunity to all customers who are either credit-worthy or have the mental ability to contract without discriminating based on marital status, religion, color, race, nationality, sex or age. From both experience and what is binding in Ghana and USA legal documents fully demonstrate that any lender who discriminates or denies access based on the above-stated characteristics shall face legal charges due to unfair opportunity. The above characteristics of borrowers must also not be included or used in credit scoring systems and other methods for credit scoring model that discriminate based on sex, religion, race, national origin, color, age or marital status. However, the recommendation was made that a lender can deny a prospective applicants access to credit on the basis of other information requested in the application documents used as a reasonable commercial basis which is uniform with the lender's common assessment for risk and practices of underwriting.

Despite the above characteristics eliminated in credit scoring analysis, its systems or models used other variables to determine the creditworthiness of existing customers and other prospective applicants. These variables make credit scoring models or systems a powerful tool to use and its benefits derived from scoring credit are enormous. Fensterstock (2005) stated that American businesses had been recorded as high users of credit scoring model for risk evaluation, which can be accounted for numerous factors due to the usage of technology. Fensterstock added that credit scoring system or model is more effective and efficient compared to personnel in a credit department to evaluate potential customers. It also helps save money.

Fensterstock (2003) mention that credit scoring development provides an objective decision, consistency with accuracy in providing results, ability to identify potential customers with high value, and finally able to lower risks at increasing credit limits. According to Jiang & Yuan (2007), Credit scoring models helps in accurately predicting bad or good debtors effectively. Guo et al. (2016: 424) “It has the ability to quantify the credit risk of each individual loan, instead of categorizing the loans into a small number of risk groups”. Credit scoring model is developed to decide whether to grant credit or not (X. Y. Liu, Fu, & Lin, 2010). Jiang & Wu (2009) in the industry of credit scoring is to aid save cost and help in making an efficient decision in credit scoring decision. This has led to a high growth of research in this area.

According to Wei & Mingshu (2013) using credit scoring model such as clustering ensembler enhances the prediction accuracy compared to the single model. They added that it saves time when compared to traditional static methods. Hence, credit scoring can only be achieved at a specific time. In addition, credit scoring model helps boost the relationship between management and customer through evidence as well as helping managers make decisions concerning default risk in time. Fayyad et al. (1996); Chen et al. (1996); Lee et al. (2006) identified that credit scoring and its classification problem have contributed greatly to business roles in the credit industry through financial forecasting, decision support, marketing strategies, process control, fraud detections and in other areas. Rimmer (2005:56-57) Reasons for the high demand of credit scoring approach in the late 1980s and early 1990s in the high street banks in the UK was due to lower costs of operation, availability of the system and central control, fast processing of data, decisions generated were unbiased. These actually led to high demand for automated decision systems to be preferred by lenders in scoring situations such as commercial loan facilities, mortgages that were classified as high-value loan portfolios. Through the development of credit scoring, mainstream credit facilities were made available to unreachable consumers with lower risk and ability to match the right products mix customers demanded. In addition, Jacobson & Roszbach (2003) stated that a lender ability to rank possible defaulters through observations is impossible. Credit scoring models are to equip the lender in his deficiency to rank all possible customers based on their risk levels. It has helped greatly in resource allocation to consumers, from challenges of allocations to the best equilibrium ever compared to the previous.

Wendel & Harvey (2003) stated that dependence on credit scoring through experiences and technology would help leverage the cost incurred on individuals employed for making credit granting decisions. In support, Diana (2005) discussed that reliance on technological credit scoring models would foretell the future analysis of credit risks. Automated credit decision systems are not to replace credit officers or managers, but to assist in the decision-making process with speed and accuracy of either disqualifying or approving applicants whose transactions falls below or above certain credit score indicators.

Leonard (1995:81-82) underscored that the development of credit scoring has helped to save time to do other reviews which would be necessary by a credit manager. He stated that the historical time of nine days for scoring through a critical analysis and decision of a credit manager was reduced to three days after using a credit scoring systems. He added that accuracy effectiveness which used to be 85% have improved to 92% which was very significant with fewer errors in the history of credit scoring. Banasiak & Kiely (2000) stated that the ultimate benefit of developing a customized credit score with cautious rules of decisions, automating it, with

validated statistics, and cheaper is due to the fact that results based on decision processing are faster and better as well as high standards of decisions.

Avery et al. (2000) identified the use of credit scoring for predicting future performance of loans for small business credit, mortgages or consumer applications. Key benefits derived from credit scoring are lowering cost of origination of loans, less subjectivity of underwriting, faster and consistency in decisions of underwriting with an accuracy of assessing risks. According to Park (2004:47), he said credit scoring model for risk assessment helps for accurate, quick and consistent decisions. It is reliable and gives unbiased ratings and not distract over time and same result with other reviewers. In addition, Park (2004:74) stated that the risk score has several uses such as being able to determine whether to approve and decline applicants who do not qualify for credit, to set limits on credits accessed by applicants, defining terms and conditions as well as effective management of portfolio for risk assessment distribution. Profit scoring model development with best models to predict consumer behavior would make the good profit from the industry and would be able to attract and select profitable customers to their business (Thomas, 2000). Due to the nature of the subprime lending sector, it has given growth to banks to depend more on management risk technology to reduce human interactions else there would huge loss. Development of credit scoring has contributed greatly to the subprime lending sector where there was inadequate information about prospective which could have resulted in inability rate their credit worthiness due to either issue of missing data in their credit history, impairment of credit or income validation issues (Quittner, 2003). Saunders & Cornett (2007) one major benefit of credit scoring is that lenders of credit are assured of making accurate predictions on the performance of borrower's loans without needing to use more resources.

Kellison & Brockett (2003:5), automobile insurance companies are facing losses due to the poor credit history of applicants. Insurance analysts' can use credit scoring to predict accurately loss costs per individual, to distinguish between the insured classes, and effectively price their policies with better-calculated risks that returned to the insurer from the policyholder. This would in effect solve the problem of huge cross-subsidies from all classes and would ensure equitable rates among the different groups as policyholders.

2.1.4 Solutions to classification problems in credit scoring

There are two major approaches to solving credit scoring problems even though there are still challenges and difficulties in using each of these approaches successfully.

2.1.4.1 Data Solutions

The rapid growth of data availability in large quantum with it complexities and networked systems being stored daily from several areas such as finance, security, internet, and surveillance has necessitated for improvement on the basic understanding of knowledge discovered in analysis most of this data to support decisions process. Data engineering methods and existing knowledge discovery have achieved great success in the application of this knowledge and techniques on several real world data by learning and understanding the problem of imbalanced data set cannot be under estimated (H. He & Garcia, 2009:1263). Class imbalance usually occurs, in the case of classification problem where there is less instance of some class than the others do. In a situation such as this, standardized classifiers are overpowered or flooded by the majority classes and tend to ignore the fewer classes (Chawla et al. 2004). In the context of credit scoring according to data set imbalanced usually occurs as the number of observations of non-defaulter being greater than the number of defaulters (Brown & Mues, 2012). As soon as the distribution class is too skewed it causes alarm for class imbalance problem (Maldonado, Weber, & Famili, 2014). The problem of class imbalance is very vital issues since it presence is a significant constraint in the attainable performance by the standard learning methods which assume the distribution of balanced class (Huang et al., 2006; Yao, 2009). These problems run across several areas such as Sentiment analysis, fraud detection, network intrusion detection, and churn prediction are all domains where data imbalance exist (Ofek, Rokach, Stern, & Shabtai, 2017). In support, Weiss (2004) highlighted that the problem of binomial classification is viewed as credit scoring problem. According to Chawla, Bowyer, Hall, & Kegelmeyer, (2002:321), "A dataset is imbalanced if the classification categories are not approximately equally represented."

Imbalance class seriously cause negative performance effects on the learning techniques of credit scoring that predict distribution balanced class (Chawla et al., 2004). Class imbalance distributions set a challenge for classifications and learning algorithms (Ofek et al., 2017). Problems also identified with the imbalanced distribution of data is the complex models learn by algorithms which have less relevance to the data and overfit the purpose of the data (Y.-M. Huang et al., 2006). Algorithms known as best standard classifiers tend to be biased towards the majority examples class, since the patterns that forecast the majority number of positive examples are weighted through the process learning in support of standard accuracy rate metric, and inconsiderate in the account of distribution class of the data. In such instances, the minority class with fewer examples have higher chances of being misclassified often than the class belonging to the majority examples (Pérez-Godoy, Fernández, Rivera, & Del Jesus, 2010). From scenario (2) in the diagram below, misclassifying the target class with fewer examples is much more expensive than misclassifying the majority class with larger examples (Chawla et al., 2002; H. He & Garcia, 2009).

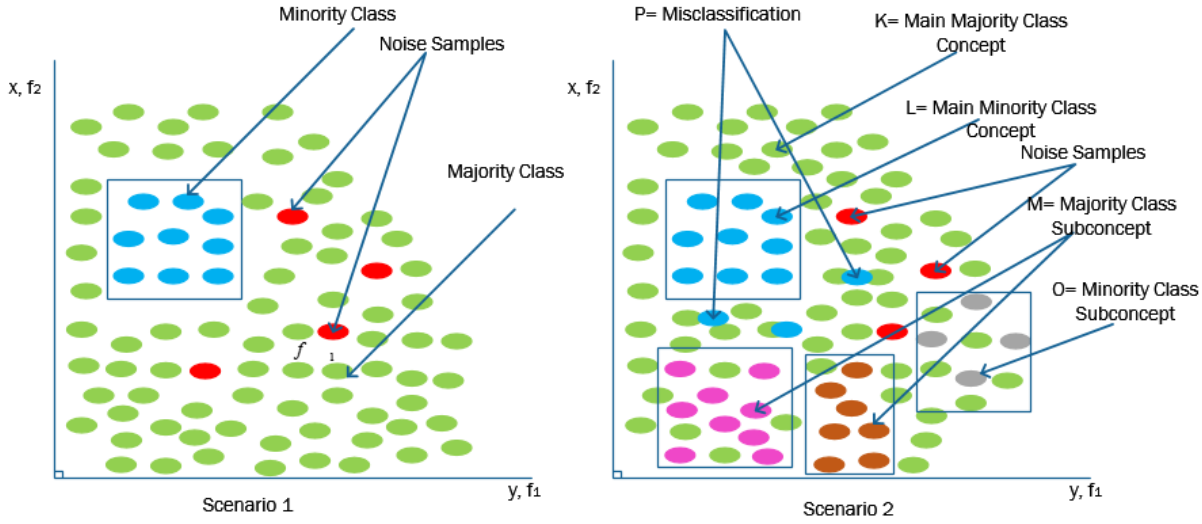


Figure 1: Typical diagram normal and complex imbalance situation Source: Author

Both scenarios above, the dominant class clusters as significantly more examples than the sub-concepts examples clusters, for the datasets to demonstrate both between-class and within-class imbalances. Scenario (1) is a between imbalanced datasets and scenario (2) is a complex imbalance dataset with both within-class and between-class, noise, overlapping, multiple concepts, misclassification and absence of representative data.

In Scenario (1), the blue balls represent the minority class whilst the green balls represent the majority Class. The red balls represent the noise in the data set with no overlapping and the concept used is only one. Scenario (2) also demonstrate data complexity with serious overlapping and multiple developments of concept which constitutes issues such as lack of representative data, overlapping, misclassification, noise, multiple concepts, and disjoints. The minority dominant class concept and subconcept class are represented by the Clusters labeled as "L" = Blue and "O"= Ash respectively. The dominant majority class concept and two sub-concepts class clusters are represented by "M" = Pink and Brown, and "K"= Green respectively. In both scenarios, the distributions depict imbalance relativity. In the case of "O" which is a subconcept of the minority data set in scenario (2) due to the lack of representative data, some inducers might ignore this subgroup. This is what we term as rare cases or instances.

One major problem with rarity is that the rare targets are difficult to find. Imbalance resulting from rare instances is lack of representation of the minority class within a specified datasets. That is to say, the target group is rare. Instances such as this, the absence of data in terms of the minority will affect the learning process irrespective of the between-class imbalance (Weiss, 2004). Such instance is the complete removal of the dominant minority cluster "L", the dataset would have a class concept of homogeneous minority, which is identified as Cluster "O". These data samples can be extremely rare or unlearned due to under-representation. In support, this minority subgroups may be comprising with several rare sub-concept instances which might result in several degrees of classification challenges (Holte, Acker, & Porter, 1989; Quinlan, 1986).

This situation affects the general performance of all classifiers since the distribution of the data is bias towards one end. This urge requires experts to solve these issues. Both industry and academia have paid attention to this problem due to the complex features of imbalanced datasets, learning from it demand new approaches, understanding, algorithms, and tools to efficiently change a large amount of raw data collected into a representation of knowledge and information. (H. He & Garcia, 2009:1263). Researchers have developed more interest in imbalanced data sets due to the cause of suboptimal performance of classification models (Chawla et al., 2004). However, application of knowledgeable expert should outwit the deficiencies of standard techniques for solving imbalanced datasets when the class of minority contains a lot of rare outliers and examples. It should be an improvement of the less class while sustaining better accuracy classification of the larger class than the standard techniques (Napierała & Stefanowski, 2015). Techniques used in solving the problem of imbalanced class datasets can be group into two: algorithmic approaches and data approaches (H. He & Garcia, 2009; Nanni, Fantozzi, & Lazzarini, 2015; S. Ramanna, Jain, & Howlett, 2013). Three approaches are mainly used; cost-sensitive learning, data resampling, and method adaptation. Among these three approaches, data resampling is the most commonly used either under-sampling the larger class or over-sampling the fewer class. In the adoption of any of the approaches, there is a trade-off between the complexity and the predictive performance (Ofek et al., 2017). Sampling technique is the most appropriate methods to be adopted if the size of data for training has limited attributes issues or data acquisition for training data is costly. In situations where one class within the distribution is very rare, all the attributes belonging to that rare class should be included in the sampling methods. Rare examples share common attributes and can be well defined within a specific domain under study (Weiss, 2004).

Data acquisition as the core of all research sometimes present the challenge of biasness and it is the researcher responsibility to identify such biases. It is also the researcher's responsibility to also address the technique adopted for solving the problem at hand either over sampling the data or under sampling it to have a balanced data set for training. According to L. Wei et al., (2007:434) Credit applicant data set from one giant commercial bank in the US. The datasets contain 5000 detailed applicant records, which comprises of 815 bad applicants and 4185 good applicants. From Crone & Finlay (2012:229-230), two datasets were used which they labeled as dataset A and dataset B. After removal of all indeterminate and outliers, the size of data retrieved from Experian UK for Dataset A contains 88,789 as observations, divided into 13,261 noted as bad applicants and 75,528 as good applicants. On the other hand, data set B was retrieved from retailer catalog from a mail order who provides revolving credit. The data sample retrieved based on their selection criteria contained 18,098 as bad creditors and 120,508 as good creditors.

The most widely adopted techniques for over sampling rare cases or imbalance numerical datasets is called SMOTE (Synthetic Minority Over-Sampling Technique). In applying SMOTE, it generates randomly new instances of the fewer classes that are in

between the boundary of both the instances of the minority class and that of the instances of the majority class (Chawla et al., 2002). Even though most researchers adopt sampling methods for addressing rare and bias data cases which are mostly expressed in discrimination ratios in defining both classes due to the problem at hand. According to H. He & Garcia (2009), ratios used and considered for imbalanced data set experiments are mostly 5:1 where the majority cases are selected per five at random whiles one per minority at random or higher ratios are preferred.

In the situation of Weng & Poon, (2006:273-275), they tested their experiment of imbalance situation on text documents (Physics and news) due to the limitation of bridging methods as a sampler. In their effort to demonstrate the effect of data imbalance situation performance on classifiers. They adopted Fisher discrimination ratios of 5:5 (minority and majority) for physics text document and 4:6 (minority and majority) news text documents. Y.-M. Huang et al. (2006), selected at random 1000 records as their targets from records of 62,621. They distributed their selection criteria in a ratio of 19:15:66 into three (3) classes. Class 1, Class 2 and Class 3 were respectively represented as 19:15:16 where class 3 were the majority group.

Data segmentation is identified by Weiss (2004:12) as another method for dealing with imbalance data or data rarity situation. Through careful segmenting of the data would result in reducing the degree of rarity in the dataset. Efforts to segment the data efficiently and effectively is to partition the dataset into different data mining subgroup problem. Shihab, Al-Nuaimy, Huang, & Eriksen, (2003:95) in data segmentation, you can target only the important data that is segmented for training the data from the lot while the less important segmented data are left unattended or discarded. Higher model performance can be achieved in homogeneity data set through segmentation of heterogeneous data set into sub-population or sub-groups (Bijak & Thomas, 2012; Chi & Hsu, 2012; LEE & ZHANG, 2003).

Examples from work of research conducted in demonstrating the accuracy effect of imbalance situations in datasets such as Weng & Poon (2006:273-275), text experiment with their selected ratio recorded 92% accuracy for physics and 95% accuracy for news with relatively balanced datasets. In the case of a reduced sample size of 20% of the minority, they recorded reduction in the performance accuracy of 89% for physics and 94% for news. In the extreme situation of 95% reduction of the minority sample size, they recorded 61% and 72% for physics and news respectively. In addition, in identifying factors that influence the accuracy of classification class problem, their finding shows that their classification accuracy without cleaning the data set is 80% for all classes. After the adjustments of the proportion of the training and cleaning of the data set test results was increased up to 64% for each class (Y.-M. Huang et al., 2006). In the case of Ofek et al. (2017:88) in solving binary class imbalance problem, they adopted under-sampling methods called fast novel clustering-based which depicted high performance in terms of prediction while the size of the lesser class instances is bound to its complexity of time. In their training stage, the algorithms group the minority classes and pick a number similar to the minority classes from the majority classes for each of the individual clusters. A classifier is selected for training each cluster. In addition, any unlabeled classes are classified as the class of the majority if it does not pair into any of the clusters.

However, support vector machine (SVM) and neural networks as algorithmic approaches have achieved great success in solving imbalance problems (Japkowicz, Myers, & Gluck, 1995; Raskutti & Kowalczyk, 2004). Others have also used other algorithms approaches in solving imbalance situation as Brown & Mues (2012) adopted several methods to analysis imbalanced credit scoring data sets using five real-world data sets to develop their classifiers for experiments. In their findings, classifiers such as random forest and gradient boosting were performing well in analyzing the credit scoring model and comparatively could cope with class imbalances in the data sets used. They also identified that with the large class imbalance problem, decision tree (C4.5) algorithms, K-nearest neighbors, and quadratic discriminant analysis were significantly worse.

To conclude, techniques for solving imbalance data set with the complexity of time have lesser prediction performance whilst more complex technique is slow significantly. In most instances, this would not demonstrate superiority (Ofek et al., 2017). Imbalance data set has received much attention with great success in solving that problem.

2.1.4.2 Feature selection

Real-world data are growing each day by the numerous activities being conducted each business day. Banks and other financial institutions gather tons of data from their weekly, monthly and annual reports. As these data increases, it is characterized with complexities, misleading or irrelevant information, and inconsistencies with an abundance of noise for any study. According to Y. Liu & Schumann (2005:1099) data used for credit scoring are collected from real-world data which has several sources with the intention for the general task. It is probable that the collected data originally would not only contain observations but would have several features. Some of these features accompanied with these observations may be unimportant to the credit risk model and some that are redundant would lead to high intercorrelation. Blum & Langley, (1997:245), machine learning aims to solve more complex and difficult task, the issues targeting on the most important information in a large amount of data has become very important aspect for researchers. For instance, mining data in a scientific or corporate environment records requires addressing many examples and features, as well as the internet and web giving large sums of poor quality information to be accessed with ease for a learning system.

Netnews, electronic mail, etc. cannot be left out in the personalization of information retrieval from information filtering systems. Classification algorithms are faced with computational delays, lower predictive accuracy and also difficulty in interpreting scoring models due to the abundance of redundant and unimportant features (Y. Liu & Schumann, 2005). In other to minimize what features or characteristics adopted for a particular study lead to the development of feature selection. Properties, attributes or characteristics are another named used for features. In gathering features with its values builds up a flat data file which depicts an application with each line depicting patterns (or instance, examples, records, case, tuple). The features collected can have continuous values or discrete, or have a complex form (H. Liu & Motoda, 1998:3). In addition to H. Liu & Motoda, (1998), ever since, feature selection has gain recognition over the years in data mining and machine learning communities by researchers and it has been an active area for solving credit scoring problem.

According to Jamali, Bazmara, & Jafari, (2012:42) feature selections has been adopted in several fields such as filtering and information retrieval, classification of text, web categorization, risk management, credit cards fraud detections and medical diagnosis. “Feature selection addresses the dimensionality reduction problem by determining a subset of available features to build a good model for classification or prediction” (Maldonado, Weber, & Famili, 2014:228). According to Blum & Langley (1997), Feature selection focuses on the most relevant features by eliminating features with no or less predictive information. It eliminates redundancy, solves dimensionality problems, ignores noisy data or irrelevant data. It improves on the effect of an application; speed

up time for the algorithm, quality of data is improved, as well as the performance of classifiers. In summary, feature selection techniques methods help to improve classification system development. In reducing features, it helps in the improvement of prediction accuracy and the cost of computation to be minimized (Zhang, 2000). H. Liu & Motoda (1998) describes the reasons for choosing feature selection may be because of challenges facing most learning algorithm design. Aside from the influences of this function on the behavior of the algorithm, it determines if the solution is applicable to the problem or how the learning algorithm would work based upon the power of the language used to describe the problem. They added that for the adaptation of feature selection is due to its convenient, primitive, widely used, independent, and its usage is general. Danenas & Garsva, (2015); Fan, Yang, & Qin, (2013), identify three (3) important reasons why they adopted feature selection for training their data in credit risk and scoring domain was due to; (a) To minimize data dimensionality of the features by reducing the complexities in the model and forming new subset, (b) to have statistically significant features to develop a new classification model based on other evaluator, and (c) to ensure all that all featured used in the model are statistically significant.

Despite several techniques have been proposed for feature selection in data mining and machine learning literature. Feature selection has two main categories, thus, filter model and wrapper model (Jain & Zongker, 1997; Kohavi & John, 1997; Ping, 2009b; Somol, Baesens, Pudil, & Vanthienen, 2005). H Liu & Yu (2005) added a third approach has the integrated model with each having a different evaluation criterion. The integrated model uses a combination of both model advantages by exploiting the model in various evaluation criterion in different search stages. Summary and comparison of these models are further explained below. The filter model approach is dependent on the general properties of training a data to be able to select a set of features without the help of a learning algorithm. In choosing filter methods, the feature set evaluator estimates features independently. The main objective of an algorithm for feature selection is to capture some important measures for individual feature selected. Filter model is derived mostly from computational analysis of the samples of data in the training data set through which each feature is assigned a score. Upon ranking all the features, during the phase two of the learning system, interest is drawn towards achieving highest classification accuracy with a minimal feature selected. The following characteristics are exhibited by filter model: a) it is not dependent on the biasness of a classifier, but on the properties of the intrinsic data, in order to use the features selected for learning several classifiers; b) Measuring information gains, dependence, distance or consistency is often cheaper (in complexity of time) compared to measuring classifier accuracy, this makes the filter model produce faster subset results, all other things being equal; and c) The simple nature of measure and less time complexity, a filter model is known for handling large data sizes compared to a classifier. Hence, situations where the classifier is unfit for to learn from large data, filter model approach can be used to reduce the dimensionality of the data for the classifier to learn from that data. However, there is a risk that the selected features by the filter model approach cannot enable full exploitation of the biasness of a learning algorithm (Huan Liu & Motoda, 1998:33-38).

In the case of wrapper model, it demands one predefined learning algorithm in the process of feature selection. Selected features are dependents upon performance effect of the learning algorithm. This model needs to re-train the classifier anytime new feature subset is selected. Wrapper model ability to identify features best suit to the predefined learning algorithm yields to the superiority of the learning performance. This model is more expensive computationally since it has to embed the learning algorithm in it selection processes (Hengpraprohm & Chongstitvatana, 2009; Huan Liu & Motoda, 1998) whiles the integrated model approach differentiates itself by the use of both wrapper and filter model at the same time by integrating four (4) procedures for feature selection. These procedures are Relief, chi-squared, InfoGain and Gain Ratio, while the first two is associated with filter model which exploit the intrinsic features of the data without applying a predefined algorithm such as the last two procedures which belong to wrapper model (Dailing Zhang & Xu, 2013:558).

Feature selection has been adopted in credit scoring models since it is a key factor or option to improve the performance of a hybrid model for accurate prediction (Ping, 2009b; Ping & Yongheng, 2011; Somol et al., 2005). To improve the performance of a classifier by reducing redundancy and selecting the independent predictive features, experts and researchers in the field have these methods. Some have adopted univariate analysis to assess effects of each feature independently on the class feature; others adopted statistical correlation analysis to identify the correlation between various independent features and to remove highly correlated ones. In addition, the procedures of feature selection are sometimes integrated into algorithm classifiers, instances such as the usage of stepwise statistical methods in discriminate analysis and regression methods, and choosing suitable desired features when building decision trees (Y. Liu & Schumann, 2005:1099).

There is no economic theory supporting which features to be used as relevant in determining creditworthiness, hence procedures for choosing any set of features as best for determining credit-worthiness is unsystematic and control by arbitrary trail (D. J. Hand & Henley, 1997). Even though, in credit scoring, it is important to use few features in deciding on approval of credits since it yields better insight and understanding of the model to the credit scorecard builder (Somol et al., 2005). On the contrary, this approach is known as unsystematic procedures, methods adopted for feature selection is the automated data mining procedure. Several experts in machine learning have come up with several selection feature methods. There are no literature or methods as efficient for solving feature selection problem in modeling credit scoring (Y. Liu & Schumann, 2005).

Feature selection is also a perfect technique for solving class imbalance situation (del Castillo & Serrano, 2004; Z. Zheng, Wu, & Srihari, 2004). Even though, Zheng et al., (2004) stated that measures adopted for the use of feature selection in imbalanced data sets are appropriate. They suggested a framework for feature selection, which is able to targets features separately for both negative and positive classes and explicitly put them together. In addition, they demonstrated ways of transforming the already existing measures to be able to consider separately positive and negative features. Few credit scoring studies have developed a scoring model using features selection function as an option to hybrid their models to enhance their base classifiers predictive accuracy.

In conclusion, it is beneficial to have a larger sample size but it is more efficient to have a smaller sample size. The problem of identifying efficient sample distribution (balancing) and size to improve significantly the performance of predictive accuracy of various algorithms on a particular data set has not been considered. It is in the same vein, that limited attention has been paid to feature selection technique as a data preprocessing issue in credit scoring industry (Crone & Finlay, 2012). Neither has there been any efficient permanent solution for solving credit scoring problem using feature selection technique.

2.1.4.3 Algorithmic solutions

Several developmental approaches and methods have been adopted in the credit scoring industry over the past years to better understand credit models that would best predict with an accuracy of the worthiness of applicants' credits with the minimal loss in the industry. Several credits models have been proposed which span from optimization approaches to statistical approach than to

algorithms and artificial intelligent systems. Artificial intelligence has been adopted and accepted widely in recent years for building models of credit scoring which results have improved significantly. Hence, this makes it more meaningful to study recent methods of artificial intelligent in solving problem in the credit scoring market (Chen & Lin, 2014: 658). The Naive Bayes Rule (NBR) was used on two real-world data sets as classification rules to groups credit applicants into bad and good. For this purpose, the classifier was benched marked against logistic regression analysis, linear discriminant analysis, k-nearest neighbors, neural networks, and classification trees. Results from each of the instances experimented, Naive Bayes Rule (NBR) perform lower in terms of it predictive accuracy (Antonakis & Sfakianakis, 2009). Harris (2013:4404), Credit scoring model was built based on support vector machine using narrow (greater than 90 due days past) and Broad (Less than 90 due days) default definitions to predict and classify risk accurately for Barbados credit union data set. In comparison, the model with the broad default definition outperforms the model with the narrow definition default.

In the findings of Ala'Raj & Abbod (2015:5) which was conducted using two data sets from German and Australia demonstrated that the test on type I and type II error in the Germany data sets ANN and LR achieved the lowest type II error by 11.4% and SVM scoring 28.3% as type I error. The implication of these states that SVM was outperforming in classifying loan default accurately unlike ANN and LR that was not perfect, in terms of good loan classification ANN and LR was superior to SVM. In addition, in the case of the Australian data set, ANN classified bad loans accurately and all other models achieved same error rate for type II error. This works showed that all three models perform well in classifying good loans in Australian data set than German data set. Recent literature emerging suggests that hybrid credit scoring models may perform better than the individual models known as the best (K. Li, Niskanen, Kolehmainen, & Niskanen, 2016). Development from credit scoring lead to various researchers having proposed various highly effective and sophisticated data mining techniques or models such as skew-t discriminant analysis (STDA), Stepwise discriminant analysis (SDA), skew-normal discriminant analysis (SNDA), sparse discriminant analysis (Sparse DA), Mixture discriminant analysis (MDA) and Flexible discriminant analysis (FDA) for screening applicants who apply for credits. These models were tested on one real-world data to identify the best performing model. To evaluate the performance in terms of predictability of each model, the model was accessed on the "bad rate among accepts (BRA)" and "total percentage of correctly classified cases (Total PCC)". The results show that only STDA, SNDA, and SDA were outperforming methods for credit scoring implementation model (Chen & Chen 2010:1).

Nanni & Lumini (2009:3029-3033) was the first to compare different methods of ensembler for credit scoring and predicting bankruptcy. Their experiment was centered on datasets from Japan, Australia and German using stand alone, AUC, bagging, random subspace (RS), rotation forest (RF), class switching as ensemble classifiers to test the performance of Multi-layer Perceptron Neural Network (MLP), Radial Basis Function SVM (RSVM), and Levenberg-Marquardt Neural Network (LMNN) and old 5-nearest neural net (5-NN). Findings show that using classification ensemble has superior performance and LMNN being the best-tested method with random subspace (RS). In Comparing LR, BPNN, linear programming to Fuzzy Art the results deduced show that Fuzzy ART has marginal accuracy with far lower type II error compared to LR, Back propagation neural networks and linear programming whereas banks interest are in type II error (M. Jiang & Lin, 2010:436). Tsai & Wu (2008: 2643-2649), demonstrated the performance of predictive accuracy of a proposed hybrid model on three datasets from German, Australia, and Japan were used for the experiment using ANN and another complex model. In theory, it proves that multiples classifiers (Combination of models) perform better compared than single classifiers. Results show otherwise, on the average performance and predictive accuracy single classifiers are best compared to the multiple classifiers.

Findings from an experiments to determine the accuracy of Hybrid credit scoring model(HCSM) against Support Vector Machine (SVM), Genetic Programming (GP), Logistic Regression (LR), Decision Tree (C4.5/5.0), and Back Propagation Neural Network (BPN) on two different datasets from German and Australia shows that HCSM accuracy classificatory rate was higher compared to other five models. In addition, test on both datasets demonstrates that SVM, BPN, GP, and LR performed very well and can be used as an alternative to each other while the C4.5 model was inferior significantly (Zhang et al. 2008: 10-11).
According to Wang et al. (2012:61) Decision tree (DT) is popular amongst classification algorithms in machine learning and data mining. Even though, it is popular it still performs poorer compared to other techniques due to its inability to deal with data noise and redundant attributes of data in credit scoring circumstances. Two-stage ensembler trees were introduced thus, Bagging-RS DT and RS-Bagging DT based on random subspace and bagging. This is to reduce noise influence, data redundant attributes and enhance the classification accuracy. Findings from this work showed that Decision Tree comparable to four single classifiers namely; linear discriminant analysis, radial basis function network, logistic regression, Multi-layer perceptron performs worse in terms of classification accuracy. Bagging-RS DT and RS-Bagging DT was better in terms of classification accuracy compared to the five single classifiers.

Neural networks and regression are identified as powerful tools for classification but difficult and exhausting to identify the best platform for a particular problem. Binary particles swarm optimization (BPS) and genetic algorithm (GA) are adopted to customize Multi-layer perceptron neural network (MLP) to enhance it prediction accuracy in terms of credit risk scorecards. Findings from this work demonstrated that both methods perform better than the single classifiers in terms of predictability even though GA (genetic algorithm) was time-consuming (Correa & Gonzalez, 2011). Lin et al. (2011) study propose three approaches with a combination of two well-known classification models, namely, Support vector machine (SVM) and K-Nearest Neighbor to identify the best combination hybrid classifiers. The different scoring combination was developed through selecting features with two classifiers and three approaches. The experiments are conducted using data sets from Irvine, University of California (UCI) to test for the accuracy of the hybrid features selection models. SVM and KNN classifiers combine with F-score, Rough Sets (RST), and linear discriminate analysis (LDA) approaches as suggested features steps for optimization to remove both redundant and irrelevant features from the analysis. After evaluation of the proposed models, F-score approach with the combine classifiers performs better in all two data sets.

In Huang et al. (2007), three strategies were adopted to build the hybrid support vector machine based models for credit scoring to analyze applicants' credit ratings from input feature of the applicant. The test was conducted on two data set to demonstrate which of the classifiers would perform better in accurately classifying applicants. Evaluation depicts that neural networks, decision tree, and genetic programming classifiers are incomparable to the SVM classifier on both data sets with few input features relatively. Ala'Raj & Abbod (2015: 6) used data set from Germany and Australia to compare the performance of each model even though they focused on a heterogeneous combination in arriving at their results. Logistic regression (LR) was rated in terms of accuracy as the highest in the German data set followed by ANN and SVM getting the lowest in type I error. In addition, the Australian dataset,

ANN (artificial neural network) had the highest accuracy with the lowest type I error scored. They identified that Logistic regression (LR) was superior to the other models by achieving the highest accuracy with the lowest type I error as well as support vector machine (SVM) in the Australian and German data set which would cost banks and financial institutions more.

Test for cross classification which was conducted by Dahiya et al. (2015: 170-172) using three data partitions demonstrated that Ensemble model was rank as the best performing model in all partition cases whiles SVM and LR model performance was better compared to all other base classifiers individually. To add, SVM was performing better among the individual models in 50:50 partition. Kambal et al. (2013: 381-383) conducted it credit scoring models experiments for Sudanese banks and included German data set. The experiments set its focus on only two models that are Artificial Neural Network (ANN) and Decision Tree (DT) and applied Principal Component Analysis (PCA) and Genetic Algorithms (GA) as feature selection technique. The findings demonstrate ANN as the best performing model compared to DT model in all the cases.

Lessmann et al. (2015: 129-134) did a comparison study of 41 classifiers using six (6) measuring performance tools on eight (8) credit scoring real-world datasets. Results depicted Logistic regression which could not significantly give accurate results in the standard industry environment while several classifiers significantly predicted risk associated to credit accurately. In addition, they suggested outperforming LR to be excluded in the industry environment since it is unacceptable methodological signal advancement given the state-of-the-art. Wang et al. (2011: 284-287) design model for assessing green credit risk using support vector machine. The real-world data set was collected to test for performance of the scoring system and the risk model. Artificial neural network and logistic regression were employed as benchmarks. Findings demonstrated SVM model for assessing risk as highly significant, performs better in specificity, sensitivity, and correctly classified cases in terms of percentage compared to the rest of the models.

In the work of West et al. (2005: 2550-2558), using mean generalization error in three datasets from German, Australia and bankruptcy data to test the best accurate predictive ensembler from three (3) most recent classifiers methods namely; boosting, bagging, and cross validation and multilayer perceptron NN as a base. The result shows that MLPNN is more robust and accurate based on all test conducted than the single model. Bagging and boosting strategy were best for Australian and bankruptcy data and bankruptcy data respectively whilst cross validation we accurate for the only German data set. Yu et al. (2008:1440-1443) focus of concern was on the multistate neural network, in addition, to proving his proposed model he attempted to compare his model to ANN, fuzzy SVM, majority voting ensemble model, SVM and logistic regression (LR). Their proposed model is consistently reliable in terms of performance and predicting accuracy over the single models, majority voting based ensemble and hybrid models. Zhang et al. (2010: 7839-7842) proposed a model vertical bagging decision trees model (VBDTM) which was built on multi-stage ANN classifier ensemble method of sampling. The results from their experiments show that VBDTM is more robust and accurate compared to other best single models such as SVM, NN, and others models tested.

To conclude, both artificial intelligence and statistical techniques have been extensively explored for credit scoring, there is no single model which has been consistent in terms of performance in all conducted works (G. Wang et al., 2011). Traditional use of combination classifiers methods is noted for achieving worse accuracy compared to good classifiers and averagely better than bad ones (Ala'raj & Abbod, 2015). From all findings shows that hybrid models are better than single classifiers in terms of robustness and effectiveness but that does not guarantee that those solutions are permanent because when tested on other data sets it is likely to perform worse than the base or single classifiers and others may perform better. Single classifiers with it best and ultimate performance deviate from one data set to the other hence it needs adjustment to give better accuracy in solving credit scoring problems. However, this work focuses on Support Vector Machine (SVM), Artificial Neural Network (ANN), Logistics Regression and the later addition, Random Forestry as choices as base classifiers which its detailed selection and mathematical modeling for this study are explained in the next section.

2.2 Theoretical Literature Review

It is a necessity in any field of science that almost all models have to be compared to one another to understand it's applicability of the model to a given situation. It is a common practice and requirement in data mining technique. This works looks at "Confusion Matrix from which other metrics would be discussed to the adaptation of "AUC" or "Area under ROC curve" which this work is centered on. In this section, we would only describe the theoretical knowledge or approach of confusion matrix and it analysis without its application, that would be handled in chapter five.

2.2.1 Introduction

According to Krzanowski & Hand (2009: chapter 1), in the area of science and dealing with a classification problem, a set of each object are identified to belong to one of two classes. Procedures used to assign this set of object to a class is realized based upon information gathered about the general object. Unfortunately, the procedure for assignment is not perfect: errors occur in the procedure, this is as results of sometimes assigning an object to a wrong class. These problems of imperfection necessitated for evaluating the performance quality of a procedure adopted. Through the evaluation processes, we can now decide whether these procedures used in assigning the objects to a class must be adopted for that purpose, must it be improved or should there be a replacement with another procedure.

Real problem examples of assigning classes cut-across several fields of study that includes:

a. Evaluating applicants for financial credit, where the objective is to assign each credit applied to a group of either "not likely to default" or "likely to default" class.
b. Speech recognition system developments, which requires that spoken words should be assigned to a class.
c. Grouping patients into either disease A or B through medical diagnosis test.
d. Incoming email filtration to identify whether they are genuine messages or spams.
e. Ranking potential applicant for a university course, with the aim of whether or not they would be likely to pass their final examination.
f. examining transactions for credit cards, to identify whether it is fraudulent or not.
g. Microarray data investigation to identify trends of a gene expression, whether it conforms to cancer or not.

The above problems listed are unlimited in nature. There are situation or scenarios where there are more than two classes involved, but under most circumstance, studies and solutions are applicable to only two class in practice where practitioner is to

choose between (yes/no, accept/reject, sick/well, wrong/right, condition absent/present, act/do not act, and so on). However, the situation involving more than to two classes are often segmented into series of two-class cases.
In classifying objects, the information used to assign this object to a class is considered as a descriptive variable of vector, features or characteristics. Each level of variable measurement would determine the type of information that one would obtain:

- categorization of values such as "eye color, skin color, etc" is known as a nominal variable.
- A nominal variable with two possible categories such as "absence or presence of pain" is known as a binary variable;
- categorization based on the ordered way such as "severe pain, moderate pain, mild pain, no pain" is also called ordinal variable.
- a possible value of distinct finite number such as "the number of staffs who wear glasses in GTUC in a class of 10" is known as discrete variables.
- In addition, a continuous variable is capable of handling either infinite or finite range such as "outcome of patient weight in a hospital.

In the case where the possible outcomes to be handled by a continuous variable is limited by the measuring device (e.g. If the results of the weights given by the machine are to the nearest grams), it must be treated generally as continuous rather than discrete due to its fundamental characteristics. Sometimes the descriptive vector variables will be univariate, thus, containing a single variable but most of the time, it would be multivariate. In the case of a single variable, it is treated as a proxy for or be approximated to some variable that can better define the classes. In a medical scenario such as taking sedimentation of erythrocyte rate. This is a medical screening used to measure inflammation. Inflammation results to sticking together of red blood cells, in other to fall faster. The goal of such a work is to assign each patient to not ill or ill possible outcome classes based on their test responses. In the case of this work, we would be dealing with a binary variable as well as possibility to assign each borrower to a class of non-defaulter/defaulter based on the experiment on the dataset.

2.2.2 Assessment performance of classifiers for credit scoring models

2.2.2.1 Methods for evaluation of classifiers

In dealing with classification problems, it is ideal to measure the performance of a classifier in terms of error or misclassification rate. For a classifier to predict accurately each class correctly it is recorded as "Success", else, it is an error. The rate of misclassification is the ratio of errors committed over the whole instance sets and the overall classifier performance it measures. The most used methods for evaluating the performance of classifiers are described below:

- **Cross-Validation (CV):** CV is a procedure for approximating the performance of a general predictive model. The basis for Cross-validation is to split data into one or more times, for evaluating each algorithm risks: Portion of the data (sample for training) is used for algorithm training, and the rest of the data (Validating sample) are used for evaluating the algorithm risk. The estimated smallest risk by the algorithm is selected by the CV. Random Sub-Sampling is the alternative to this methods.
- **Random Sub-Sampling (RSS):** Repeating severally of a hold-out technique to enhance the estimation of the performance of classifiers is known as Random Sub-Sampling. The RSS is subject to encounter similar problems of the hold out technique since it does not use enough data for training. Adopting this technique, means you do have control over the repeated times of usage of record for training or testing. This shows that some data might be used more often for training than others.
- **Holdout Method:** Holdout depends on a single data split. This technique is considered as the simplest of CV. We separate the dataset into two main sets, which are testing set, and training set. The estimator function fits only the training sets using its function. Through this process, the estimator is queried to predict using the testing set data on output values it has never seen before. The errors it commits through these processes are accumulated to calculate the absolute mean test error, which is adopted for evaluating the model. This technique is preferable to the residual technique and has a shorter time for computation. However, the variance can be very high after the process of evaluation. The evaluation of this technique solely depends on the selection criteria based upon the data points that fall in the testing set and training set, which can significantly influence the evaluation, based on how the division is determined.
- **Leave-one-out Technique:** If K-fold Cross Validation is denoted as "*K*" and the number of data points in a set denoted as "*N*". With "*K*" equal to "*N*" (*K=N*), implies for *N* separate times, all the data are trained by the approximator function except one point that is used for prediction. To evaluate this model you must compute the average error. The ratings given by the (LOO-XVE) leave-one-out CV error is good, but its computation looks very expensive at first instance. The good side of this technique is that amateur in this field can easily adapt this technique to make LOO (Leave-one-out) predictions as easy as any regular predictions without in-depth knowledge into predictions. The computational time for leave-one-out cross validation error (LOO-XVE) is shorter compared to residual error and it is the best way for evaluating models.
- **K-fold Cross-Validation:** This technique is a procedure for enhancing holdout method. This technique requires that the dataset should be divided into *k* sub units, and the hold-out technique is reiterated *k* times. In each of these process, one of the *k* sub units is adapted for testing sets and the other sub units *k-1* are combined to form the training set. Through these processes, all the averages of the errors for k-trails are calculated. This technique is not hindered by how the divisions of the data are achieved. Each data points are used as testing sets only once and used in the training set *k-1* times. As *k* increases the estimates, resulting in the variance reduces. Even though this technique has its good side, computations for algorithm evaluation has to be k times and in each instance, the training algorithm has to be started from initial *k* times. This technique variant is to divide randomly the data into training and testing set *k*, unlike times. The benefit for the adaptation of this technique is that you have control over choosing how large you would want your testing set to be and the number of trails you want to average over.
- **Bootstrapping:** This technique distinguishes itself from all the above mention techniques that select or divides its sampling into testing and training set without reusing the sample dataset or replacing them. In bootstrapping technique, all the sets

selected for training the algorithms are replaced back into the original data set pool to have equal chances of being used for analysis.

2.2.2.2 Confusion Matrix

With reference to Krzanowski & Hand (2009: Chapter 1-3) using the assumption that v is the threshold value of V in a specific classification rule, where each individual is allocated to a population P, thus, where the scores of classification s exceed v else to the N population. To assess the efficiency of the classifier we must be able to calculate the possibility of misallocation of a population. In this sense, we might probably be able to tell in the future the rate at which misallocation might when classifying individual objects. From this, we would be able to derive four possibilities with their corresponding rates for a classifier:

1. The possibility that objects within the P group are classified correctly. That would be represented as true positive rate $tp = p(s \geq v|P)$;
2. The possibility that objects within the N group are misclassified. That would be represented as false positive rate $fp = p(s > v|N)$;
3. The possibility that objects within the N group are classified correctly. That would be represented as true negatives rate $tn = p(s \leq v|P)$;
4. The possibility that objects within the P group are misclassified. That might be represented as false negative rate $fn = p(s < v|P)$.

From this scenario, to attain a perfect classification then one must be able to identify a threshold V=v where all individuals members in the class of P(Positive) might have a score greater than v, and all those in class N (Negative) might have a score less than or equal to v.

From Table 1 below is a typical representation of the above binary situation where we have two instances: true and false instance. That results in the possibility of four (4) classification outcomes indication: a true negative (Correct Rejection), a false negative (Misses), a true positive (Hits) and a false positive (False alarms). Hence, tab. 1 is described in a contingency table, which is also known as confusion matrix. The contingency table positions the predicted models for classification the rows with the observed phenomenon of the classification on the columns. In classifying a model, the correct classification that is true negatives and positives lies on the diagonal of the contingency table. Model errors are specified by other fields. A perfect model is depicted with only the true negatives and positives filled out fields whiles the other fields would be equal to zero (0). Some model evaluators or metrics have can be deduced from the confusion matrix. These metrics are discussed below.

Table 1 Contingency Table (Confusion Matrix)

		ACTUAL VALUE (as confirmed by experiment)	
		POSITIVES (YES)	NEGATIVES (NO)
PREDICTED VALUE (Predicted by test)	POSITIVES (YES)	**TP** TRUE POSITIVE	**FP** FALSE POSITIVE (Type I error)
	NEGATIVES (NO)	**FN** FALSE NEGATIVE (Type II error)	**TN** TRUE NEGATIVE

Previous studies in credit scoring industry have adopted to accuracy as the most used evaluator to measure performance of a classifier either in solving multi-class or binary classification issues (Chawla et al., 2004; Desai, Crook, & Overstreet, 1996; Gu, Zhu, & Cai, 2009; Hossin, Sulaiman, Mustapha, Mustapha, & Rahmat, 2011; Ranawana & Palade, 2006; VENKAT & KIM, 1987). In using accuracy to measure, the quality of solution provided by a classifier is as a result of the percentage of predicted correctly over the total classes. This is presented in Equation 1.1

$$\boldsymbol{ACCURACY} = \frac{TP + TN}{TP + FP + TN + FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots EQN.(1.1)$$

Accuracy cannot be measured without adding error rate since is a key factor needed to evaluate the incorrect prediction that is resulted from the percentage of incorrect prediction over the total classes. This can be represented in equation 1.2.

$$\boldsymbol{ERROR\ RATE} = \frac{FP + FN}{TP + FP + TN + FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots EQN.(1.2)$$

Most researchers in choosing an optimal solution and discriminating among algorithms have used both metrics extensively due to its advantage. These metrics computations are easy with fewer complexities; it has good standings in measuring multi-label and multi-class problems; very simple for scoring with simple interpretation for understanding by a novice. However, these two metrics has its weakness as evaluators and discriminatory processors. A major weakness of the accuracy metrics is biased towards the majority class and unable to distinguish overall performance accurately since the minor class turns to be at disadvantage (Weiss, 2004). In imbalance situation, all or majority of the instances in the minority class are predicted wrongly unlike the majority class

(H. Han, Wang, & Mao, 2005). This makes it very difficult being able to identify the optimal classifier. In addition, these metrics are less informative due to its simplicity (MacKay, 2005).

The sensitivity ratio is also another metrics derived from the confusion matrix. This ratio is calculated from the percentage of positive class accurately predicted or classified. The sensitivity ratio is also known as Hit rate or recall or true positive rate from other studies. This ratio is represented by Equation 1.3.

$$\boldsymbol{SENSITIVITY\ RATIO} = \frac{TP}{TP + FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..EQN.(1.3)$$

The sensitivity ratio is the best suit for evaluating performance problems with imbalance dataset problem since that can be used for either minority or majority class. Studies from (Zekic-Susac, Sarlija, & Bensic, 2004) demonstrated the application of using sensitivity ratio in both classes that are bad and good applicant cases. This shows that this metrics can be used for either bad-good applicant or majority-minority classes based upon your targeted class at a time for each instance.

Precision metrics is a derivative from the confusion matrix. It measures the accuracy of a class in question. It can either be used for positive or negative class depending on the class given. It is represent based on equation (1.4)

$$\boldsymbol{PRECISION} = \frac{TP}{TP + FP}\ (Positive\ cases)\ OR\ \frac{TN}{TN + FN}\ (Negative\ cases) \ldots.EQN.(1.4)$$

Anytime F-measures is adopted then it gives a signal that precision and sensitivity ratio have used in their analysis. F-measure metric is adapted to observe or watch both precision and sensitivity ratio at the same time (Van Rijsbergen, 1979). If all factors influencing precision and sensitivity ratio are held constant, then we expect their weights to be equal. This can be written in an equation as (1.5).

$$\boldsymbol{F - MEASURE} = \frac{2 * PRECISION * SENSITIVITY\ RATIO}{PRECISION + SENSITIVITY\ RATION} \ldots\ldots\ldots..EQN.(1.5)$$

Even though G-Mean and F-Measure are good improvements in terms of accuracy, their lapses make it ineffective in answering various generic questions about evaluating classification. A typical example, in this case, is, *which ways can we compare the different classifiers in terms of performance over a range of sample distribution?* (H. He & Garcia, 2009). On the other hand, F-measure may be low due to low accuracy from the minority class even though the general accuracy might be very high (H. Han et al., 2005).

Table 2 Confusion Matrix for Cost-Sensitive (Type I and Type II error)

		Actual Condition	
		Positive (Risk Free)	**Negative (Risk)**
Experimental Results	**Positive (Risk Free)**	True Positive (TP)	Type I Error (FP)
	Negative (Risk)	Type II Error (FN)	True Negative (TN)

From the table 2 above is the cost sensitive learning that in practice is called the type I error and type II error. These are graded in the Cost and savings in terms of predicting the actual positives and actual negatives. The type I error represented as the actual negatives or bad applicants being presented as good or positive applicants. This type of error would be a great loss than savings to the credit firm and it is more serious to commit this type of error. In the case of type II error, which is represented as the actual positives or good applicants being presents as negatives or bad applicant would be savings to the credit firm since the results from the analysis would be to deny applicant access to credit requested. This can be elaborated from the below topic and equations.

Cost Sensitive Learning: Elkan (2001) interpreted the cost sensitive basic rule for machine learning very well from the confusion matrix. This knowledge was later adopted and applied in credit scoring industry by (Abdou, El-Masry, & Pointon, 2007) using West (2000) lay down procedures as a guide. Cost Sensitive learning function applied by Abdou et al., (2007) can be express as in equation 1.6

NB: Let $\pi_G = \beta_G$, $\pi_B = \beta_B$

$$CSL = P_{B-G} * C_{B-G} * \beta_B + P_{G-B} * \beta_G * C_{G-B} \ldots\ldots\ldots\ldots\ldots\ldots\ldots EQN.(1.6)$$

In equation 1.6, the equation variables can be denoted as: β_B is the probability prior to a bad applicants, β_G is the probability prior to a good applicants, P_{B-G} is the possibility of predicting bad applicants as a good applicants, P_{G-B} is the possibility of predicting good applicants as bad applicants, C_{B-G} is the cost involved in predicting true bad applicants as good applicants, C_{G-B} is the cost involved in predicting true good applicants as bad applicants.

From the above function Abdou et al., (2007) explain that β_G and β_B is a derivation from the below equations:

$$\beta_G = \frac{fn+tp}{fp+tp+fn+tn} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots EQN.\ (1.6.1)$$

$$\beta_B = \frac{fp+tn}{fp+tp+fn+tn} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots EQN.\ (1.6.2)$$

From Elkan (2001) point of view to this function is to minimize failures and record high efficiency if you want to adopt this as cost sensitive measures. To determine how well an algorithm performs is the size of the values from this cost function. The smaller the value the better the accuracy and vice versa.

From West (2000) definition, P_{G-B} and P_{B-G} are actual rate of false negatives and actual rate of false positives respectively, and is a derivation from the equations below:

$$P_{G-B} = \left(\frac{fn}{tp + fn}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.EQN.(1.6.3)$$

$$P_{B-G} = \left(\frac{fp}{fp+tn}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . EQN.(1.6.4)$$

Substituting equation 1.6.1, 1.6.2, 1.6.3 and 1.6.4 into equation 1.6. A rewrite of the whole equation would be:

$$CSL = \left[\frac{fp}{fp+tn}\right] * \left[\frac{fn+tn}{fp+tp+fn+tn}\right] * (C_{B-G}) + \left[\frac{fn}{tp+fn}\right] * \left[\frac{fn+tp}{fp+tp+fn+tn}\right] * (C_{G-B}) \ldots . \ldots . EQN.(1.7)$$

Simplification of equation (1.7) would be rewritten as:

$$CSL = \left[\frac{fp}{fp+tp+fn+tn}\right] * (C_{B-G}) + \left[\frac{fn}{fp+tp+fn+tn}\right] * (C_{G-B}) \ldots\ldots\ldots\ldots EQN.(1.8)$$

From equation (1.8), the function is said to be true when good applicants are treated or selected as positive applicants. In the cost sensitive learning, we can demonstrate bad applicants being selected as good applicants or positive class would be represented in an equation as:

$$CSL = \left[\frac{fp}{fp+tp+fn+tn}\right] * (C_{G-B}) + \left[\frac{fn}{fp+tp+fn+tn}\right] * (C_{B-G}) \ldots\ldots\ldots\ldots EQN.(1.9)$$

From equation (1.9), bad applicants are now predicted as positive; in this case, all good applicants are predictions from selecting bad applicants as good applicants, which is represented by fn whilst all bad applicants are predictions from selecting good applicants as bad applicants, which is represented as fp.

In cost sensitive learning, it is always an advantage to predict good applicants in a class of positives, researchers argues that there is a lower risk in predicting positive (good) applicants as negative (bad) applicants than situation when bad applicants are predicted as good applicants with higher risks (Nayak & Turvey, 1997). In addition, in the results of mistakes when bad applicants are predicted as good would lead to the following lost consequences to the lender: profit on the principal, principal, administration fees, coverage of insurance, taxes on property and legal fees (Nayak & Turvey, 1997:286). However, in the case of denying a good applicant with good credit history as bad applicants has not really being debated on the cost involved in making such mistakes except the cost of loss of profit, which would be accrued, from the principal when offered the credit.

From the concluding results for cost learning metrics from Dr. Hofmann (West, 2000:1147), Abdou et al. (2007) and West (2000) agreed and adopted five (5) as a relative factors for instances of predicting bad applicants as good applicants as against prediction of good applicants as bad applicants. This work would focus on also minimizing misclassification errors based on the algorithm selected. In addition, in measuring model performance the use of a scalar gives a poor summary of model performance especially in the case of adopting non-parametric models such support vector machine (SVM) or artificial neural networks (ANN) and some of the metrics developed from contingency table are sensitive to anomalies of data such as skewness of class. ROC was an intervention to resolve some of these issues which most of the evaluation metrics have failed to address.

2.2.2.3 Receiving Operating Characteristics (ROC) Curve

The definition and explanation of this metric would be discussed briefly from receiving operating characteristics (ROC) curve to the derivation of the area under (ROC) curve. This metrics is well noted for its wide usage and popularity within the data mining environment. Receiver Operating Characteristics (ROC) curve name originated from the usage of signal detection curve theorem (Egan, 1975; Green & Swets, 1966), with the aim of detecting a particular signal existence even though there may be genuine losses of few of these signals there should be a concurrent possible false alarm. In the theorem of signal detection, the objective is to assign individual tasks to a class of signals and non-signals. The word usage of "characteristics" in ROC refers to the behavior characteristics of a classifier over the operational potential ranges (Krzanowski & Hand, 2009).

The ROC curve is a graphical representation of the false positive rate and true positive rate by plotting these on the horizontal and vertical axis respectively depending on the variation of the threshold t. It actually a complete representation of the performance of a classifier. It is the summarization of information in a single curve using the functions from the cumulative distribution scores of the two classes. Receiving Operating characteristics (ROC) curve has a long history, and many researchers have rediscovered this evaluation metrics in different disciplines. This demonstrates and indicates it naturalness and extreme practicability in representing a classification rule. For references on how ROC curves have been applied in medicine, psychology, theory of electrical signal detections then we must read (Bamber, 1975; Egan, 1975; Green & Swets, 1966; J. A. Hanley, 1989; James A Hanley & McNeil, 1982; Metz, 1978; Peterson, Birdsall, & Fox, 1954; Swets & Pickett, 1982). For more recent studies on machine, data mining, medicine, credit scoring using ROC curves or AUC as evaluation metrics refer to (Ataman & Street, 2005; Bradley, 1997; Fawcett, 2005; Hernandez-Orallo, Ferri, Lachiche, & Flach, 2004; Pepe, 2003; Rosset, 2004; Swets, 1996; Walter, 2005; Y. Zheng & Heagerty, 2004; X. H. Zhou, Obuchowski, & McClish, 2011).

From table three (3) we take a look at all researchers conducted between the years 1964 to 2007 on receiving operating characteristics (ROC) curve and area under curve (AUC) analysis that amounts to 9710 articles. This table does not only show articles using the analysis but also demonstrate its popularity when it comes to comparing models and efficiency of systems applications in various fields. As time goes on the more researchers understand and appreciate the efficiency of this metrics in their analysis looking at the rate at which its popularity grows.

Table 3 ROC/AUC analysis articles between 1964-2007 *Source:* (Krzanowski & Hand, 2009: Chapter 1, p. 15)

Dates (To, From)	**Number of articles**
1964 - Below	2
1967 - 1964	7
1971 - 1968	8
1975 - 1972	9
1979 - 1976	18

1983 - 1980	28
1987 - 1984	41
1991 - 1988	192
1995 - 1992	854
1999 - 1996	1582
2003 - 2000	2506
2007 - 2004	4463

In other to construct ROC curve two metrics needs to be used, that is the sensitivity ratio or true positive rate which is plotted on the vertical axis against false positive rate which is plotted on the horizontal axis. Equation 1.3 has to be recalled while we introduce equation 2.

$$\boldsymbol{SENSITIVITY\ RATIO\ OR\ TRUE\ POSITIVE} = \frac{TP}{TP + FN} \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots EQN.(1.3)$$

$$\boldsymbol{FALSE\ POSITIVES\ RATE} = \frac{FP}{FP + TN} \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots EQN.(2)$$

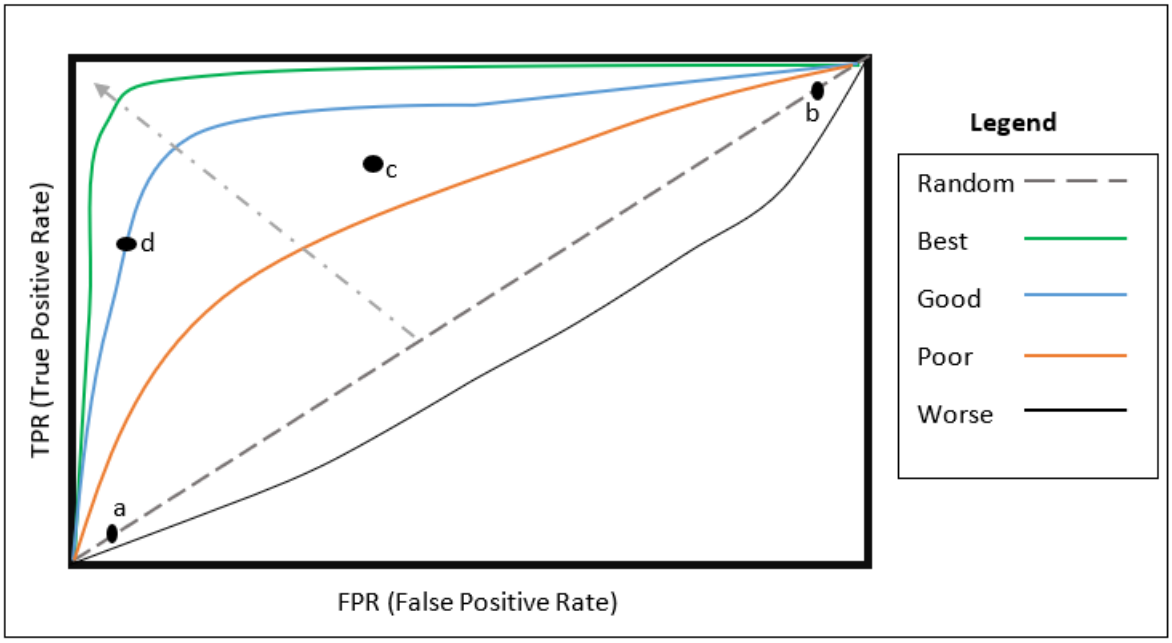


Figure 2 ROC/ AUC (ROC) Curve *Source: Author* *NB: Graph not to scale.*

Figure 2 is a hypothetical graph of ROC/AUC curves, which is not drawn to scale. This is just for illustration of this metrics. From this graph, the "*dash gray lines"* shows the random performance of a classifier with higher false positive than true negative response. The gray arrow pointing to the top left corner or left diagonal is a region or area used to determine the performance of a classifier. The higher it goes the better the performance of a classifier. Point "a" on the optimal line depicts extreme and always would exhibit negative classification. On the other hand, point "b" also depicts extreme positive and always exhibit positive classification. Point "c" on the top right is an area where the dominance of true positive rates with quite a number of errors of false positives. In the extreme scenario, all classifiers that fall within the "c" region would classify all instances as positives. These demonstrate that we would not escape any positive class or member within the process that includes a large number of false positives being included in the positive class. Any classifier that falls within that area towards the right has a worse performance than the random classifier since it exhibits more false positives than true positive results. Point "c" is also called region.

Point "d" any classifier that falls within this area performance faulted with a small number of errors as false positives. In the extreme scenario, all classes would be classified as negatives class or members with no false positives neither would it results in true positives. This area is also called a conservative area. From the legend, a classifier may be defined as best or better when it falls on the path of the "*green line" or* better than that which means almost to *100% (0.1).* The *"orange line"* denotes as poor performance, which indicates that a classifier identified within this threshold is performing poorly. Meanwhile, the *"blue line"* depicts a good performance of a classifier that is better than the *"orange line" or* almost to best performing classifier since there was a little chance of it being able to touch the edges of the graph but optimal when compared to the *"green line".* The *"black line"* denote worse performance of a classification model. This demonstrate that this algorithm or classification model is not accurate and it predicting results is full or errors. ROC curve is not sensitive to skewness. It also gives a full visualization of all models performances, which is more convenient and easy to identify and rank them based on perfect, optimal or sub-optimal.

2.2.2.3.1 Derivation of area under the ROC Curve (AUC)

The area under the ROC curve is a summary index possibly the most widely used for metric for classifying models which were studied in (Bamber, 1975; Bradley, 1997; Green & Swets, 1966; J A Hanley & McNeil, 1983). The area under the ROC curve was

adopted as a measure to compare and the contrast between learning algorithms (Provost & Domingos, 2003; Rakotomamonjy, 2004) while other researchers adopted this metric to construct learning optimization model (David J. Hand & Till, 2001; J. Huang & Ling, 2005; Rosset, 2004). The area under the ROC curve (AUC) value gives a full description of performance ranking of a classifier than probability and threshold metrics. According to David J. Hand & Till (2001) when dealing with binary situation, the value of AUC can be estimated as;

$$AUC = \left[\frac{S_p - \frac{n_p(n_n+1)}{2}}{n_n n_p}\right] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots EQN.(2.1)$$

Where, n_p denoted as the number of positive class members, S_p denoted as the summation of all ranked positive class members and n_n is denoted as the number of negative class members. In practice and in theory, AUC has been proven as a more better metrics when evaluating classification models in terms of performance and identifying the best fit solution during the training stage when compared to accuracy metric (J. Huang & Ling, 2005). Even though AUC is superior to many evaluation metrics and adopted in discriminating against models in binary class situation, in event of multiclass problem, it has a very high cost of computation especially when it is to discriminate and evaluate a bulk-generated solutions. This lead to the formation of time complexities in multiclass problem when adopting AUC to reduce the high computational cost. According to Provost & Domingos (2003) time complexity is $O(|C|n\log n)$ and David J. Hand & Till (2001) also define time complexity as $O(|C|^2\, n\log n)$.

However, with reference to Hossin & Sulaiman (2015), there are other evaluation metrics such as mean square error (MSE), hybrid discriminatory metric and other visualization based metrics. These metrics are not discussed in this work because of it not relevant to this work but it formula can found in the summary table.

Table 4 Summary of formulae *source: Author*

FOCUS OF EVALUATION	USED METRICS	FORMULA
To measure the ratio of correct predicted specify over the sum total of evaluated instances	Accuracy (acc)	$\frac{tp+tn}{tp+fp+tn+fn}$
To measure misclassification error is the total sum ratio of incorrectly predicted instances over the total sum of evaluated instances.	Error Rate (err)	$\frac{fp+fn}{tp+fp+tn+fn}$
This fraction is used to measure correctly classified positive patterns.	Sensitive Ratio (Sn)	$\frac{tp}{tp+fn}$
This fraction is used to measure correctly classified negative patterns.	Specificity (Sp)	$\frac{tn}{tn+fp}$
To measure correctly the predicted positive patterns from the total sum positive class patterns predicted.	Precision (P)	$\frac{tp}{tp+fp}$
Between precision and recall values is the harmonic mean metric.	F-Measure (FM)	$\frac{2*p*Sn}{p+Sn}$
To have relatively balanced rates, *tp* rate and *tn* rate can be maximized from this metric.	Geometric-mean (G-Mean)	$\sqrt{tn*tp}\ OR\ \sqrt{(\frac{tn}{tn+fp})*(\frac{tp}{fn+tp})}$
The effectiveness of averages of all classes.	Average Accuracy (AvAcc)	$\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+tn_i+fp_i+fn_i}}{l}$
This is used to measure the average error rate of all classes.	Average Error Rate (Averr)	$\frac{\sum_{i=1}^{l}\frac{fp_i+fn_i}{tp_i+tn_i+fp_i+fn_i}}{l}$
Per-class average precision.	Average Precision	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}$
Per-class average recall.	Average Recall (Avr)	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}$
Calculating the average of F-Measure per class.	Average F-Measure (AvFM)	$\frac{2*p_M*r_M}{p_M+r_M}$
This metrics is adopted when measuring relativity of random classifiers. It combines both Precision and sensitivity	Kappa Statistics (*K*)	$\frac{P_a-P_e}{1-P_e}$

This metrics is used when measuring misclassification errors and to minimize the error margins to the lowest or small value. This is to enhance accuracy.	Cost-Sensitive Learning (CLS)	$CSL = \left[\frac{fp}{fp+tp+fn+tn}\right] * (C_{G-B}) + \left[\frac{fn}{fp+tp+fn+tn}\right] * (C_{B-G})$
To rank all classifier performance.	Area Under the ROC (AUC) Curve	$AUC = \left[\frac{S_p - \frac{n_p(n_n+1)}{2}}{n_n n_p}\right]$
Reference: C_i is denoted as classes; true positive for class C_i - tp_i; true negative for class C_i-tn_i; false positive for class C_i-fp_i; false negative for class C_i-fn_i; M denoting macro-averages; $P_a = \left(\frac{tn+tp}{N}\right)$; $P_e = \left(\frac{fn+tp}{N}\right) * \left(\frac{fn+tp}{N}\right)$; S_p Sum of all positive instances; n_p number of positive instances; n_n number of negative instances.		

From the previous pages of this work, we can measure the performance of a classification rule upon several ways. It is ideal to choose a particular criterion for which performance measure corresponds to the core performance of the chosen application in question. However, often, it is a challenge to point a particular aspect as a focus, and there is no way an individual would be able to know precisely circumstances upon which a classification rule would be tested or when in the future a classification rule would be verified for validity circumstance. For this reason, it would be of good to have demonstrated several ways through summarizing and displaying the performance of a classification rule over varieties of conditions. This calls for the need to which we have to justify in table 2.2.3 in using the evaluation metrics selected.

2.3 Conclusion

In summary, most finding and solutions provided from works reviewed has a concentration on credit scoring performance. These solutions for performance and efficiency within the credit industry includes both data methods as well as algorithm methods. Problems of imbalance data have been a major concern for credit scoring imperfectness in the credit industry. For this work purpose, we have a higher representation of good credit applicant than bad applicant of finished data sets from applicants' credit approval. We would focus on using K- folds cross-validation and cross validation since it is able to deal with more than one-class learning for training and validation or testing the accuracy of the classifiers selected for a prediction on our dataset.

Metrics for evaluation would be used in the testing phase as an evaluator in measuring the effectiveness of a classifier produced with the unseen data whiles in the training or validation phase it serves as an optimization criterion for classification algorithms. Thus, metric used at this stage is to discriminate and to choose the most favorable solution that can yield most accurate prediction of a specific classifier for future evaluation.

Finally, in this work, we would adapt to the usage of multiple metrics instead of a single metric in analyzing the performance of each algorithm from a broader viewpoint. We would use H or F-measures, Sensitive ratio, Cost sensitive learning, kappa statistics and Area under the curve (AUC in ROC) in the application of this work as evaluation metrics since they contribute differently towards the efficiency of algorithm performance. In evaluating the algorithms, the Sensitive ratio would be adopted to evaluate the algorithm performance of effectively predicting the class of positives. F-measure would be used to monitor the precision and sensitivity ratio simultaneously. The Cost sensitivity learning would be adopted to minimize the misclassification errors to the minimal. The kappa statistics would be adopted to measure relativity of the random classifier. Lastly, the focus of this work which is Area under the curve (ROC) adapted to determine which of the four algorithms selected failed to correctly predict both classes accurately by comparing using the (AUC=0.5) as random guessing.

2.2.3 Justifications of the adaptation of evaluation metrics for these studies adopted

Table 5 Summary for other works and evaluation metrics adopted. *source: Author*

RESEARCHER	YEAR COVERED	ESTIMATION TECHNIQUES	METHODOLOGY ISSUES	FINDINGS	COUNTRY USED
Zhou & Lai (2009)	2009	Area under ROC curve (AUC)	An LS-SVM weighted model for credit scoring using ROC (AUC) Curve maximization was proposed with a direct search optimization.	Their test was centered on two datasets from Germany and Australia. Their results compared with other five most widely used methods shows that WLSSVM model is more stable in terms of performance and is a good option for constructing credit models using AUC maximization.	German, Australian
Brown & Mues (2012)	2012	Area under the receiver operating characteristic curve (AUC)	The AUC was adopted to measure and compare with Nemenyi's post hoc test and Friedman's average rank test to test the difference in ranks and significance of the individual classifiers.	Gradient boosting and random forest classifiers are more stable in terms of performs in the context of credit scoring and comparatively able to deal with class imbalances in data sets. In addition, C4.5 algorithm decision tree, k-nearest neighbors and quadratic discriminant analysis techniques are significantly performing worse than the most widely used classifiers in the case of large class imbalance.	Benelux, Germany, Australian
Verbraken et al. (2014)	2013-2014	Area under the ROC curve (AUC)		In a proposed test for accuracy, selection criteria for model parameter and determining cutoff values using government institution loans granted to customers as data. In their findings, their proposed model for classification based on profit measure performs wells compared to other approaches in terms of monetary value and accuracy in the set test, and this also helps model deployments.	Belgium, Chile, United Kingdom

Ala'raj & Abbod (2015)	2015-2016	Area under the ROC curve (AUC)	This estimation technique was employed in a demonstration of the discrimination and separation ability of each model and to measure their performance and sensitivity in being able to classify good loans as well as bad loans.	the result clearly demonstrates that consensus methods are better compared to MARS, LR, and traditional methods combinations based on all five datasets. In terms of stable RF is good, which is followed by all other datasets, while logistic regression, in spite of goodness, was performing worse compared to classical combinations. Aside from that, MARS comparatively to logistic regression and traditional combination and base classifiers performs well.	United Kingdom
Ala'raj & Abbod (2016)	2016	Area under the ROC curve (AUC)	The AUC was used to determine which of the models used in the work best predict the classes in analyzing in binary classifications.	Their results show clearly that ConsA was the best among other classical combiners and classifiers. In terms of stability RF was good and placing second in all the datasets, and DT was good but comparatively worse than some of the classical combinations. In their conclusion based on their evaluation, LR, PROD, Max, NB and WAVG, SVM, MIN, and NN are worse significantly compared to ConsA approach.	German, Australian, Japanese, Iranian, Polish, Jordanian, UCSD
Blanco et al. (2013)	2013	Area under ROC curve (AUC)	The used in the estimation technique in order to rectify classification problems. The AUC was evaluated using the aid of R using the ROCR Library.	The result of this work depicts that in credit scoring multilayer-perceptron would be best used in Microfinance institution since its ability to perform better in terms of accuracy and lower cost misclassification when compared to QDA, LR and LDA models.	Peruvian
Pompella & Dicanio (2016)	2005-2014	Area under ROC curve (AUC)	AUC was used to measure dissimilarities using column matrix discriminating power as a related indicator.	The findings depict incoherent stances and probably misappropriated rated banks. Their arguments based on their analysis shows that regulators of financial institutions for implementation could adapt their models and methods.	Bloomberg

Koutanaei et al. (2015)	N/A (Two year period)	Area under ROC curve (AUC)	The AUC was adopted for classification and setting parameters in the context of blended credit scoring methods.	In stage three PCA algorithm was identified as the best FS algorithm based on the classification algorithm employed for preparing the dataset from each FS algorithm. Finally, in terms of classification accuracy (ANN) artificial neural network and Adaptive Boosting (AdaBoost), adaptive boosting was higher.	Iran
Leung et al. (2007)		Area under ROC curve (AUC)	This work uses Gini coefficient and ROC curve estimates which were calculated from the AUC as a standardized tool to measure model performance in terms of it prediction accuracy.	From their findings using ROC to compare aritificial intelligence technique through the adaption of natural immune systems, called simpe artificial immune systems (SIAS). SIAS was found to be a better classifer when tested on three different data sets.	Australian, German, Thomas (2002)

III. METHODOLOGY

This section introduces four learning algorithms, which are used for classifying and predicting our data. The four learning algorithms are embodied with three base models as the targets with an ensembler which is adopted as classifications algorithms due to the wide application of these models in the industry of credit scoring (Harris, 2015; Louzada, Ara, & Fernandes, 2016; Tomczak & Zieba, 2015; Xiao, Xiao, & Wang, 2016). These models are logistic regression, Support vector machine, artificial neural networks, and random forestry. These models are very popular in the credit industry that is applied to screening the credit worthiness of applicants. We used this chapter to illustrate all general mathematical formulas behind these algorithms, which are the basis for next chapter or empirical results.

3.1 Logistic Regression

Ohlson (1980) was the first to apply logistic regression in default prediction model classification. In credit risk estimation logistic regression model was first used by Wiginton (in Li et al. 2016: 344). Under the classification models, logistic regression model deals mainly with the binary context in a specific classification modeling with posterior estimates probability of positive class such that sigmoid logistic is a linear function for features vector (Bishop, 2006). On the other hand, logistic regression is classified as a multivariate analysis which is based on single or multiple attributes of independent variables used to predict and analyze the dependent variables attributes, and is mainly used to gain knowledge into the probability relationship of the several conditions of dependent variable and the values of the independent variables (Campus, 2010: V4-736).

According to Bekhet & Eletter (2014: 23) "Logistic regression (LR) is a predictive model widely used in classification". Thomas (in Bekhet & Eletter, 2014: 23) Logistic Regression is a linear function where the variables targeted is a nonlinear function of its probability of being good is higher. He added that independent variables are sensitive to correlations in classifying results for modeling logistic regression. Hence, strongly correlated variables are would not be needed when developing a model.

The estimated model method adopted for logistic regression analysis is the maximum likelihood. The model adopted for logistic model is usually featured as;

$$l_i = L(y_i = 1 \mid X_{1i}, X_{2i}, X_{3i} \dots \dots X_{ki}) = \frac{\exp(\propto + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \dots\dots\dots + \beta_k X_k)}{1+\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 \dots\dots\dots \beta_k X_k)} \quad \dots\dots\dots\dots\dots \text{(a)}$$

$$= \frac{e^{\propto + \sum_{j=1}^{k} \beta_j X_{ji}}}{1 + e^{\propto + \sum_{j=1}^{k} \beta_j X_{ji}}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \text{(b)}$$

Assume "corporate or non-risky credit firms" should be zero (0), then "non-risky credit firm's" ($y_i = 0$) conditional likelihood should be: $P(y_i = 0 | X_1, X_2, \dots \dots . X_{ki}) = 1 - p_i$. Through this we would derive a new observation of likelihood:

$P(y_i) = p_i^{y_i} (1 - p_i)^{1-y_i}$ ..... (c)

Where $y_i = 1$ or $y_i = 0$. This treat all observations as independent, Hence, joint likelihood would:

$L(\theta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$ ..... (d)

Simplification of this calculation would be the logarithmic likelihood function, which is as follows:

$$In[L(\theta)] = In\left[\sum_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}\right]$$

$$= \sum_{i=1}^{n} [y_i In(p_i) + (1 - y_i) In(1 - p_i)]$$

$$= \sum_{i=1}^{n} \left[y_i In\left(\frac{p_i}{1 - p_i}\right) + In(1 - p_i)\right]$$

$$= \sum_{i=1}^{n} \left[y_i(\alpha + \beta X_i) + In\left(1 - \frac{e^{\propto + \sum_{j=1}^{k} \beta_j X_{ji}}}{1 + e^{\propto + \sum_{j=1}^{k} \beta_j X_{ji}}}\right)\right]$$

$$= \sum_{i=1}^{n} \left[y_i(\propto + \beta X_i) - In(1 + e^{\alpha + \sum_{j=1}^{k} \beta_j X_{ji}})\right] y \dots \text{(e)}$$

The functions of the logarithms-likelihood and the estimated maximum likelihood is derived parameters of β_j and α where (j=1, 2, 3,...k). Through these functions, we can then derive the following likelihood:

$$\frac{\partial In[L(\theta)]}{\partial \alpha} = \sum_{i=1}^{n} \left[y_i - \frac{e^{\propto + \sum_{j=1}^{k} \beta_j X_{ji}}}{1 + e^{\propto + \sum_{j=1}^{k} \beta_j X_{ji}}}\right] \dots\dots\dots\dots\dots\dots \text{(f)}$$

$$\frac{\partial In[L(\theta)]}{\partial \beta} = \sum_{i=1}^{n} \left[y_i - \frac{e^{\propto + \sum_{j=1}^{k} \beta_j X_{ji}}}{1 + e^{\propto + \sum_{j=1}^{k} \beta_j X_{ji}}}\right] X_i \dots\dots\dots\dots\dots\dots \text{(g)}$$

The solutions derived from the calculated equations of k+1 is the estimated of β_j where (j=1, 2, 3…k) and α value. This estimation process is the maximum likelihood estimation.

3.2 Artificial Neural Network (Ann)

According to Fensterstock (2001), neural network is the least well-known methods in credit scoring. However, it may be considered as one of the most powerful technology tools in decision support available for solving credit scoring issues. Neural network should not be view as a competitor to the conventional computing or as a replacement but rather it is a complementary technique. It is also identified with its nature of performing complex tasks and do not require programming. NN is also characterized

by its ability to generalize from examples, learn from experiences, extract very important patterns from noisy data and requires less time to develop. Its special feature of dealing with situations, which was not thought of during developmental stages, makes it a very strong tool. Haykin (1994), artificial neural network is the paradigm of processing information which as biological nervous systems as basis including brains in processing information. Fensterstock (2003:12) Neural Networks systems are built on methods that analyze past data and has the capacity to categorize existing or potential accounts into one of the various classes for varying risk credit, such as: indeterminate; good; charge-off; delinquent and bankrupt. Fensterstock (2001), added that this system is a product of artificial intelligence algorithms application which allows the system through analyzing of past data, to identify the relationship between characteristics of accounts and their possibilities of defaults (recognition patterns). It is also developed with the ability to determine automatically the characteristics which are of importance in predicting defaults with more flexibility compared to standardized statistical methods.

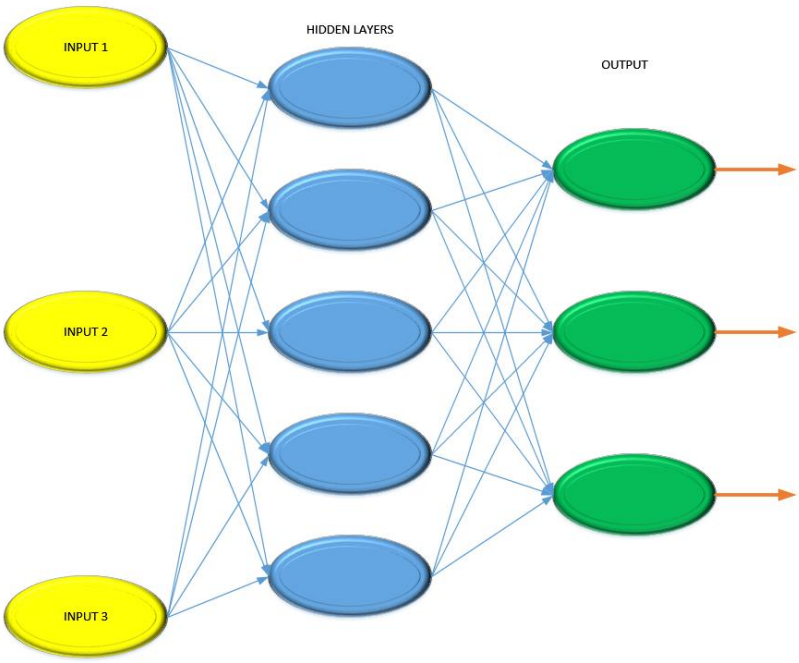
"Artificial neural networks (ANNs) are nonlinear mapping structures based on the function of the human brain" (Lek & Gugan 1999:65) whiles Akko (2012), also explains neural networks as the development of working principles through nonlinear modeling in simulating human brains. This is the transfer of human being's ability to learn, into an environment of computerized systems. According to Mao (1996), artificial neural networks are enormously parallel systems with a tremendous number of interconnected processors. Basheer & Hajmeer (2000) also explains artificial neural networks as the new computing tool that has been extensively using in solving the most complex problems in our real world. Feed-forward is a type of algorithm or model within artificial neural networks family. In credit scoring or risk analysis the most widely used algorithm within artificial neural networks is the feed-forward neural network (Pazhoheshfar & Saberi n.d.; Ala'Raj & Abbod, 2015; Ala'raj & Abbod, 2015; Ala'raj & Abbod, 2016; Oreski & Oreski, 2014; Chen & Lin, 2014; Information & Engineering, 2004; Luo et al., 2016; Zhao et al., 2015; Li et al., 2009; Zhong et al., 2014; Zhang et al., 2007; Bekhet & Eletter, 2014; Jackson & Wood, 2013; Wilson, 1998; Brown & Mues, 2012; Azayite & Achchab, 2016; Zhou et al., 2013; Li et al., 2016; Ghodselahi & Amirmadhi, 2011).

This work is no exception; hence, the feed-forward neural network would be adopted to build this model. The feed-forward neural network is able to identify and map any kind of nonlinear function right from input to output. A standardized feed-forward neural network is characterized by three (3) hierarchical network layers, which are the hidden layers, output layers, and the input layers. The network structure can be illustrated through the following assumptions:

Assumption:

i. Input vector pattern is $X^m = (x_1^m, x_2^m, x_3^m, x_4^m \dots\dots\dots\dots x_n^m)$
ii. Vector output desired is $K^m = (k_1^m, k_2^m, k_3^m, k_4^m \dots\dots\dots\dots k_n^m)$
iii. Hidden layers (Middle layer output) is $V^m = (v_1^m, v_2^m, v_3^m, v_4^m \dots\dots\dots\dots v_n^m)$
iv. Vector output layers is $Y^m = (y_1^m, y_2^m, y_3^m, y_4^m \dots\dots\dots\dots y_n^m)$

Where $m = 1,2,3 \dots\dots n$. The connections weights between the input layers and the middle layer is $(w_{ij}), where\ i = 1,2,3, \dots\dots n; and\ j = 1,2,3, \dots\dots n$; the connections weight between hidden layer to output layer is $(p_{jt}), where\ j = 1,2,3, \dots\dots n; and\ t = 1,2,3, \dots\dots n$; the value of the output threshold of the middle layer of each unit is $(\theta_j), where\ j = 1,2,3, \dots\dots n$; the value of the threshold of output layer of each unit is $(\gamma_t), where\ t = 1,2,3, \dots\dots n$.



3.3 *Figure 3 Feed-forward artificial neural network* *(Li et al. 2016; PP. 346)*

Support Vector Machine

Results from experiments conducted by Huang et al. (2007) depicts that SVM is a promising model and addition to the data mining techniques in existence.

When confronted with the choice of using SVM, you must be prepared to address the following problem that is how to choose the subset of the optimal feature input and the parameters best suit for the kernel (Ping, 2009). SVM is a specific type of algorithm learning system, which is characterized by decision function through a capacity control mechanism, which uses kernel functions, and solution of sparsity (W. Huang, Nakamori, & Wang, 2005). According to Jndel (2010), Support vector machine is a state-of-

the-art recognition pattern algorithms which are well developed through generalization theory and optimization but cannot be applicable to the brain.

According to Shawe-Taylor et al. (1998), minimization risk structure principles are based on the ability of the SVM being able to select data or find traces of data to differentiate and build the best hyperplane optimal which may be the two distinct types of segmentation.

Supposing, we have a set of training M = $[m_1, m_2, m_3, \dots\dots m_v]$ and it corresponding set label N = $[n_1, n_2, n_3, \dots\dots n_v]$, these sample can be expressed as: $[(m_i, n_i)], m_i \in R^d, n_i \in [+1, -1], i \in [1, 2, \dots\dots, v]$. Where "d" represent the number of dimension of input space, and "v" is the size of the sample number. In the conditions for separation liner, hyperplane must be optimal to differentiate the classes into two samples. The formula for hyperplane can be described as follows;

$$(\omega \cdot m_i) + b = 0 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (a)$$

The data would be classified using the below formulas:

$$(\omega \cdot m_i) + b \geq 0, \; n_i = 1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (b)$$

$$(\omega \cdot m_i) + b < 0, \; n_i = -1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (c)$$

Optimal hyper plane normal direction is express as "ω".

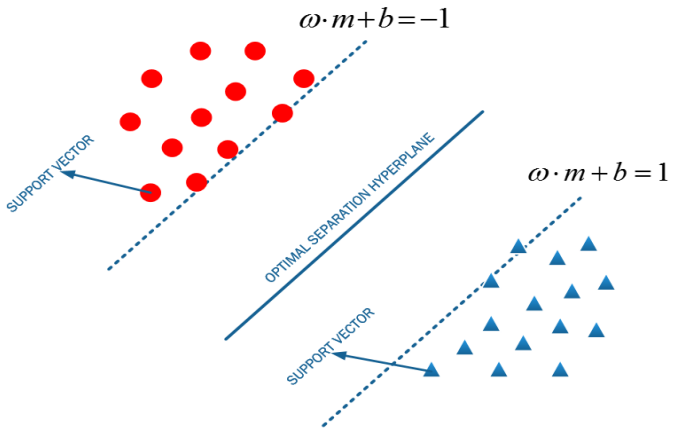


Figure 4 build of optimal separate hyperplane (Zhou et al. 2013; pp. 174)

Hyperplane can be said to be optimal when there are no errors classified in the training data as well as the intervals of the hyperplane data in terms of distance and how furthest are the hyperplane. Alternatively, a situation such as linear separation, hyper plane optimization can be reached as solving programming problem in quadratic. Taking into account the samples training, the bias "b" and optimal weights "ω" are the significant indicators to be identified. The cost function weight minimized can be derived as follows:

$$\min S(\omega) = \frac{1}{2}\|\omega\|^2, \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (d)$$

Where;

$$n_i[(\omega \cdot m_i) + b] - 1 \geq 0, \; i = 1,2,3, \dots\dots, l \dots\dots\dots\dots\dots\dots\dots\dots (e)$$

Lagrange multiplier can be applied to solve the function of optimization $S(\omega)$ must be a quadratic, the condition constraints must be linear and the question must be a typical problem of quadratic programming. Then, LaGrange multiplier can be adopted as a method to solve the problem as follow;

$$L(\omega, b, \alpha) = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{l} \alpha_i [n_i[(\omega \cdot m_i) + b] - 1], \dots\dots\dots\dots\dots\dots\dots\dots (f)$$

Where ($i = 1,2,3, \dots\dots\dots, l$) $0 \leq \alpha_i$ is the lagrange multiplier whereas the value extreme of "L" is saddle point, the optimal solution is the derivation of "L". Hence, the final derivation function is

$$plabel = Sgn(\omega^* \cdot m + b^*), \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (g)$$

Where b^* and ω^* are the optimal solution. Even though, optimal solution is reached, linearly separable of all the data is impossible. Instances, where separable of all the data linearly becomes impossible, alternative options can be used to solve impartibility to achieve linear separation thus, by mapping all the data on a high dimensional space. In the process of mapping the data, kernel functions has to imported into SVM (Cristianini & Shawe-Taylor, 2000). According to (Lyu, 2005; Minh, Niyogi, & Yao, 2006; S. Zhou & Gan, 2004), Kernel functions can be used to calculate samples of the inner product and the kernel function $K(m_i, m_j)$ must satisfy the condition Mercer theorem. The most widely used kernel functions in SVM are sigmoid kernel function, radial basis functions, and the polynomial kernel function. After the added kernel function, we would derive the new formula as;

$$plabel = Sgn\left(\sum_{i=1}^{\mu} \omega_i K(m_i, m) + b\right), \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (h)$$

The number of support vectors and weight of the vector are represented as "μ" and "ω_i" respectively.

3.3.1 Why the adaptation of kernel tricks

This section is an extract from Cristianini & Shawe-Taylor (2000b), under support vector machine which is to introduce the technique or trick of the kernel which is the foundation of support vector machine performance. One astonish feature of the SVM is that the learning theoretic issues are independent of the approximation theoretic issues. Representation of Kernel function is to offer other option of solution through which data can be projected into high feature dimensional space in other to increase the power of computation for machine linear learning.

The introduction of machine linearity in the dual process representation is to make it feasible to perform these implicit steps. Examples from Cristianini & Shawe-Taylor (2000b: Chapter 2), highlighted that all the training examples would never appear in isolation but would always be part of the inner products between pairs of examples. The advantage of adopting the dual process representation of machine learning is derived from the fact that in representing all the number of tunable parameters they are not dependents on the number of variables used. By an introduction of appropriate kernel function to replace the inner product, one can perform implicit nonlinear mapping onto a high feature dimensional space without increasing the number of tunable parameters, so far as the inner product of the vector feature computed by the kernel corresponds to the two inputs (Cristianini & Shawe-Taylor, 2000b: Chapter 2).

One interesting thing about the kernel properties that, it can be studied in general or in a self-intended way towards one trick to be able to apply them to different theories of learning algorithms. According to Hamel (2009:107) Kernel trick is by using the most appropriate function of the kernel that can be adopted as an advantage to map into feature spaces without necessary paying the price of computing the explicit mappings since the input space computations are always simplified by the computations in the feature space. In selecting kernel function for this study, we have so judiciously in other to control the model complexities. To construct an appropriate model for a study or data set, it all lies in finding the right kernel. For the purpose of this study, we would only discuss four kernel functions based upon reviews on their performance and their mathematical expressions. For full details and better understanding of kernel function integration into SVM you should refer (C.-C. Chen & Li, 2014; Cristianini & Shawe-Taylor, 2000a; Hamel, 2009; Noble, 2006; Rakotomamonjy, 2004; S. Ramanna et al., 2013; Schölkopf & Smola, 2002; Suykens & Leuven, 2003).

For an appropriate mapping given the equation $\vartheta : R^v \rightarrow R^u$ with $u \geq v$, then we would have functions as;

$$K(y,m) = \vartheta(y).\vartheta(m) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (i)$$

Where $y, m \in R^v$ are denoted as kernel or kernel functions (Hamel, 2009:107). The steps involved in mapping the datasets from the original source of imputation space to the feature space is identified as kernel trick (Hamel, 2009:103). The kernel tricks is a general approach in it usage than solely limiting its application to support vector machines. This means that all classifiers that merely depends on the products of dot inputting data can adopt this technique. Only four of these kernel tricks would be mention here based upon it wide usage even though there are several of these kernels functions or tricks.

3.3.1.1 Radial Basis Function (RBF) or Gaussian Kernel

The RBF is able to map the input space into infinite dimensional feature space. It is more friendly to use and fits to greatly into various shapes of the border for decision (X. Wang, Tian, & Cheng, 2007). This kernel function or trick is the most widely adopted in most works using the expression;

$$K(y,m) = \exp\left[-\frac{||y-m||^2}{2\sigma^2}\right] or\ \exp[-\mu||y-m||^2] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (j)$$

3.3.1.2 Linear Kernel

This kernel function is linear in function and would not make any changes to support vector machine with existing linear function. Their product would be the same and this kernel is express as:

$$K(y,m) = \vartheta(y).\vartheta(m) = y.m^T \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (k)$$

3.3.1.3 Polynomial of degree "d"

The polynomial kernel has been widely used and adapted for several works including (Cortes & Vapnik, 1995) where they used this kernel function for recognition of handwriting numbers when there were faced with shifting of the optical character recognition. This kernel contributed greatly in identifying the problem for solution. In their recommendations, it is one of the best and most reliable kernel function due to its strength in detecting handwriting. This kernel function can be express as:

$$K(y,m) = (\mu + y.m^T)^d \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (l)$$

3.3.1.4 Sigmoid Kernel

This kernel is also known for its strength within certain datasets and that can be referred from (Dong, 2007; X. Y. Liu et al., 2010). This kernel function can be written in an expression as:

$$K(y,m) = tanh(\mu + (y.m^T) + \delta) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (m)$$

From the above expressions and definitions, $\delta,\ \mu,\ \sigma,\ \vartheta, d$ are all parameters of the kernel.

3.4 Random Forestry (RF)

Random forestry, which was carved from a context of random tree ensembles as inductive conformable predictor having been noted as favorable algorithm performing well across various datasets problems (Bhattacharyya, 2013; Fernández-Delgado, Cernadas, Barro, Amorim, & Amorim Fernández-Delgado, 2014). Its construction is centered on three core steps in building this ensembler for classification. Firstly, the random forest manipulates the training datasets in gaining ensemble diversity. It is through this that bootstrap sampling technique is adopted to create a list of sets for learning. Secondly, an inducer is introduced through the model of random tree through different datasets training to generate a base classifier. Through these steps, small individual groups are generated at each node through the input features being selected randomly. This algorithm usually adopt the value of an integer with the greatest value that should not be greater than $log_2\rho + 1$ even though It has options for the user to predefine the size of the group. We define "ρ" as the number of input features. Through this, the best split point or feature would be picked to split on even

though those trees generated are not pruned. Finally, the algorithm resort to the technique of majority voting through the combination of base classifiers.

According to Breiman (2001:6) who have contributed greatly to the construction and modification of random forest ensemble as a classifier define this algorithm as collection of structured classifiers$\{h(X, \Theta_k), k = 1, \ldots\ldots\}$. Where $\{\Theta_k\}$ is defined as identically independent distributed vectors of random where each vote is casted by each unit of tree for the most popular class at input x. He added that after the generated large numbers of tree, they vote for the most popular class and that is what he term as procedures of random forest. The correlation among the trees and the individual strengths of the decision tree classifiers determines the error rate of the forest (Breiman, 2001). An increase in the error rate is dependent on the increase in the correlation whiles increases in the accuracy of the forest is solely dependent on the increasing the individual strength of the trees. This depicts that if one or more of the input variables are highly significant as predictors, these features would be selected in many of the trees from which they would be correlated. For correlation to be avoided within the trees, random forest adopts a learning algorithmic modified tree which would select randomly features of subsets instead of selecting the whole features to identify the best split for each node. This procedures and features make random forest a unique ensembler from bagging technique as a tree. The random forest is embodied with properties and features as out-of-bag (OOB) error estimates, importance of variables, proximity matrix with full details and its operations can be referred from (Breiman, 1996, 2001; G. Wang et al., 2012).

3.5 Conclusion

This chapter dealt with all the mathematical formulation of each model was adapted for our study. We discuss on logistics regression from a simplified linear model to the derivation of maximum likelihood estimation, artificial neural networks (ANN or NN) as another model for classification which we adapted using the forward feed technique only. In addition is the support vector machine with it kernel tricks for adjustment of the model in making it work better. Only one kernel tricks would be used in the next chapter for our analysis based on the comparison of their results and which one that kernel gives the most optimal solution. In the later part of this study, we introduce the random forest as a combined classifier and ensembler based upon it strong performance within the few years and it robustness among several data set as an alternative to all the three models proposed if it should fail in the testing stage of this work.

IV. EXPERIMENTAL SETUP, DATA ANALYSIS AND DISCUSSIONS OF RESULTS

In this chapter, we discuss the source of our datasets, description, variables selected and the sampling techniques adopted to select the dataset used for this study. this chapter also deals with the sensitive aspect of this works. It's introduce the experimental stages of this work to the final stage of analyzing our data set.

4.1 Experimental Setup

This research would be supported with data mining techniques using Logistic Regression (LR), Artificial Neural Networks (ANN) and Support Vector Machine (SVM) to measure the accuracy of their prediction accuracy on credit defaults using our target population. Random sampling was applied in selecting the dataset for this study. We adopted RStudio version 3.3.2 (2016-10-31) "Sincere Pumpkin Patch" The R Foundation for Statistical Computing, Platform: x86_64-w64-mingw32/x64 (64-bit) as a statistical and data mining tool to mine the datasets selected. This software is installed and run on Lenovo laptop, System Model 20384, Microsoft Windows 10 Home Single Language, Intel(R) Core(TM) i3-4010U CPU @ 1.70GHz, 1700 MHz, 2 Core(s), 4 Logical Processor(s), Installed Physical Memory (RAM) 8.00 GB. Figures 5 and 6 below are our experimental design and detailed illustrated design model respectively for this study.

From Fig. 5 shows a simple experimental design of the process through which we shall select our data sets right from the finished or unprocessed data into a refined stage where we would have our final dataset, which would be imputed into our method selection. From the method of selection, we shall build all classifiers that are logistics regression, support vector machine, and artificial neural networks train them and testing. In the process where these through algorithms fails to predict accurately, we should initialize our processes again from selection method and this time used random forest ensembler as an alternative classification algorithm. Figure 6 gives detailed illustrations of the process through from input of our raw data to finish. Our raw data would go through the cleaning, transformation and feature selection for our analysis. We adopted two selection method or technique for this processes that resulted from two analysis. The first phase shall be through 10 folds cross validation technique where each time our model is run it should be done ten (10) times to ensure efficiency and accuracy of our selected classifiers. The second phase is to segment our data set into 70/15/15 partition where we would use 70% of our dataset for tanning our classifiers and 15% for validation and 15% for testing of the accuracy of the classifiers selected. If all three selected base classifiers should predict accurately and optimally then we proceed to our analysis and conclusion but if it fails to yield the optimal accuracy expected then we shall start the processes again by choosing the alternative combined model or ensembler known to be performing extremely well over the years that is random forest ensembler. In both analyses, random forest is an option if all selected algorithms fail to yield the best optimal solutions.

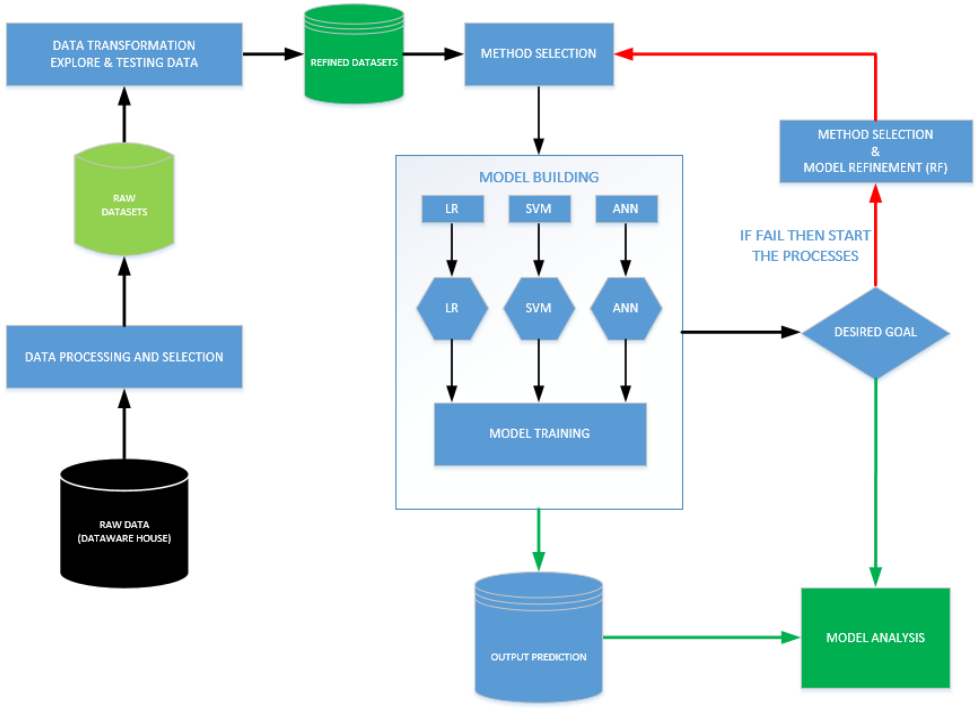


Figure 5 Experimental Design *Source: Author*

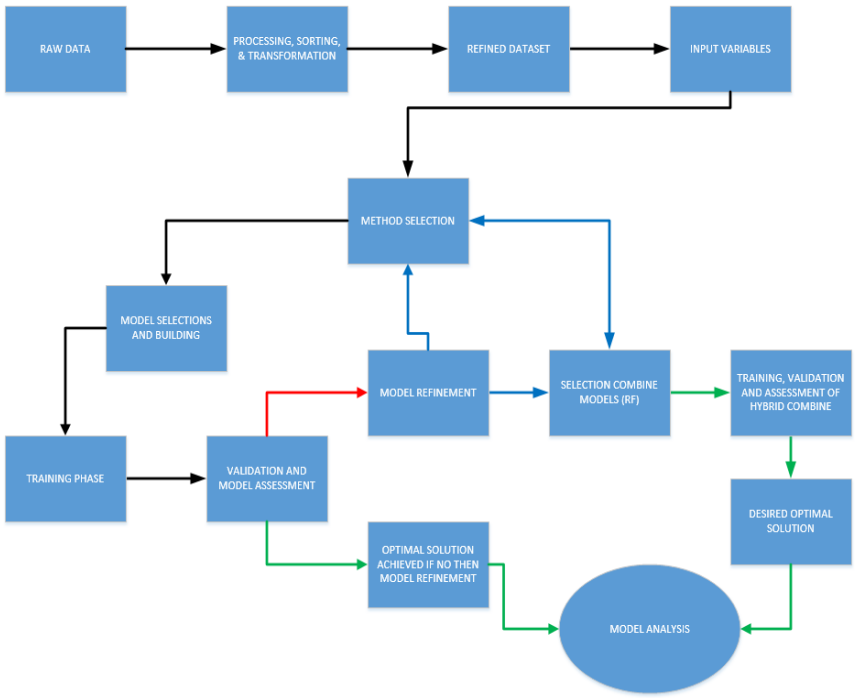


Figure 6 Detailed illustration of the experimental design *Source: Author*

4.1.1 Dataset Acquisition And Description

Our data set is a secondary data from borrowers' lending's between 2009 to 2015 fiscal year that is collected from Commercial Banks XYX in Ghana. This dataset contains 66 variables with 44747 instances for both non-defaulters and defaulters. Sorting our data set into zero (0) overdue date account for 29423 instances and non-zero overdue data accounts for 15324 instances. We selected at random 15000 instances from the total instances of 44747 which constitute 8700 to be non-defaulters and 6300 to be defaulters. In other to ensure and enforce accuracy in predicting defaulters accurately with reference to Ala'raj & Abbod (2015:103) "the better the accuracy of the classifier, the more impact it has on others". We mean that a classifier may be performing better or would be recorded as having high accuracy rate does not mean other classifiers are poor because they would be close to performing better. For this study concern, we used German and Australian data sets from UC Irvine Machine Learning Repository (UCI) (https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/) as a benchmark to enforce effectiveness and efficiency results for accuracy. From table 6 depicts all characteristics of our datasets used for this study.

Table 6 Datasets descriptions for studies

Datasets	Attributes	Total Instances	Non-Defaults	Defaulter	Instances Selected	Missing Values
GHANA	66	44747	29423	15324	15000	YES
GERMAN	20	1000	700	300	1000	N/A
AUSTRALIAN	14	690	307	383	690	YES

All variables within these three datasets have been clearly tabulated within the next pages.

Table 9 shows all the variables in our data set from Ghana that is used for this study. The following characteristics are found within this dataset: (a) Three (3) data type: Category, Numerical and Constant; (b) Sixty Six (66) variables or attributes; (c) Twenty Five (25) Numerical values; (d) Forty-One (41) Categorical values.

Table 7 German dataset variables description

VARIABLES	DATE TYPE		VARIABLES	DATA TYPE
Status of Existing Checking Account	Categorical		Age	Numerical
Duration in Month	Numerical		Housing	Categorical
Credit History	Categorical		Job	Categorical
Loan Purpose	Categorical		Telephone	Categorical
Credit Amount	Numerical		Property	Categorical
Savings Accounts / Bonds	Categorical		Foreign Worker	Categorical
Present employment since	Categorical		Number of Dependents	Numerical
Installment Rate (%) of Disposal Income	Numerical		Present Residence Since	Numerical
Number of Existing Credits at this Bank	Numerical		Other Debtors / Guarantors	Categorical
Other Instalment Plans	Categorical		Personal Status and Sex	Categorical

Table 8 shows all variables descriptions and data type within the German data set that was collected from UC Irvine Machine Learning Repository (UCI) (https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/) as one of the bench marking the Ghanaian dataset. This dataset has the following characteristics: (a) Two (2) data type: Category and Numerical; (b) Twenty (20) variables or attributes; (c) Seven (7) Numerical values; (d) Thirteen (13) Categorical values.

Table 8 AUSTRIALIAN DATASET VARIABLE DESCIPTION

VARIABLES	DATA TYPE		VARIABLES	DATA TYPE
A1	Categorical		**A8**	Categorical
A2	Numerical		**A9**	Categorical
A3	Numerical		**A10**	Numerical
A4	Categorical		**A11**	Categorical
A5	Categorical		**A12**	Categorical
A6	Categorical		**A13**	Numerical
A7	Numerical		**A14**	Numerical

Table 8 shows all variables descriptions and data type within the German data set that was collected from UC Irvine Machine Learning Repository (UCI) (https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/) as one of the bench marking the Ghanaian data set. This dataset has the following characteristics: (a) Two (2) data type: Category and Numerical; (b) Fourteen (14) variables or attributes; (c) Six (6) Numerical values; (d) Eight (8) Categorical values.

Table 9 Ghanaian dataset variable description

VARIABLES	DATA TYPE	VARIABLES	DATA TYPE	VARIABLES	DATA TYPE
Customer Code	Numerical	Age	Numerical	Open Market Value	Numerical
Loan Code	Categorical	Sex	Categorical	Other Deductions	Numerical
Date of Birth	Categorical	Contract	Categorical	Tax	Numerical
Customer Telephone/Fax	Categorical	Nationality	Categorical	Other Income	Numerical
Customer Occupation	Categorical	Net Income	Numerical	Pension	Numerical
Customer Residence Status	Categorical	Date Added	Constant	Number of Children	Numerical
Customer Marital Status	Categorical	Date Modified	Constant	Salary Income	Numerical
Customer Identification Card Type	Categorical	Post City	Categorical	Industry Category	Categorical
Gross Monthly Income	Numerical	Astatus	Categorical	Date Added	Constant
Number of Dependents	Numerical	Date Applied	Categorical	Outstanding Principal Bal.	Categorical
Current Credit Grade	Categorical	Loan Purpose	Categorical	Default Flag	Categorical
Months (Duration) Current Employer	Numerical	Payment Type	Categorical	M_Totnet_IIS_YTD	Numerical
Employment City	Categorical	DelinStatus	Categorical	Current Arrears Status	Numerical
CustomerSegDsc	Categorical	LastCrTxndt	Categorical	Last Amount Paid Date	Categorical
Agreement Number	Numerical	LastDrTxndt	Categorical	InsurnanceYN	Constance
Applied Received Date	Categorical	Next Due Date	Categorical	Last Amount Paid	Numerical
Loan Approved Date	Categorical	Default Start Date	Categorical	Total Instalments	Numerical
Loan Account Open Date	Categorical	M_Eligible Amount	Numerical	Current Total Tenor	Numerical
Disbursement Date	Categorical	M_Instalment Amount	Numerical	Repayment Frequency	Categorical
Maturity Date	Categorical	Original Loan Amount	Numerical	Original Tenor	Numerical
Instalment Start Date	Categorical	Overdue Days	Numerical	Highest Arrears Status	Categorical
IntCrBureauStatus	Categorical	Instalment Remaining	Numerical		

4.1.2 Data Transformation and Cleaning

This is the stage where we would be refining and defining our datasets for learning by our models selected for training and testing or validation of the performance in terms of predictive powers.

4.1.2.1 Feature Selection

From our summary of our datasets, we shall adopt feature selection as a mode to select only significant attributes or variables for this study. All other variables as outliers within the datasets would be ignored for this study since they would not affect our results.

4.1.2.2 Defining of variables for Ghanaian dataset

All selected variables or attributes are defined in the table below. This is one of the most important stage of our data set where we are cleaning and defining the variables by replacing them with dummy variables or integers for each classification. This stage prepares our datasets for learning by our selected algorithms. Our data set fitness is dependents on how well we prepare these datasets. The other two datasets (German and Australian) used to bench mark our analysis has been refined and transformed for analysis. This can be referred from table 17 at appendix

4.1.2.3 Missing Values

In most real world datasets, we are faced with the problem of missing values. This problem could be as a result of either personnel in charge in the collection processes did not collect the data from the beginning or started the collection process and breaks or some of the information requested from the applicants could not available. We have several ways of dealing with this problem since its treatment could either influence our results positively or negatively or cause biasness towards one class even though this case rarely occurs. Situations where missing values would affect a big portion of the total observations within the data sets, these values cannot be ignored or disregarded.

Practitioners in general, treat category variable differently from numerical values. In most cases, new category label "N/A" is assigned to the missing values or adapting the mode value as a replacement to the missing values for the category. On the other

hand, numerical missing values are replaced by the median value or mean value. In addition, another option for treating the missing value in the dataset is to assign a code "@" in a situation where there are text and "666667" where it requires numerical variables. This would give room to the algorithms to decide whether those code or numerical value assigned are of important or not.

This approach may look quite discretional at first glance, it is logical from the point of view practitioners. Majority of the learning algorithms is unable to treat or work with data sets containing missing values. In other to address this issue of missing values in our observations when using learning algorithms then we must use the assumption of replacing them with the most probably occurred values. In real world application and interpretation of the usage of this technique is; if we cannot obtain some information from the applicant, then we must use the default value.

4.1.2.4 Treating Large Values

All large values within our datasets are treated as natural logarithms values due to skewness problem. This is a common practice and acceptable in the field of research.

4.2 Data Analysis and Discussions of results

Our results would be presented in two phases where phase one uses 10 folds cross validation methods and phase two uses data segmentation where we have 70% of the total datasets being used for training of the model and 15% for validation and 15% for testing of the accuracy of our selected models. We only reported on the testing report since that was our target. There were chances where would have gotten good results from our validation results since that stage is seen by our algorithms compared to the testing sets which was entirely new for the algorithms to predict.

4.2.1 Demography of Ghana Dataset

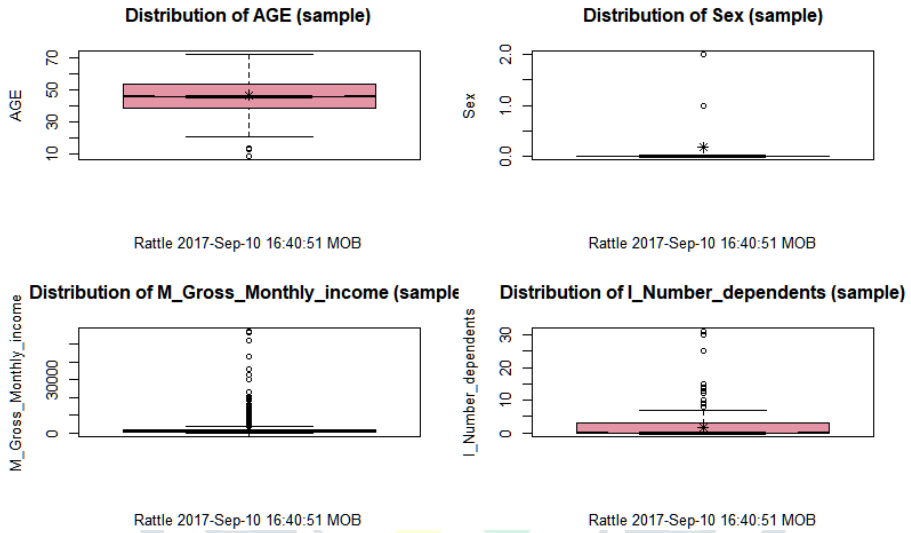


Figure 7 Box Plot variables that are significant

From fig.7 and 8 above we have interesting distributions of our datasets using the box plots. We can see that Age is significant in terms of relations with people who default in term of loans. Majority of the people applying for loans are between the ages of 40year to 50years which can be inferred from fig.9 of the histogram plotting of the age distribution that has an influence on the group of people who are likely to default. Sex box plot distribution depicts that sex has an influence on those who are likely to default when given loans and from fig.9 below demonstrate that majority of the applicants going in for loans and are likely to defaults are in the range of Zero (0) which are males compared to females.

From fig.7 of the box plotting of gross income has the influence on those who are likely to defaults if given loans. From the box plot, gross income is skewed towards zero with reference to fig.9. Gross monthly income of applicants whose salary are less than or close to 0. This demonstrates that majority of the applicant whose gross income is less than 10000 have the maximum likelihood of defaulting on loans are high. From fig.7 of the box plot distribution of number of dependents has an influence on the likelihood of loan applicants being able to pay their or default. The box plotting depicts that those who with dependents close to zero (0) have lower chances of defaulting than those above zero (0) which can be referred from fig.9.

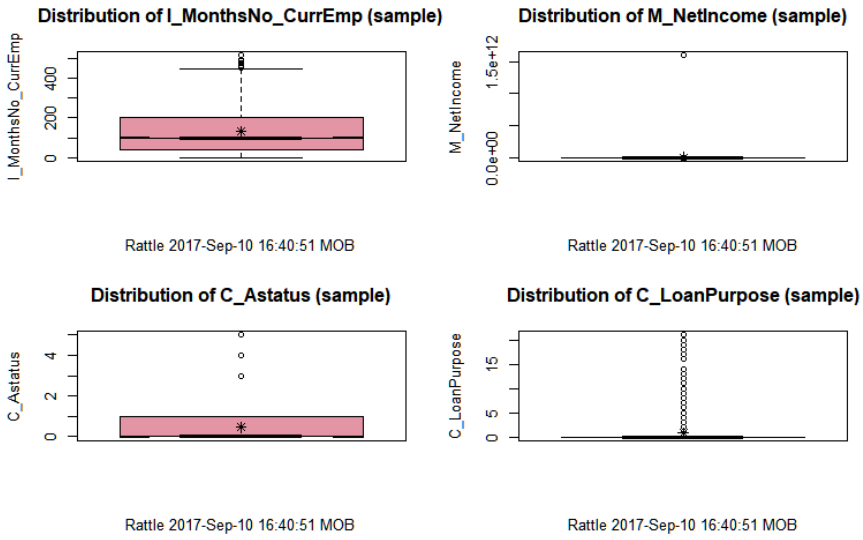


Figure 8 Cont'd of Box Plot variables that are significant

From fig.8 of the box plots distribution depicts numbers of years an applicant have stayed on their current jobs. From the box plots, it shows that the longer people stayed in their jobs the chance of them getting closer to their retiring age might be higher than those employed and younger. The inference from fig.10 shows the histogram distribution of applicants requesting for loans. It is clear that majority of the applicants requesting for loans are below 100 months of employment with a higher likelihood of repayments compared to those above 100.

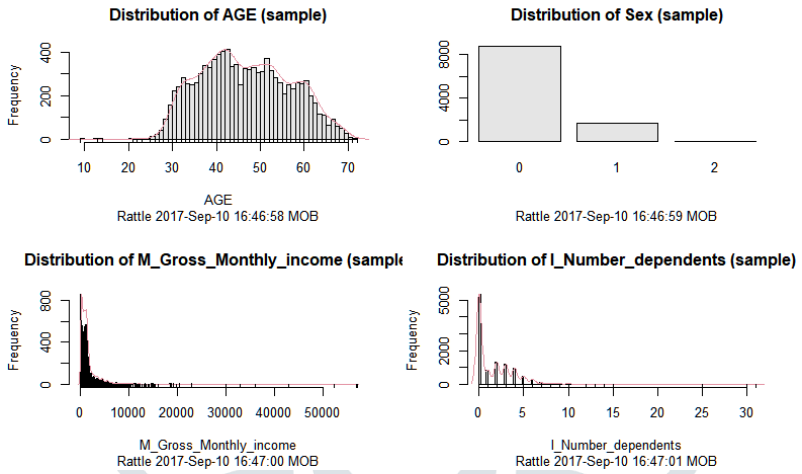


Figure 9 Histogram Distribution of Box Plot graphs

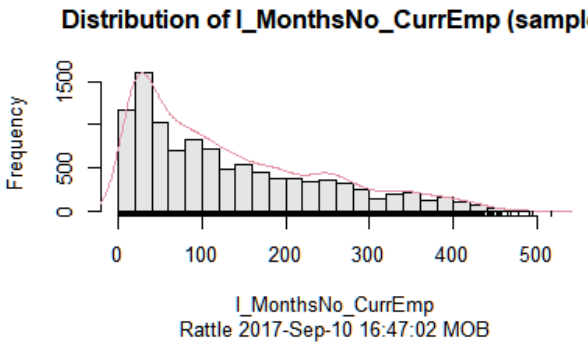


Figure 10 Cont'd Histogram Distribution of Box Plot graphs

4.2.2 Test for Statistical Significance

According to Vicente, Marqués, & Sánchez (2014), it is not evidenced enough to prove that classification algorithm or model achieves better performance or results than others, due to the testing results of different measuring performance or through the data mining technique of splitting the data set. For a complete evaluation performance, it would be right to propose testing of hypothesis to ensure that all difference in experimental test in terms of performance are significant statistically and not to be attributed to any

random effects of splitting data sets. How to choose the appropriate test for specific experiments would be based on several factors, which includes the number of datasets and number of classification algorithms to be compared for that particular study.

From Demšar (2006), stated that in machine learning statistical test can either be a non-parametric with examples such as Kolmogorov-Smirnov, Wilcoxon Signed Rank, Friedman test, and Wilcoxon Rank-Sum and parametric such as paired F-test and T-test. Suggestion from Demšar (2006) noted that non-parametric test should be considered than parametric tests since they can be unsafe statistical and inappropriate conceptually. It is safer and appropriate to adapt non-parametric test approach in machine learning compared to parametric approach since they do not presume homogeneity of variance or data normality (Demšar, 2006). In accordance to the above-stated points, we adapt the Wilcoxon rank sum test which is also known as Mann-Whitney test for this study to compare all the performance rankings for the four (4) classification models measured across all the three datasets.

This is a two-sample nonparametric test. The null hypothesis states that there is no shift in the distribution of the samples and alternative hypothesis states that there is a shift in the distribution of the samples or locations. This test does not consider if the two samples are distributed normally but take into account whether their distribution is of the same shape. For every test, the null hypothesis would be rejected if the p-value is less than 0.05 and accept the alternative with a confidence level of 95%.

Hypothesis 1 "AGE"

p-value < 2.2e-16

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 2 "SEX"

p-value = 0.008454

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 3 "GROSS MONTHLY INCOME"

p-value < 2.2e-16

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 4 "NUMBER OF DEPENDENT"

p-value = 0.0003289

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 5 "MONTHS OF EMPLOYMET"

p-value = 1.026e-11

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 6 "NET INCOME"

p-value < 2.2e-16

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 7 "CUSTOMER ACCOUNT STATUS"

p-value < 2.2e-16

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 8 "LOAN PURPOSE"

p-value < 2.2e-16

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 9 "LAST AMOUNT PAID"

p-value < 2.2e-16

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 10 "ELIGLE AMOUNT"

p-value < 2.2e-16

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 11 "MONTHLY INSTALL AMOUNT"

p-value < 2.2e-16

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 12 "ORIGINAL LOAN AMOUNT"

p-value < 2.2e-16

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 13 "OVER DUE DAYS"

p-value < 2.2e-16

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 14 "RESIDENCIAL STATUS"

p-value = 0.008454

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

Hypothesis 2 "REST OF VARIABLES"

p-value < 2.2e-16

H_0: There are no shift in distribution of the variable : Reject null hypothesis if $p < 0.05$

H_1: There are shift in the distribution of the variable : Accept if $p < 0.05$ with confidence of 0.95.

4.2.3 Predictive Modelling (Probability of Default Model)

Table 10 PENALISED DATASET FOR PROBABILITY TESTING

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-2.09E-02	2.16E-02	-0.97	0.332209	
AGE	9.40E-03	5.56E-04	16.905	< 2e-16	***
Sex	-1.86E-02	8.78E-03	-2.115	0.034431	*
M_Gross_Monthly_income	-4.18E-05	4.47E-06	-9.354	< 2e-16	***
I_Number_dependents	-6.96E-03	1.73E-03	-4.024	5.76E-05	***
I_MonthsNo_CurrEmp	-5.56E-04	4.27E-05	-13.008	< 2e-16	***
M_NetIncome	-1.79E-13	1.18E-13	-1.512	0.130529	
C_Astatus	1.22E-01	5.87E-03	20.795	< 2e-16	***
C_LoanPurpose	6.42E-04	1.59E-03	0.403	0.687011	
M_LastAmtPaid	-1.40E-05	1.55E-06	-9.027	< 2e-16	***
M_EligibleAmt	2.33E-07	4.04E-08	5.775	7.94E-09	***
M_InstallAmt	-2.51E-05	1.20E-05	-2.102	0.03561	*
M_OrigLoanAmt	5.18E-06	6.37E-07	8.123	5.07E-16	***
I_OverdueDays	1.34E-03	2.48E-05	53.856	< 2e-16	***
I_OrigTenor	-4.48E-04	4.60E-04	-0.973	0.330494	
I_CurrTotalTenor	1.25E-02	6.60E-04	18.889	< 2e-16	***
I_TotInstallments	-1.22E-02	4.56E-04	-26.636	< 2e-16	***

M_OutStandPrincipalBal_base	-5.95E-06	6.49E-07	-9.163	< 2e-16	***
M_NoOfChildren	-3.84E-03	2.43E-03	-1.579	0.114305	
M_Salary_Income	2.23E-06	2.82E-06	0.792	0.42834	
M_Other_Income	-1.34E-05	7.72E-06	-1.736	0.08262	.
M_Pension	-3.31E-04	6.78E-05	-4.887	1.04E-06	***
M_Tax	1.09E-04	2.18E-05	5.004	5.71E-07	***
M_other_Deductions	1.46E-05	1.34E-05	1.089	0.276135	
IMN_C_Occupation	-2.47E-03	6.41E-04	-3.844	0.000122	***
IMN_C_Residence_Status	-1.50E-02	3.17E-03	-4.731	2.27E-06	***
IMN_C_Marital_0tatus	-1.14E-02	2.92E-03	-3.891	0.0001	***
IMN_C_Nationality	1.69E-02	6.27E-03	2.698	0.006991	**
IMN_C_PaymentType	1.07E-02	6.12E-03	1.743	0.081295	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.327 on 10471 degrees of freedom
Multiple R-squared: 0.562, Adjusted R-squared: 0.5608
F-statistic: 479.8 on 28 and 10471 DF, p-value: < 2.2e-16

From tab.10 above and the significant codes expressed below, '***' is 99.9% significant, '**' is 99% significant, '*' is 95% significant and '.' is -95% significant. The closer the adjusted R-squared to 1 the better the accuracy of the model. The residual standard error must be close to zero (0) shows strong relationships of these variables. From this, we can now deduce our significant variables from the above table as (1) Age was positively correlated to default, implying that as people gross older their responsibility increases through family ties and as well as nearness to retirements. The chance of defaulting is higher compared to a young age. (2) Sex was negatively correlated to default which implies that most of the applicant was male due from our definition of data set. The likelihood of males also defaulting is higher than men. (3) Gross monthly income was negatively related to default meaning that the lesser your income the likelihood of an applicant defaulting is higher. (4) Number of dependents is also negatively correlated which demonstrates that those not defaulting were having fewer or no dependents at all.

This enables such applicants to meet their loan agreement terms compared to those with dependents. (5) Months of current employments are negatively related to defaulting which implies that the lesser you have to work the higher tendency of defaulting on a loan is high. (6) customer account status was positively related to default which depicts that the type of account an applicant is used in determining whether he or she would default on the loan or not. (6) last amount paid with respect to if you have ever received loan facility from the bank. (7) Eligible amount indicates the amount an applicant is able to access. (8) Monthly installment amount is significant in determining if an applicant would default or not. (9) Original loan amount is the amount given when after agreement by the lender to offer the loan to the applicant. (10) The current total tenor is the time duration used to settle the loan if given. (11) Total installments is the number of times payment would be made to service the loan before the agreed time expires (12) Outstanding principal balance if you have ever taken loans. (13) Monthly pension deduction, (14) tax deductions, (15) occupation, (16) Residential status, (17) Marital status, and (18) Nationality.

The probability model of defaults can be expressed as either a simplified linear model or a recall from chapter (3.2);

$$PM = \theta + \beta_1\gamma_1 + \beta_2\gamma_2 + \beta_3\gamma_3 + \beta_4\gamma_4 \dots\dots\dots\dots\dots\dots\dots\dots \beta_n\gamma_n \dots\dots\dots\dots\dots enq.(3)$$

Where $Where\ ; PM\ is\ predictive\ modelling, \theta\ is\ the\ intercept, \beta_i\ estimated\ values\ of\ the\ each\ variable, \gamma_i\ is\ the\ value\ of\ each\ of\ the\ variable.$

OR

$$RECALLL \dots\dots\dots\dots\dots\dots \frac{\partial In[L(\theta)]}{\partial \beta} = \sum_{i=1}^{n}\left[y_i - \frac{e^{\propto + \sum_{j=1}^{k}\beta_j X_{ji}}}{1 + e^{\propto + \sum_{j=1}^{k}\beta_j X_{ji}}}\right]X_i \dots\dots\dots\dots\dots\dots\dots (g)$$

Recall expression deal with individual variable but for simplicity purposes and novice readers of this work, we would adopt Eqn. 3 modeling where we express the following parameters as;

$$\begin{aligned} PM = {} & (-2.09\text{E}-02) + (9.40\text{E}-03) * (Age) + (-1.86\text{E}-02) * (\text{Sex}) + (-4.18\text{E}-05) * (\text{Gross}_{\text{MNInco}}) + (-6.96\text{E}-03) \\ & * (\text{NUM}_{\text{dependent}}) + (-5.56\text{E}-04) * (\text{MN}_{\text{curEMP}}) + (1.22\text{E}-01) * (\text{C}_{\text{Astatus}}) + (-1.40\text{E}-05) \\ & * (\text{M}_{\text{LastAmtPaid}}) + (2.33\text{E}-07) * (\text{M}_{\text{EligibleAmt}}) + (-2.51\text{E}-05) * (MN_{InstalAmt}) + (5.18\text{E}-06) \\ & * (\text{Orig}_{\text{LoanAmt}}) + (1.25\text{E}-02) * (\text{Curr}_{\text{TotTenor}}) + (-1.22\text{E}-02) * (\text{Total}_{\text{Instalm}}) + (-5.95\text{E}-06) \\ & * (\text{Outstand}_{\text{PrincialBal}}) + (-3.31\text{E}-04) * (\text{M}_{\text{Pension}}) + (1.09\text{E}-04) * (\text{M}_{\text{Tax}}) + (-2.47\text{E}-03) \\ & * (\text{Cus}_{\text{Occupation}}) + (-1.50\text{E}-02) * (\text{CUS}_{\text{Residential}}) + (-1.14\text{E}-02) * (\text{CUS}_{\text{MartialStatus}}) + (1.69\text{E}-02) \\ & * (\text{CUS}_{\text{nationality}}) \dots\dots\dots\dots \text{enq.}(4) \end{aligned}$$

From the above, a predictive model or probability model of a customer not defaulting should be less than one (1) and default should be equal to one (1) or above. Mathematically, the expression can be written as;

$PM < 1 :\ then\ the\ probability\ of\ customer\ not\ defauling\ on\ loans\ are\ high.$

$PM => 1 :\ then\ the\ probability\ of\ customer\ defauling\ on\ loans\ are\ high.$

4.2.4 Phase One: 10 Folds Cross-Validation Technique

Error Matrix

In other to examine the prediction accuracy averagely by all three single classifiers and the fourth which is an ensembler, we adopted the type II and type I errors, which was discussed extensively in chapter two of this work using **confusion matrix.**

Type II Error (False negatives): This gives errors predicted by incorrectly presenting good credit applicants as bad applicants. This type of error is risk free.

Type I Error (False Positives): It demonstrates the prediction rate of errors of chosen model, by classifying incorrectly bad credit applicants as good even though this would cause severe replication on the bank or financial institution.

From Table 11 below demonstrates misclassification errors for 10 cross-validations which access the performance of each model based on the rate of misclassifying an applicant in terms of defaults.

Our focus would be on the false positive rate since that would lead to severe replication on the bank activities. Type I error must be lowest as possible to reduce all associated risks since it occurrence would lead to bankruptcy and fold-up of the banks compared to false negative which could save the bank by denying applicant loans. A bank who offer loans to potential defaulter would suffer from losing their principal or/and their interest on the invested principal either in the process or after the closure of the loan. In addition, the bank would lose additional cost such as legal fees or administration fees, forceful possession costs, maintenance fees and secured disposing of assets and agreements based lowering or waiving some of the interest or principal leads to increase of type I error on behalf of the lender (Nayak & Turvey, 1997).

Comparatively, even though in the Ghana dataset we have average accuracy of the Logistic regression (LR) and support vector machine (SVM) performing quite well in terms average accuracy with 94%, logistic regression has the lower type I error with 3% with an average of error of 6% which is very impressive and could save the bank a lot of money compared to the two models. The random forest (RF) ensembler performance very well in the Ghanaian data sets. On the other hand, artificial neural networks had the highest false negatives with 14% and an average error of 17%, this means in actual we would have good borrow who would have paid their loans even though they are classified as negative or potential defaulters. This, in reality, would save the bank lot of money since they are going to reject good credit worthy applicant's access to loans. This could lead to downsizing of their customers and may lead to severe loss of investments.

Table 11 MISCLASSIFICATIONS ERRORS (10 Folds Cross-Validation)

MODELS	TYPE II ERROR	TYPE I ERROR	OVERALL ERROR	AVE. ERROR	ACCURACY
GHANA					
LR	430 (0.04)	**354 (0.03)**	7%	6%	**94% (0.94)**
SVM	27 (0.00)	741 (0.06)	6%	8%	94 % (0.94)
ANN	1678 (0.14)	517 (0.04)	18%	17%	82% (0.82)
RF	**0 (0.000)**	**0 (0.00)**	**0**	**0**	**100% (1)**
GERMAN					
LR	126 (0.16)	120 (0.15)	31%	36%	69% (0.69)
SVM	53 (0.07)	184 (0.23)	30%	44%	70% (0.70)
ANN	137 (0.17)	**110 (0.14)**	31%	36%	**69% (0.69)**
RF	74 (0.09)	162 (0.20)	29%	41%	70% (0.70)
AUSTRALIAN					
LR	20 (0.04)	79 (0.14)	18%	20%	82% (0.82)
SVM	29 (0.05)	**73 (0.013)**	18%	20%	**82% (0.82)**
ANN	5 (0.01)	192 (0.35)	36%	40%	64% (0.64)
RF	20 (0.04)	79 (0.14)	18%	20%	82% (0.82)

From above, the German credit data set artificial neural network had the lowest type I error with 14% and an average error of 36%. Even though it performs sub-optimal in terms of accuracy, it would save the bank more than of the other single classifiers including Random Forest (RF) as an ensembler and hybrid model performed poorly compared to the other single classifiers. In the Australian data set support vector machine performed extremely in terms of type error with 13% and average accuracy of 20% with 82% level of predicting accuracy. LR and FR were second to SVM with type I error 14% and average accuracy of 82%.

To conclude, the above demonstrates that Logistic regression performs much better in terms of Ghana datasets when comparing single classifiers and hybrid models or ensembler could perform much better by reducing the risks in terms of cost sensitivity issues or losses that the bank would suffer from in they should offer some loans. ANN performs much better than all the single classifiers including RF in the case of Australian data sets whiles SVM performance better in the Australian data set.

Risk Curves

Table 12 Risk and Recall Curves (10 Folds Cross-Validation)

MODELS	ACCURACY
GHANA	
LR	95% (0.951)
SVM	**98% (0.980)**
ANN	87% (0.874)
RF	100%
GERMAN	

LR	**75% (0.746)**
SVM	74% (0.738)
ANN	73% (0.728)
RF	74% (0.736)
AUSTRALIAN	
LR	**93% (0.932)**
SVM	93% (0.928)
ANN	72% (0.723)
RF	93% (0.929)

The risk curves (Precision) and recall curves are two-dimensional graph plotting where financial rewards and financial risks as variables are plotted along the horizontal axis and vertical axis respectively. Deductions from these curves help to identify if the banks would be able to avoid risk and gain all their rewards. It is through this that risk is ranks as either disastrous risk, bearable risk or critical risk. According to Williams (2008), Risk curves or Charts is also known as cumulative gain charts. Adaptation of risk charts or curves to measure models or algorithms of non-compliance of loans and fraud cases gives more advantages over any other situation where the application is to predict rain or for any other matter. From the figure 11 below, it is represented with the accuracy or recall curves, which is represented by a green line and the strike line blue, which is represented by the blue line. From the strike line, we have 42% being identified for each of the algorithms which imply that for every caseload 42% of the population would have a higher risk of not paying their loans or the banks would not be able to recover their loans. The target is to avoid issues of default and non-complaisance of the loans then we must we must first consider applicants.

From table 12 above is not to identify errors with the algorithms but to able to avoid risk. In the Ghanaian dataset comparing the single classifiers, we identify SVM to have higher accuracy with 98%, LR 95%, ANN 87% and RF 100% that is also represented below in fig.7 below using the green line. This indicates that in each caseload 98%, 95%, 87%, and RF 100% would need to audit very well else this would lead to default scenarios and banks would not be able to recover their loans. It also implies that proper attention must be given to each instance in other to avoid risk and non-compliance of the loan terms. From table 12 in the case of Australian data sets, we had 45% has a strike line representing the risk scores from our diagram in the appendix. For every situation or caseloads, we would have 45% of the applicant who would not comply with the loan terms and would lead to default issues or applicant with scores higher than 45% must be audit properly. In the presentation of the Green line which is also found in table 12 demonstrate that 93% for LR, SVM, and RF requires audit properly and 72% for ANN would need a proper audit.

From the diagram in the appendix, 30% as a strike line was identified by the German data set as risk scores. In every situation of scoring applicants, as well as 30% of those caseloads, have higher chances of defaulting the loans as well as those with higher scores. Comparing all our single classifiers, we had LR, SVM, and RF with accuracy scores of 75% for LR and 74% for both SVM and RF which indicates that 75% and 74% of the total applicant applying for loans must be an audit and needs attention else the banks would result in severe risk cases.

Conclusion on the above would mean that in a scenario where we pick 10 applicants representing 100% we would have four applicants representing 42% of risk scores for Ghana data set defaulting on their loans. Three applicants representing 30% in the German situation also defaulting on their loans out of the ten applicants being used as caseloads as well as four applicants in the case of the Australian dataset.

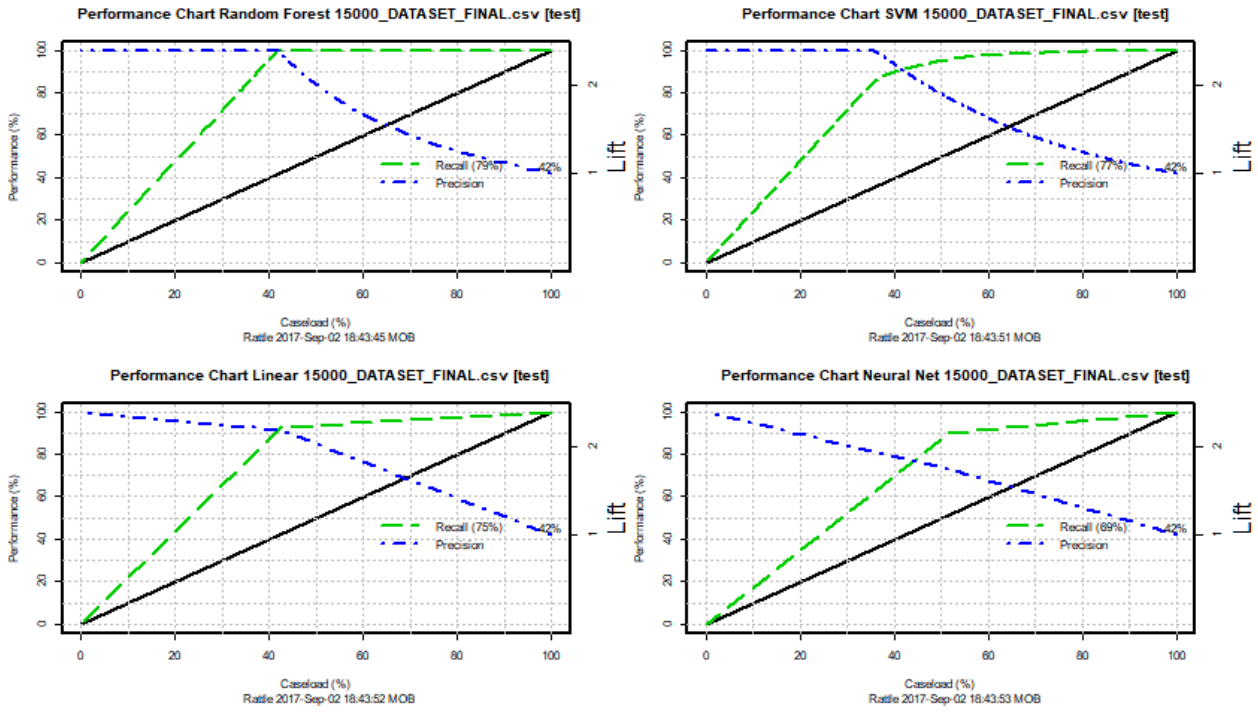


Figure 11 Risk and Recall Curves for Ghanaian Dataset (10 Folds Cross-Validation)

Area Under The Roc Curve

Several researchers have adopted area under the ROC to evaluate classifiers based upon it popularities and accuracy. From above discussion in chapter two of this work, we have underlined that ROC curves as a better evaluation metric and preferred by most researchers than accuracy when evaluation of classifiers are considered. From Garcia Pagans (2015: Chapter 7) Most works has adopted an area ranges below from excellent to poor ratings of classifier if they fall within setting range with a threshold of 0.5%

- 1 to 0.90 as "Excellent"
- 0.90 to 0.80 as "Good"
- 0.70 to 0.80 as "Correct"
- 0.70 to 0.60 as "Poor"
- 0.60 to 0.50 as "Bad"

Table 13 Area under the ROC curve (10 Fold Cross-Validation)

MODELS	AUC (ROC CURVES)
GHANA	
LR	93% (0.9340)
SVM	**97% (0.9729)**
ANN	83% (0.8282)
RF	100% (1)
GERMAN	
LR	**69% (0.6925)**
SVM	**68% (0.6830)**
ANN	67% (0.6706)
RF	**68% (0.6811)**
AUSTRALIAN	
LR	**90% (0.9037)**
SVM	**90% (0.8982)**
ANN	61% (0.6084)
RF	**90% (0.9003)**

From table 14 demonstrate each performance the individual algorithms or classifier using the Area under (ROC) curve from the Ghanaian data set SVM gain 97% followed by LR with 93% are excellent in terms of performance. From fig.10, it can be seen clearly that there would be more gains when we should choose SVM has our main classifier with less false positive or LR with 93 significantly as an alternative would not put the company into disrupt. ANN with 83% accuracy with higher false positive rate would lead to disaster if that should be chosen as a classifier and can be rated as good.
In the German data sets LR with 69% and SVM and RF achieving 68% each depicts poor performance and ANN with 67% indicates better other algorithms or improved models should be suggested when this dataset is adopted whereas in the case of Australian LR, SVM, RF was rated 90% as excellent classifier for each with higher true positive than false positive. From the diagram at the appendix indicates any of this models when selected for this data set would do much more better in terms of predicting each class accurately than ANN with 61% true positive which have false positive would lead to disastrous situations of applicants are offered loans.

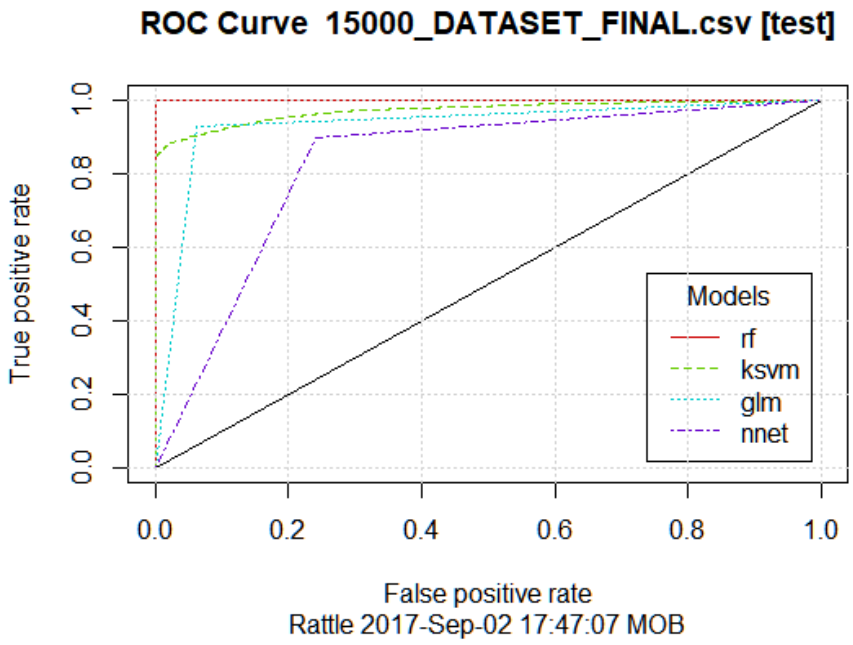


Figure 12 ROC Curve for Ghanaian Dataset (10 Fold Cross-Validation)

Predicted Versus Observed

The predicted versus observed graph is very important linear models such regressions models which would not be better option of discrete value than predicting a continuous value. This graph plots observed values on the horizontal axis and predicted values on the vertical axis. The two lines, which are plotted diagonally, the blue line is a linear fit to the real point's whiles the other is a perfect fit, that is if the predicted values are the same as the real values observed.

In this case, the Pseudo R-squared measures the resemblance by mimicking of the R-squared. This is calculated as the square of the correlation between the observed values and the predicted values. An algorithm would be much better preferred if the Pseudo R-square is closer to one (1).

From fig.11 below RF =0.9992, SVM=0.7951, LR = 0.7503 and ANN = 0.4201 and from above explanation Random forest would be much more preferred to be selected as an algorithm than any other classifiers since it is able to mimic the observed values and predict accurately without misclassifying.

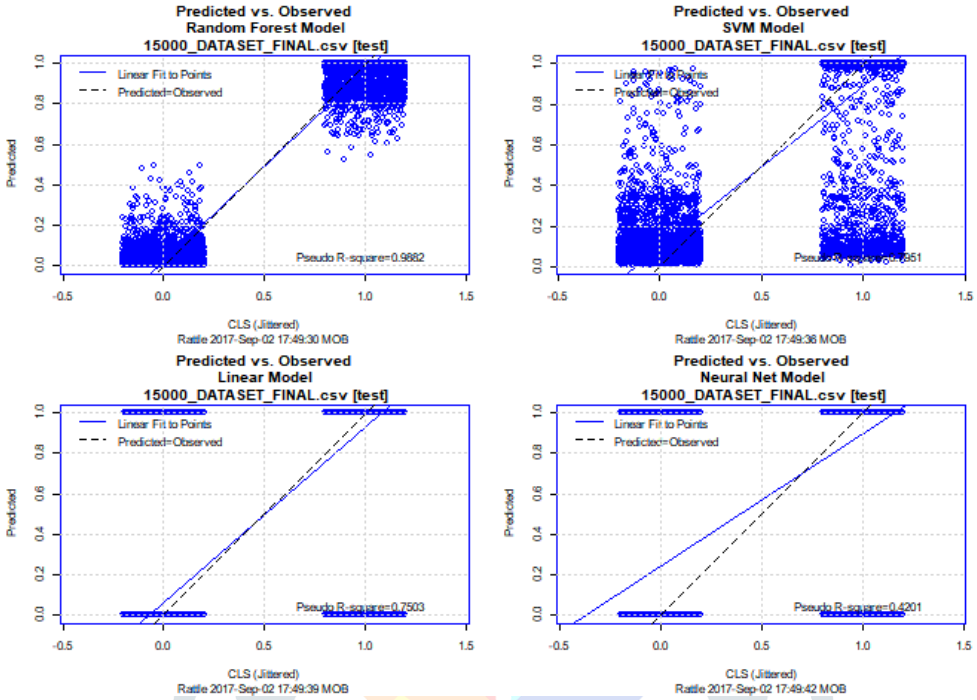


Figure 13 Predicted vs. Observed for Ghanaian Dataset (10 Folds Cross-Validation)

4.2.5 Phase two: 70/15/15 cross-validation partitioning

Phase two of this work partition our data sets into 70% for training our algorithms, 15% of the dataset for validation and the rest for testing which is used an estimation criterion for performance of the all four algorithms.

Error Matrix

Table 14 MISCLASSIFICATION ERRORS (70/15/15 Cross-Validation partition)

MODELS	TYPE II ERROR	TYPE I ERROR	OVERALL ERROR	AVE. ERROR	ACCURACY
GHANA					
LR	322 (0.14)	**34 (0.02)**	16%	14%	**85% (0.85)**
SVM	2 (0.00)	107 (0.05)	5%	6%	95% (0.95)
ANN	299 (0.13)	96 (0.04)	18%	16%	83% (0.83)
RF	0 (0)	0(0)	0	0	100% (1)
GERMAN					
LR	11 (0.07)	**24 (0.16)**	23%	30%	**76% (0.76)**
SVM	8 (0.05)	25 (0.17)	22%	29%	78% (0.78)
ANN	14 (0.09)	**24 (0.16)**	25%	31%	74% (0.74)
RF	6 (0.04)	**24 (0.16)**	20%	37%	**80% (0.80)**
AUSTRALIAN					
LR	9 (0.09)	10 (0.10)	18%	18%	81% (0.81)
SVM	9 (0.09)	**7 (0.07)**	15%	16%	**84% (0.84)**
ANN	6 (0.06)	8 (0.08)	13%	14%	86% (0.86)
RF	4 (0.04)	**7 (0.07)**	11%	11%	**89% (0.89)**

From Table 14 in other to minimize cost that would lead to severe implication on the side of the bank we must consider all the results of the type I error of this stage. In the Ghanaian dataset, we had 2% of type I error as the minimum type I error with an average error of 14% and accuracy of 85% whereas SVM having a highest average accuracy of 95% with 2% of type I error but higher type II error. SVM cannot be chosen over LR even though it had the best average accuracy. RF as an option in case if all models fail should be adopted performed extremely well-handling misclassification errors with 0%.

In the case of German data set, LR, ANN, and RF performed equally with a type I error of 16%. Even though this is not too good but it is acceptable. LR, ANN, and RF had 30%, 31% and 37% as average error respectively with an accuracy of 80% for RF, 76% for LR and 74% for ANN. LR would be preferred to all singles classifiers and RF could be used an alternative due to higher accuracy classify but also has the highest average error of 37% which makes less competitive as an option.
Comparing the performance of the classifiers in the Australian data set, SVM had the lowest type I error with 7% and average 16%, which is significant in our study with an average accuracy of 84% and RF 7% as type I, error with an average of 11% with 89% as average accuracy. This demonstrates that SVM performs better in the Australian data set when compared to the singles classifiers and RF can be preferred to that if we need higher accuracy rate.

To conclude, from the above demonstrate that LR is a perfect suit for running credit scoring in Ghanaian banking industry with RF as an alternative if we want to enforce high efficiency and effectiveness in classifying credit applicants. In the German data set LR as a choice is preferable even though RF can be adopted as an option but it the average error is 37% which makes it less competitive in competing with the LR in the German data set. SVM can be preferred to all other single classifiers but RF performs well with minimum average errors with higher accuracy.

Risk Curves

Table 15 Risk and Recall Curves (70/15/15 Cross-Validation partition)

MODELS	ACCURACY
GHANA	
LR	90% (0.896)
SVM	**100% (0.996)**
ANN	88% (0.879)
RF	**100%**
GERMAN	
LR	86% (0.859)
SVM	**87% (0.867)**
ANN	80% (0.802)
RF	**87% (0.868)**
AUSTRALIAN	
LR	94% (0.944)
SVM	93% (0.933)
ANN	**95% (0.947)**
RF	**97% (0.967)**

From table 16 of our risk and recall curves and figure 12 we have a strike line of 43% being a target as a risk score. This would mean that for every caseload or situations we would have 43% of the population not complying with the terms of the loans in the Ghana situation. From table 12, we had SVM and RF higher accuracy scores with 100%, 90% for LR and ANN 88% indicates that for every caseload 90% of the applicants must be audit properly in terms of LR, 88% must be audit properly when we adopt ANN and SVM and RF requires all applicant must be audited properly before loans are granted to applicants.

In the case of Australian dataset, we had 40% as a strike line and 33% as a strike for German data set. The risk scored by these two situations would need much more audit and attention to be paid when selecting each caseload in terms of default situations and scoring applicants. In the German situation LR, SVM, ANN, and RF requires 86%, 87%, 80% and 87% respectively of the cases should be investigated properly before loans are offered to those applicants. In the case of the Australian data set, 94%, 93%, 95% and 97% are the requirements estimated by each algorithm thus LR, SVM, ANN and RF respectively to be properly investigated before loans are granted to those applicants.

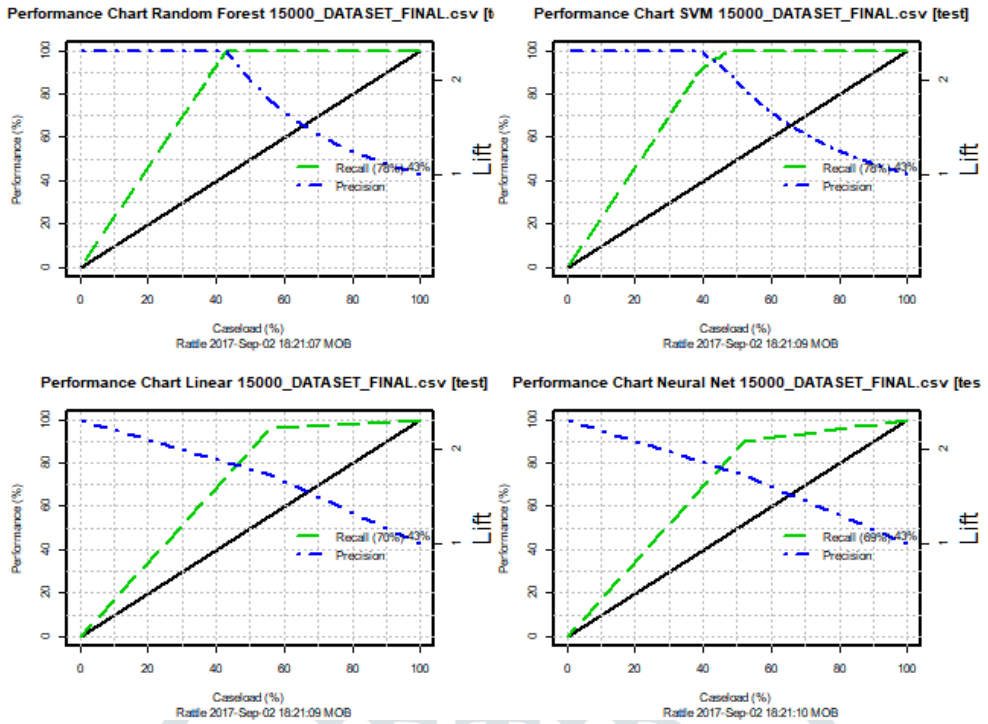


Figure 14 Risk and Recall Curves for Ghanaian Dataset (70/15/15 Cross-Validation partition)

To conclude, in every caseload we would have 43% for Ghanaian situation, 33% for German situation and 40% for Australian of the applicants defaulting as a risk score and applicants with higher scores should also be audited properly.

Area Under The Roc Curve

Table 16 Area under the ROC curve (70/15/15 Cross-Validation partition)

MODELS	AUC (ROC CURVES)
GHANA	
LR	86% (0.8567)
SVM	**99% (0.9942)**
ANN	83% (0.8337)
RF	**100% (1)**
GERMAN	
LR	82% (0.8234)
SVM	**83% (0.8336)**
ANN	75% (0.7526)
RF	**83% (0.8345)**
AUSTRALIAN	
LR	**92% (0.9174)**
SVM	**92% (0.9026)**
ANN	**92% (0.9226)**
RF	95% (0.9513)

From table 18 using the threshold of 0.5 for the area under the ROC curve with ratings below in terms of performance (Garcia Pagans, 2015):
From the Ghanaian dataset in table 18, we had SVM giving excellent results with 99% accuracy with less or zero false positives rate whiles the other single classifiers are performing good between 83-86%. RF as an alternative was 100% in terms of predictive powers and classifying true positive and false positive rates accurately.

- 1 to 0.90 as “Excellent”
- 0.90 to 0.80 as “Good”
- 0.70 to 0.80 as “Correct”
- 0.70 to 0.60 as “Poor”
- 0.60 to 0.50 as “Bad”

In the case of German data set, SVM had 83% and RF as alternative in case if all models should fail to predict accurately had 83% which both classifiers. ANN performed poor with 75% accuracy meaning it would have a lot of false positive if this classifier should be an option. In the Australian dataset, all three single classifiers performed equally with an average error of 92% which is excellent meaning any of this classifier is selected for classification problem using Australian data set would have less of disastrous issues. RF out performed all three single classifiers.

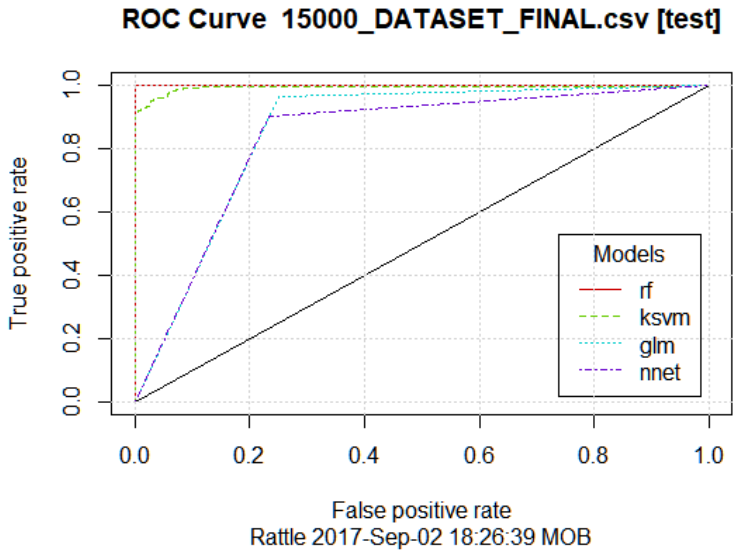


Figure 15 ROC Curve for Ghanaian Dataset (Data Segmentation)

Predicted Versus Observed

The predicted versus observed graph is very important linear models such regressions models which would not be the better option of discrete value than predicting a continuous value. This graph plots observed values on the horizontal axis and predicted values on the vertical axis. The two lines, which are plotted diagonally, the blue line is a linear fit to the real point's whiles the other is a perfect fit, that is if the predicted values are the same as the real values observed.

In this case, the Pseudo R-squared measures the resemblance by mimicking of the R-squared. This is calculated as the square of the correlation between the observed values and the predicted values. An algorithm would be much better preferred if the Pseudo R-square is closer to one (1).

From fig.11 below RF =0.9972, SVM=0.8748, LR = 0.5063 and ANN = 0.4378 and from above explanation Random forest would be much more preferred to be selected as an algorithm than any other classifiers since it is able to mimic the observed values and predict accurately without misclassifying.

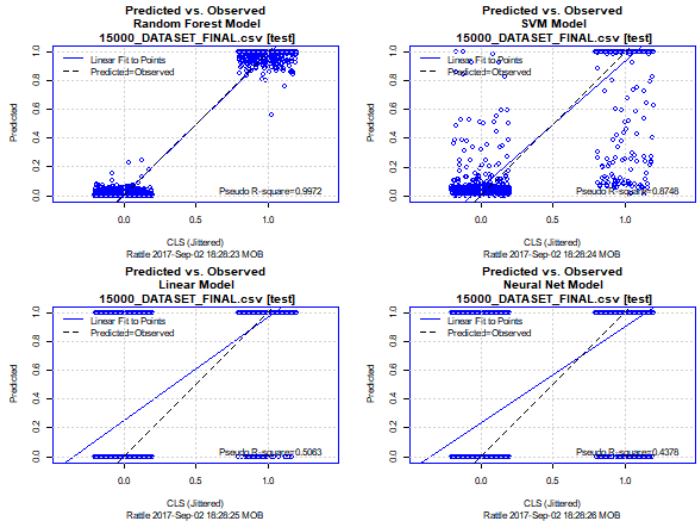


Figure 16 Predicted vs. Observed for Ghanaian Dataset (Data Segmentation)

V. FINDINGS, LIMITATIOSN AND RECOMMENDATIONS

5.1 Discussion, Findings, and Conclusion

In this work, we developed a linear discriminate model for predicting defaults probability of a loan application if given certain information or variables. In our experimental work we compared three basic major classifiers performance which are artificial neural networks (ANN), support vector machine (SVM) and logistic regression (LR) with random forest as an ensembler and alternative in case if all the three models or algorithms fail in terms of classifying the binary situation using finished data from commercial bank XYZ in Ghana.

From our linear discriminate analysis using logistic we found the following variables were significant such as; male constituting 86.7% of our total population having access to loans with ease compared to females, this could be as a results of accessibility to loans by males are easier than female and influence of cultural differences (Godquin, 2004). Marital status as a significant variable demonstrates that married people are less likely to default than singles, which is, have been a consistent through literature (Dinh & Kleimeier, 2007; Kočenda & Vojtek, 2011). We discovered that income status was also a predictor of either defaulting or not as well as default behaviour is not determine by absolute income which is also confirms (Ofori, Fianu, Omoregie, Odai, & Oduro-Gyimah, 2014). Other variables would be due to cultural difference and environmental set ups and requirements for granting loans. We might have other different variables captured in this study as significant since our main purpose was to classify credit applicants into class of defaults and non-defaulters.

Our work demonstrate the performance of predictive accuracy of a three single classifier and a hybrid model on three datasets from Ghana, German, and Australia used for the experiment using LR, ANN, SVM and RF as another complex model. In theory, it proves that multiples classifiers (Combination of models) perform better compared than single classifiers. Results show otherwise, on the average performance and predictive accuracy that single classifiers are best compared to the multiple classifiers or some hybrid classifiers when tested on the German and Australian datasets and this support the finding of (C. F. Tsai & Wu, 2008).

Evaluation depicts that artificial neural networks, support vector machine, and logistic regression are incomparable to the random forest classifier on the Ghanaian data sets with fewer input features relatively which also confirms (Bhattacharyya, 2013; Fernández-Delgado et al., 2014). Among the single classifiers SVM has outstanding performs in terms of classifying binary problem is proven from our analysis and (C.-L. Huang et al., 2007). In addition, test on all three datasets demonstrates that RF, SVM and LR performed very well and can be used as an alternative in each case of our studies whiles ANN model was inferior significantly. Even though D. Z. D. Zhang et al., (2008) added GP, BNP and C4.5 models in their experimental findings, they demonstrated that SVM and LR can be used as alternatives.

We also found out that by comparing two different data mining techniques through the selection methods, the accuracy of each algorithm was improved when it has more samples of instances for training in the case of partitioning the data set into 70/15/15 of our algorithms compared to 10 fold cross-validation. We adapted two data partitions to demonstrate accuracy and efficiency of our models in the Ghanaian environment where ensemble model was rank as the best performing model in all partition cases whiles SVM and LR model performance was better compared to ANN in all cases. To add, SVM was performing better among the individual models in 10 folds cross validation. Also, test demonstrated by Dahiya et al., (2015) achieved the same results even though their partition were different.

This work gives solutions applicable and appropriate for the benefits of all banks and financial institutions. Aspects of the bank's credit operations through the implementation of systems such as credit scoring systems which would give banks and financial institutions advantage over their competitors such as high efficiency and effectiveness of operations, increase market shares and profitability, losses and cost reductions in operations, and professional image development. Hence, a commercial adaptation of credit scoring tools should be able to compete in the banking and financial industry as well as building competitive advantage over their competitors through the use of complex and advanced but simple tools as risk reductions tools.

5.2 Limitation

This study did not consider the default probability or credit worthiness of applicants who were rejected in the process of applying for loans. In other words, our sample data set contains information of only applicants who were granted the loans and defaulted. This could result to biasness in our models or estimation technique developed, however, it was applicable to some other researchers such as (Kočenda & Vojtek, 2011; Ofori et al., 2014). Time was a major constraint since we had to work around the clock to produce this work within the shorted possible time. We were not able to acquire a representative data set from all the ten regions for our analysis due to confidential reason and integrity of banks to protect their customers as well as financial resources

5.3 Recommendations

We recommend that work such as this would need much time to complete and researchers must have dedicate good time into it. Researchers must use representative data sets to avoid biasness in their analysis. Prior permission must be seek for targeted banks and financial institutions as a source of data sets should be contacted before the commerce of the research.

The limitedness of a single classifier does exist which did not meet our expectation in terms of performance of these classifiers. We believe instead of choosing single classifiers or best single classifiers, ensembler or clustering of algorithms one must better understand the problem and challenges they want to address in specific. In the adaptation of a final classifier, the necessary solutions and decision needed for constructing and integrating problem-solving mechanism with it complexities should be considered when developing better classifiers.

This study can be developed through various improved ways in the future research. Firstly, we believe researchers must look for more relevant variables when collecting their data with intention of increasing the accuracy of predictive models. Secondly, newer classification models can be developed with advanced methodologies such as hybrids supports vector machines, complex classification and regression trees, linear discriminate with complexities and higher efficiency, other hybridizations of artificial neural networks such as learning quantization vector, Bayesian neural network learning algorithms, radial basis function, and fuzzy adaptive resonance.

VI. Acknowledgment

I thank those financial institutions who gave me the data for my analysis and other assistance given to me through resources and guidance.

References

[1] Abdou, H., El-Masry, A., & Pointon, J. (2007). on the Applicability of Credit Scoring Models in Egyptian Banks. *Banks and Bank Systems*, *2*(1), 4–20.

[2] Abdou, H., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications*, *35*(3), 1275–1292. https://doi.org/10.1016/j.eswa.2007.08.030

[3] Abellán, J., & Mantas, C. J. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *41*(8), 3825–3830. https://doi.org/10.1016/j.eswa.2013.12.003

[4] Akko, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*, *222*(1), 168–178. https://doi.org/10.1016/j.ejor.2012.04.009

[5] Ala'Raj, M., & Abbod, M. (2015). A systematic credit scoring model based on heterogeneous classifier ensembles. In *INISTA 2015 - 2015 International Symposium on Innovations in Intelligent SysTems and Applications, Proceedings* (pp. 1–7).

https://doi.org/10.1109/INISTA.2015.7276736

[6] Ala'raj, M., & Abbod, M. F. (2015). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, *104*, 89–105. https://doi.org/10.1016/j.knosys.2016.04.013

[7] Ala'raj, M., & Abbod, M. F. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems with Applications*, *64*, 36–55. https://doi.org/10.1016/j.eswa.2016.07.017

[8] Aliehyaei, R., & Khan, S. (2014). Ant Colony Optimization, Genetic Programming and a hybrid approach for credit scoring: A comparative study. *IEEE*. https://doi.org/doi:10.1109/SKIMA.2014.7083391

[9] Altman, E. I., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the U.S. market. *Abacus*, *43*(3), 332–357. https://doi.org/10.1111/j.1467-6281.2007.00234.x

[10] Anderson, R. (2007). *The Credit Scoring Toolkit - Theory and Practice for Retail Credit Risk Management and Decision Automation* (First). Oxford: Oxford University Press Inc., New York.

[11] Andrés, J. De, Lorca, P., Sánchez-lasheras, F., & Cos-juez, F. J. De. (2012). Bankruptcy Prediction and Credit Scoring: A Review of Recent Developments Based on Hybrid Systems and Some Related Patents. *Recent Patents on Computer Science*, *5*, 11–20. https://doi.org/10.2174/1874479611205010011

[12] Antonakis, a. C., & Sfakianakis, M. E. (2009). Assessing naïve Bayes as a method for screening credit applicants. *Journal of Applied Statistics*, *36*(5), 537–545. https://doi.org/10.1080/02664760802554263

[13] Ataman, K., & Street, W. (2005). Optimizing area under the ROC curve using ranking svms. *Kdd'05*. Retrieved from http://dollar.biz.uiowa.edu/~street/research/kdd05kaan.pdf

[14] Avery, R. B., Bostic, R. W., & Calem, P. S. (2000). Credit scoring: Statistical issues and evidence from credit-bureau files. *Real Estate Economics*, *28*(3), 523–547. Retrieved from http://socsci2.ucsd.edu/~aronatas/project/academic/Avery et al credit scoring.pdf

[15] Avery, R. B., Bostic, R. W., Calem, P. S., & Canner, G. B. (1996). Credit Risk Credit Scoring and the Performance of Home Mortgages. *Federal Reserve Bulletin*, 621–648. https://doi.org/10.1111/j.1467-9906.2007.00373.x

[16] Azayite, F. Z., & Achchab, S. (2016). Hybrid Discriminant Neural Networks for Bankruptcy Prediction and Risk Scoring. *Procedia Computer Science*, *83*(Ant), 670–674. https://doi.org/10.1016/j.procs.2016.04.149

[17] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

[18] Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*(4), 387–415. https://doi.org/http://dx.doi.org/10.1016/0022-2496(75)90001-2

[19] Banasiak, M. J., & Kiely, G. L. (2000). Predictive Collection Score Technology. *Business Credit*, *102*(2), 18. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=2805888&site=ehost-live

[20] Basel Committee on Banking Supervision. (2000). Principles for the management of credit risk. *Risk Management Group of the Basel Committee on Banking Supervision*, (September).

[21] Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods*, *43*(1), 3–31. https://doi.org/10.1016/S0167-7012(00)00201-3

[22] Bekhet, H. A., & Eletter, S. F. K. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, *4*(1), 20–28. https://doi.org/10.1016/j.rdf.2014.03.002

[23] Bhattacharyya, S. (2013). Confidence in predictions from random tree ensembles. *Knowledge and Information Systems*, *35*(2), 391–410. https://doi.org/10.1007/s10115-012-0600-z

[24] Bijak, K., & Thomas, L. C. (2012). Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*, *39*(3), 2433–2442. https://doi.org/10.1016/j.eswa.2011.08.093

[25] Bishop, C. M. C., & Nasrabadi, N. (2006). Pattern Recognition and Machine Learning. In *Pattern Recognition and Machine Learning* (Vol. 16, pp. 1–738). https://doi.org/10.1117/1.2819119

[26] Blanco, A., Pino-Mejías, R., Lara, J., & Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications*, *40*(1), 356–364. https://doi.org/10.1016/j.eswa.2012.07.051

[27] Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*(1–2), 245–271. https://doi.org/10.1016/S0004-3702(97)00063-5

[28] Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

[29] Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*(421), 123–140. https://doi.org/10.1007/BF00058655

[30] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[31] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

[32] Campus, S. (2010). Research on Application of Personal Credit Scoring based on BP-Logistic Hybrid Algorithm **[illegible]**. *IEEE*, (Iccasm), 735–739.

[33] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

[34] Chawla, N. V, Japkowicz, N., & Kotcz, A. (2004). Editorial : Special Issue on Learning from Imbalanced Data Sets. *ACM SIGKDD Explorations Newsletter*, *6*(1), 1–6. https://doi.org/http://doi.acm.org/10.1145/1007730.1007733

[35] Chen, B., & Lin, Y. (2014). Applications of Artificial Intelligence Technologies in Credit Scoring : a Survey of Literature. *2014 10th International Conference on Natural Computation*, 658–664.

[36] Chen, C.-C., & Li, S.-T. (2014). Credit rating with a monotonicity-constrained support vector machine model. *Expert Systems with Applications*, *41*(16), 7235–7247. https://doi.org/10.1016/j.eswa.2014.05.035

[37] Chen, H. C., & Chen, Y. C. (2010). A comparative study of discrimination methods for credit scoring. *40th International Conference on Computers and Industrial Engineering: Soft Computing Techniques for Advanced Manufacturing and Service Systems, CIE40 2010*. https://doi.org/10.1109/ICCIE.2010.5668170

[38] Chen, M. C., & Huang, S. H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications*, *24*(4), 433–441. https://doi.org/10.1016/S957-4174(02)00191-4

[39] Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6), 866–883. https://doi.org/10.1109/69.553155

[40] Chi, B. W., & Hsu, C. C. (2012). A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. *Expert Systems with Applications*, *39*(3), 2650–2661. https://doi.org/10.1016/j.eswa.2011.08.120

[41] Correa, A. B., & Gonzalez, A. M. (2011). Evolutionary algorithms for selecting the architecture of a MLP Neural Network: A credit scoring case. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 725–732. https://doi.org/10.1109/ICDMW.2011.80

[42] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

[43] Cristianini, N., & Shawe-Taylor, J. (2000a). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. *Book*. Cambridge University Press.

[44] Cristianini, N., & Shawe-Taylor, J. (2000b). *An Introduction to Support Vector Machines and other kernel based learning methods*. *Ai Magazine* (Vol. 22). https://doi.org/citeulike-article-id:114719

[45] Cristianini, N., & Shawe-Taylor, J. (2000c). *An Introduction to Support Vector Machines and other kernel based learning methods*. New York, NY, USA: Cambridge University Press. https://doi.org/citeulike-article-id:114719

[46] Crone, S. F., & Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, *28*(1), 224–238. https://doi.org/10.1016/j.ijforecast.2011.07.006

[47] Crook, J. N., Edelman, D. B., & Thomas, L. C. (2006). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*(3), 1447–1465. https://doi.org/10.1016/j.ejor.2006.09.100

[48] Dahiya, S., Handa, S. S., & Singh, N. P. (2015). Credit scoring using ensemble of various classifiers on reduced feature set. *Industrija*, *43*(4), 163–174. https://doi.org/10.5937/industrija43-8211

[49] Danenas, P., & Garsva, G. (2015). Selection of Support Vector Machines based classifiers for credit risk domain. *Expert Systems with Applications*, *42*(6), 3194–3204. https://doi.org/10.1016/j.eswa.2014.12.001

[50] del Castillo, M. D., & Serrano, J. I. (2004). A multistrategy approach for digital text categorization from imbalanced documents. *ACM SIGKDD Explorations Newsletter*, *6*(1), 70. https://doi.org/10.1145/1007730.1007740

[51] Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, *7*, 1–30. https://doi.org/10.1016/j.jecp.2010.03.005

[52] Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *95*(1), 24–37. https://doi.org/10.1016/0377-2217(95)00246-4

[53] Diana, T. (2005). Credit Risk Analysis And Credit Scoring-- Now And In The Future. *Business Credit*, *107*(3), 12–16. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=16336649&site=ehost-live

[54] Dinh, T. H. T., & Kleimeier, S. (2007). A credit scoring model for Vietnam's retail banking market. *International Review of Financial Analysis*, *16*(5), 471–495. https://doi.org/10.1016/j.irfa.2007.06.001

[55] Dong, Y. D. Y. (2007). An Application of Support Vector Machines in Small-Business Credit Scoring. In *Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007)* (pp. 112–112). https://doi.org/10.1109/ICICIC.2007.128

[56] Doori, M. Al, & Beyrouti, B. (2014). Credit Scoring Model Based on Back Propagation Neural Network Using Various Activation and Error Function. *Ijcsns*, *14*(3), 16–24. Retrieved from http://paper.ijcsns.org/07_book/201403/20140303.pdf

[57] Doumpos, M., Niklis, D., Zopounidis, C., & Andriosopoulos, K. (2015). Combining accounting data and a structural model for predicting credit ratings: Empirical evidence from European listed firms. *Journal of Banking and Finance*, *50*, 599–607. https://doi.org/10.1016/j.jbankfin.2014.01.010

[58] Dukiü, D., Dukiü, G., & Kvesiü, L. (2011). A Credit Scoring Decision Support System. *Proceedings of the ITI 2011 33rd Int. Conf. on Information Technology Interfaces, June 27-30, 2011, Cavtat, Croatia*, 391–396.

[59] Egan, J. P. (1975). *Signal detection theory and ROC-analysis*. New York: Academic Press.

[60] Elkan, C. (2001). The Foundations of Cost-sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2* (pp. 973–978). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1642194.1642224

[61] Equal Credit Opportunity Act (ECOA). (2013). CFPB Consumer Laws and Regulations ECOA CFPB Consumer Laws and Regulations, *6*(June), 1–35. Retrieved from http://files.consumerfinance.gov/f/201306_cfpb_laws-and-regulations_ecoa-combined-june-2013.pdf

[62] Fan, Y.-Q., Yang, Y.-L., & Qin, Y.-S. (2013). Credit scoring model based on PCA and improved tree augmented Bayesian classification. *IET International Conference on Information and Communications Technologies (IETICT 2013)*, 169–175. https://doi.org/10.1049/cp.2013.0051

[63] Fawcett, T. (2005). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[64] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. https://doi.org/10.1145/240455.240464

[65] Fensterstock, A. (2001, March). The Application of Neural Networks to Credit Scoring. *Business Credit*, *103*(3), 58. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=4187184&site=ehost-live

[66] Fensterstock, A. (2003, March). Credit Scoring Basics. *Business Credit*, *105*(3), 10–12. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=9215760&site=ehost-live

[67] Fensterstock, A. (2005, March). Credit Scoring And The Next Step. *Business Credit*, *107*(3), 46–49. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=16336667&site=ehost-live

[68] Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., & Amorim Fernández-Delgado, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, *15*, 3133–3181. https://doi.org/10.1016/j.csda.2008.10.033

[69] Finlay, S. M. (2006). Predictive models of expenditure and over-indebtedness for assessing the affordability of new consumer credit applications. *Journal of the Operational Research Society*, *57*(6), 655–669.

https://doi.org/10.1057/palgrave.jors.2602030

[70] Garcia Pagans, F. (2015). *Predictive Analytics Using Rattle and Qlik Sense*. Retrieved from https://books.google.com.au/books?id=fu0RCgAAQBAJ

[71] Ghana Borrowers and Lenders Act. Ghana Borrowers and Lenders Act, 2008, Pub. L. No. 773 (2008). http://www.bu.edu/bucflp/files/2012/01/Borrowers-and-Lenders-Act-No.-773.pdf. Retrieved from http://www.bu.edu/bucflp/files/2012/01/Borrowers-and-Lenders-Act-No.-773.pdf

[72] Ghodselahi, A., & Amirmadhi, A. (2011). Application of Artificial Intelligence Techniques for Credit Risk Evaluation. *International Journal of Modeling and …*, *1*(3). Retrieved from http://www.ijmo.org/papers/43-A10212.pdf

[73] Giesecke, K. (2012). Credit risk modeling and valuation: An introduction. *Available at SSRN 479323*, (June), 1–40. https://doi.org/10.2139/ssrn.479323

[74] Godquin, M. (2004). Microfinance repayment performance in Bangladesh: How to improve the allocation of loans by MFIs. *World Development*, *32*(11), 1909–1926. https://doi.org/10.1016/j.worlddev.2004.05.011

[75] Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. *John Wiley & Sons* (Vol. 1). https://doi.org/10.1901/jeab.1969.12-475

[76] Gu, Q., Zhu, L., & Cai, Z. (2009). Evaluation Measures of the Classification Performance of Imbalanced Data Sets BT - Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, October 23-25, 2009. Proceedings. In Z. Cai, Z. Li, Z. Kang, & Y. Liu (Eds.) (pp. 461–471). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04962-0_53

[77] Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research*, *249*(2), 417–426. https://doi.org/10.1016/j.ejor.2015.05.050

[78] Hamel, L. (2009). *Knowledge Discovery with Support Vector Machines*. https://doi.org/10.1002/9780470503065

[79] Han, H., Wang, W., & Mao, B. (2005). Borderline-SMOTE : A New Over-Sampling Method in. *Lecture Notes in Computer Science*, *3644*, 878–887.

[80] Han, L., Han, L., & Zhao, H. (2013). Orthogonal support vector machine for credit scoring. *Engineering Applications of Artificial Intelligence*, *26*(2), 848–862. https://doi.org/10.1016/j.engappai.2012.10.005

[81] Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

[82] Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, *45*(2), 171–186. https://doi.org/10.1023/A:1010920819831

[83] Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Critical Reviews in Diagnostic Imaging*. Retrieved from http://europepmc.org/abstract/med/2667567

[84] Hanley, J. A., & McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating (ROC) Curvel Characteristic. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

[85] Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*(3), 839–843. https://doi.org/10.1148/radiology.148.3.6878708

[86] Harris, T. (2013). Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions. *Expert Systems with Applications*, *40*(11), 4404–4413. https://doi.org/10.1016/j.eswa.2013.01.044

[87] Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications*, *42*(2). https://doi.org/10.1016/j.eswa.2014.08.029

[88] Hayashi, Y. (2016). Application of a rule extraction algorithm family based on the Re-RX algorithm to financial credit risk assessment from a Pareto optimal perspective. *Operations Research Perspectives*, *3*, 32–42. https://doi.org/10.1016/j.orp.2016.08.001

[89] Haykin, S. (1994). Neural networks-A comprehensive foundation. *New York: IEEE Press. Herrmann, M., Bauer, H.-U., & Der, R.* https://doi.org/10.1017/S0269888998214044

[90] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[91] He, J., Zhang, Y., Shi, Y., Member, S., & Huang, G. (2010). Domain-Driven Classification Based on Multiple Criteria and Multiple Constraint-Level Programming for Intelligent Credit Scoring. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, *22*(6), 826–838.

[92] Hengpraprohm, S., & Chongstitvatana, P. (2009). FEATURE SELECTION BY WEIGHTED-SNR FOR CANCER MICROARRAY DATA CLASSIFICATION. *International Journal of Innovative Computing, Information and Control*, *5*(12 (A)), 4627–4635.

[93] Hernandez-Orallo, J., Ferri, C., Lachiche, N., & Flach, P. a. (2004). ROC Analysis in Artificial Intelligence. In *1st Int. Workshop on ROC Analysis in Artificial Intelligence (ROCAI 2004), Valencia, Spain* (pp. 71–80).

[94] Holte, R. C., Acker, L., & Porter, B. (1989). Concept Learning and the Problem of Small Disjuncts. *IJCAI'89 Proceedings of the 11th International Joint Conference on Artificial Intelligence*, *1*, 813–818. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.2549&rep=rep1&type=pdf%5Cnhttp://ijcai.org/Past Proceedings/IJCAI-89-VOL1/PDF/130.pdf

[95] Hossin, M., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *5*(2), 1–11. https://doi.org/10.5121/ijdkp.2015.5201

[96] Hossin, M., Sulaiman, M. N., Mustapha, A., Mustapha, N., & Rahmat, R. W. (2011). A hybrid evaluation metric for optimizing classifier. In *Conference on Data Mining and Optimization* (pp. 165–170). https://doi.org/10.1109/DMO.2011.5976522

[97] Hsieh, N. C., & Hung, L. P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications*, *37*(1), 534–545. https://doi.org/10.1016/j.eswa.2009.05.059

[98] Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856. https://doi.org/10.1016/j.eswa.2006.07.007

[99] Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310. https://doi.org/10.1109/TKDE.2005.50

[100] Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector

machine. *Computers & Operations Research*, *32*, 2513–2522. https://doi.org/10.1016/j.cor.2004.03.016

[101] Huang, Y.-M., Hung, C.-M., & Jiau, H. C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications*, *7*(4), 720–747. https://doi.org/10.1016/j.nonrwa.2005.04.006

[102] Hussain Ali Bekhet and Shorouq Fathi Kamel Eletter. (2012). Credit Risk Management for the Jordanian Commercial Banks: A business Intelligence Approach. *Australian Journal of Basic and Applied Sciences*, *6*(9), 188–195.

[103] Ince, H., & Aktan, B. (2009). A Comparison of Data Mining Techniques for Credit Scoring in Banking: a Managerial Perspective. *Journal of Business Economics and Management Journal*, *10*(3), 233–240. https://doi.org/10.3846/1611-1699.2009.10.233-240

[104] Information, E., & Engineering, E. (2004). ANN-GA Approach of Credit Scoring for Mobile Customers. *Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems Singapore*, 1–3.

[105] J??ndel, M. (2010). A neural support vector machine. *Neural Networks*, *23*(5), 607–613. https://doi.org/10.1016/j.neunet.2010.01.002

[106] Jackson, R. H. G., & Wood, A. (2013). The performance of insolvency prediction and credit risk models in the UK: A comparative study. *British Accounting Review*, *45*(3), 183–202. https://doi.org/10.1016/j.bar.2013.06.009

[107] Jacobson, T., & Roszbach, K. (2003). Bank lending policy, credit scoring and value-at-risk. *Journal of Banking and Finance*, *27*(4), 615–633. https://doi.org/10.1016/S0378-4266(01)00254-0

[108] Jain, A., & Zongker, D. (1997). Feature Selection: Evaluation, Application, and Small Sample Performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(2), 153–158. https://doi.org/10.1109/34.574797

[109] Jamali, I., Bazmara, M., & Jafari, S. (2012). Feature Selection in Imbalance data sets. *International Journal of Computer Science (IJCSI))*, *9*(3), 42–45.

[110] Japkowicz, N., Myers, C., & Gluck, M. (1995). A novelty detection approach to classification. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (pp. 518–523). Retrieved from http://www.ijcai.org/Past Proceedings/IJCAI-95-VOL 1/pdf/068.pdf

[111] Jiang, M. H., & Yuan, X. C. (2007). Personal credit scoring model of non-linear combining forecast based on GP. *Proceedings - Third International Conference on Natural Computation, ICNC 2007*, *4*(1), 408–414. https://doi.org/10.1109/ICNC.2007.551

[112] Jiang, M., & Lin, S. (2010). Construction and application of ART on personal credit scoring. *Proceedings - 2010 International Conference on Digital Manufacturing and Automation, ICDMA 2010*, *1*(1), 434–436. https://doi.org/10.1109/ICDMA.2010.202

[113] Jiang, Y., & Wu, L. H. (2009). Credit Scoring Model Based on Simple Naive Bayesian Classifier and a Rough Set. *Ieee*, (2007), 1–4.

[114] Jorion, P. (2003). Financial Risk Manager Handbook. *Wiley Finance*. https://doi.org/10.1017/CBO9781107415324.004

[115] Kabir, M. N., Worthington, A., & Gupta, R. (2015). Comparative credit risk in Islamic and conventional bank. *Pacific Basin Finance Journal*, *34*, 327–353. https://doi.org/10.1016/j.pacfin.2015.06.001

[116] Kambal, E., Osman, I., Taha, M., Mohammed, N., & Mohammed, S. (2013). Credit scoring using data mining techniques with particular reference to Sudanese banks. In *2013 International Conference on Computing, Electrical and Electronic Engineering (Icceee)* (pp. 378–383). https://doi.org/10.1109/ICCEEE.2013.6633966

[117] Kellison, B., & Brockett, P. (2003). Check the score: credit scoring and insurance losses: Is there a connection? *Texas Business Review Special Issue*, 1–5. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=9539729&site=ehost-live

[118] Kennedy, K., Mac Namee, B., Delany, S. J., O'Sullivan, M., & Watson, N. (2013). A window of opportunity: Assessing behavioural scoring. *Expert Systems with Applications*, *40*(4), 1372–1380. https://doi.org/10.1016/j.eswa.2012.08.052

[119] Kiruthika, & Dilsha, M. (2015). A Neural Network Approach for Microfinance Credit Scoring. *Journal of Statistics and Management Systems*, *18*(1–2), 121–138. https://doi.org/10.1080/09720510.2014.961767

[120] Kočenda, E., & Vojtek, M. (2011). Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data. *Emerging Markets Finance and Trade*, *47*(1015), 80–98. https://doi.org/10.2753/REE1540-496X470605

[121] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

[122] Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, *27*, 11–23. https://doi.org/10.1016/j.jretconser.2015.07.003

[123] Krzanowski, W. J., & Hand, D. J. (2009). *ROC Curves for Continuous Data*. https://doi.org/doi:10.1201/9781439800225.fmatt

[124] Lee, T.-S., Chiu, C.-C., Chou, Y.-C., & Lu, C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, *50*(4), 1113–1130. https://doi.org/10.1016/j.csda.2004.11.006

[125] LEE, T. H., & ZHANG, M. (2003). BIAS CORRECTION AND STATISTICAL TEST FOR DEVELOPING CREDIT SCORING MODEL THROUGH LOGISTIC REGRESSION APPROACH. *International Journal of Information Technology & Decision Making*, *2*(2), 299–311. https://doi.org/10.1142/S0219622003000665

[126] Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, *23*(3), 245–254. https://doi.org/10.1016/S0957-4174(02)00044-1

[127] Lek, S., & Gu??gan, J. F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, *120*(2–3), 65–73. https://doi.org/10.1016/S0304-3800(99)00092-7

[128] Leonard, K. J. (1995). The development of credit scoring quality measures for consumer credit applications. *International Journal of Quality & Reliability Management*, *12*(4), 79–85. https://doi.org/10.1108/02656719510087346

[129] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

[130] Leung, K., Cheong, F., & Cheong, C. (2007). Consumer credit scoring using an artificial immune system algorithm. *2007 IEEE Congress on Evolutionary Computation, CEC 2007*, 3377–3384. https://doi.org/10.1109/CEC.2007.4424908

[131] Li, F. C., Wang, P. K., & Wang, G. E. (2009). Comparison of the primitive classifiers with extreme learning machine in credit scoring. *IEEM 2009 - IEEE International Conference on Industrial Engineering and Engineering Management*, *2*(4), 685–688. https://doi.org/10.1109/IEEM.2009.5373241

[132] Li, K., Niskanen, J., Kolehmainen, M., & Niskanen, M. (2016). Financial innovation: Credit default hybrid model for SME lending. *Expert Systems with Applications*, *61*, 343–355. https://doi.org/10.1016/j.eswa.2016.05.029

[133] Lin, C. C., Chang, C. C., Li, F. C., & Chao, T. C. (2011). Features selection approaches combined with effective classifiers in credit scoring. *IEEE International Conference on Industrial Engineering and Engineering Management*, 752–757. https://doi.org/10.1109/IEEM.2011.6118017

[134] Liu, H., & Motoda, H. (1998). *Feature selection for knowledge discovery and data mining*. *Springer Science+Business Media New York*. https://doi.org/10.1007/978-1-4615-5689-3

[135] Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *Ieee Transactions on Knowledge and Data Engineering*, *17*(4), 491–502. https://doi.org/10.1109/TKDE.2005.66

[136] Liu, X. Y., Fu, H., & Lin, W. W. (2010). A Modified Support Vector Machine model for Credit Scoring. *International Journal of Computational Intelligence Systems*, *3*(6), 797–804.

[137] Liu, Y., & Schumann, M. (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society*, *56*(9), 1099–1108. https://doi.org/10.1057/palgrave.jors.2601976

[138] Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: A systematic review and overall comparison. *Surveys in Operations Research and Management Science*, *21*(2), 117–134. https://doi.org/10.1016/j.sorms.2016.10.001

[139] Luo, C., Wu, D., & Wu, D. (2016). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, (September), 1–6. https://doi.org/10.1016/j.engappai.2016.12.002

[140] Lyu, S. L. S. (2005). Mercer kernels for object recognition with local features. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR05*, *2*(October), 223–229. https://doi.org/10.1109/CVPR.2005.223

[141] MacKay, D. J. C. (2005). *Information Theory, Inference, and Learning Algorithms David J.C. MacKay*. *Learning* (Vol. 100). https://doi.org/10.1198/jasa.2005.s54

[142] Maldonado, S., Weber, R., & Famili, F. (2014). Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines. *Information Sciences*, *286*, 228–246. https://doi.org/10.1016/j.ins.2014.07.015

[143] Mao, J. (1996). Why artificial neural networks? *Communications*, *29*, 31–44. https://doi.org/10.1109/2.485891

[144] Marqués, A. I., García, V., & Sánchez, J. S. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, *39*(11), 10244–10250. https://doi.org/10.1016/j.eswa.2012.02.092

[145] Marron, D. (2007). *"Lending by numbers": credit scoring and the constitution of risk within American consumer credit*. *Economy and Society* (Vol. 36). https://doi.org/10.1080/03085140601089846

[146] Mester, L. J. (1997). What Is the Point of Credit Scoring? *Business Review (Federal Reserve Bank of Philadelphia)*, (February 1997), 3–16.

[147] Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*(4), 283–298.

[148] Minh, H. Q., Niyogi, P., & Yao, Y. (2006). Mercer's theorem, feature maps, and smoothing. *Learning Theory*, *4005*(Lecture Notes in Computer Science), 154--168. https://doi.org/10.1007/11776420_14

[149] Nanni, L., Fantozzi, C., & Lazzarini, N. (2015). Coupling different methods for overcoming the class imbalance problem. *Neurocomputing*, *158*, 48–61. https://doi.org/10.1016/j.neucom.2015.01.068

[150] Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *36*(2 PART 2), 3028–3033. https://doi.org/10.1016/j.eswa.2008.01.018

[151] NANNI, L., & LUMINI, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *36*(2), 3028–3033. https://doi.org/10.1016/j.eswa.2008.01.018

[152] Napierała, K., & Stefanowski, J. (2015). Addressing imbalanced data with argument based rule learning. *Expert Systems with Applications*, *42*(24), 9468–9481. https://doi.org/10.1016/j.eswa.2015.07.076

[153] Nayak, G. N., & Turvey, C. G. (1997). Credit risk assessment and the opportunity costs of loan misclassification. *Canadian Journal of Agricultural Economics-Revue Canadienne D Economie Rurale*, *45*(3), 285–299. https://doi.org/10.1111/j.1744-7976.1997.tb00209.x

[154] Niklis, D., Doumpos, M., & Zopounidis, C. (2014). Combining market and accounting-based models for credit scoring using a classification scheme based on support vector machines. *Applied Mathematics and Computation*, *234*, 69–81. https://doi.org/10.1016/j.amc.2014.02.028

[155] Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

[156] Ofek, N., Rokach, L., Stern, R., & Shabtai, A. (2017). Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem. *Neurocomputing*, *243*, 88–102. https://doi.org/http://dx.doi.org/10.1016/j.neucom.2017.03.011

[157] Ofori, K. S., Fianu, E., Omoregie, K., Odai, N. A., & Oduro-Gyimah, F. (2014). Predicting Credit Default among Micro Borrowers in Ghana. *Research Journal of Finance and Accounting*, *5*(12), 96–105. Retrieved from http://www.iiste.org/Journals/index.php/RJFA/article/view/13574

[158] Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Systems with Applications*, *41*(4 PART 2), 2052–2064. https://doi.org/10.1016/j.eswa.2013.09.004

[159] Park, S. (2004). Solving the Mystery of Credit Scoring Models. *Business Credit*, *106*(3), 43–47. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=12599227&site=ehost-live

[160] Patra, S., Shanker, K., & Kundu, D. (2008). Sparse maximum margin logistic regression for credit scoring. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 977–982. https://doi.org/10.1109/ICDM.2008.84

[161] Pazhoheshfar, P., & Saberi, M Azadeh, A. (2011). Improving Accuracy of Artificial Neural Networks for Credit Scoring Models Using Voting Algorithm. *Intelligent and Advanced Systems (ICIAS), 2010 International Conference on*, (2011), 1–4.

https://doi.org/10.1109/ICIAS.2010.5716158

[162] Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. *Cancer Research*. Retrieved from http://www.amazon.com/dp/0198565828

[163] Pérez-Godoy, M. D., Fernández, A., Rivera, A. J., & Del Jesus, M. J. (2010). Analysis of an evolutionary RBFN design algorithm, CO2RBFN, for imbalanced data sets. *Pattern Recognition Letters*, *31*(15), 2375–2388. https://doi.org/10.1016/j.patrec.2010.07.010

[164] Peterson, W. W., Birdsall, T. G., & Fox, W. (1954). The theory of signal detectability. *Information Theory, Transactions of the IRE Professional Group on*, *4*(4), 171–212. https://doi.org/10.1109/TIT.1954.1057460

[165] Ping, Y. (2009a). Hybrid classifier using neighborhood rough set and SVM for credit scoring. *2009 International Conference on Business Intelligence and Financial Engineering, BIFE 2009*, 138–142. https://doi.org/10.1109/BIFE.2009.41

[166] Ping, Y. (2009b). Hybrid fuzzy SVM model using CART and MARS for credit scoring. In *2009 International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2009* (Vol. 2, pp. 392–395). https://doi.org/10.1109/IHMSC.2009.221

[167] Ping, Y., & Yongheng, L. (2011). Neighborhood rough set and SVM based hybrid credit scoring classifier. *Expert Systems with Applications*, *38*(9), 11300–11304. https://doi.org/10.1016/j.eswa.2011.02.179

[168] Pompella, M., & Dicanio, A. (2016). Ratings based Inference and Credit Risk: Detecting likely-to-fail Banks with the PC-Mahalanobis Method. *Economic Modelling*. https://doi.org/10.1016/j.econmod.2016.08.023

[169] Provost, F., & Domingos, P. (2003). Tree Induction for Probability Based Ranking. *Machine Learning*, *52*(3), 199–215.

[170] Qin, R., Liu, L. L., & Xie, J. (2010). An Application of Improved BP Neural Network in Personal Credit Scoring. *2010 Second International Conference on Computer Modeling and Simulation*, 238–241. https://doi.org/10.1109/ICCMS.2010.147

[171] Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1023/A:1022643204877

[172] Quittner, J. (2003). Subprime's Tech Dilemma. *Bank Technology News*, *16*(1), 19. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=8832050&site=ehost-live

[173] Rakotomamonjy, A. (2004). Optimizing Area Under Roc Curve with SVMs. In *Modelling, Compu- tation and Optimization Conference* (pp. 1–9).

[174] Ramanna, S., Jain, L. C., & Howlett, R. J. (2013). *Emerging Paradigms in Machine Learning*. (P. S. Ramanna, P. R. J. Howlett, & P. L. C. Jain, Eds.). Springer Heidelberg New York Dordrecht London. https://doi.org/10.1007/978-3-642-28699-5

[175] Ranawana, R., & Palade, V. (2006). Optimized Precision - A New Measure for Classifier Performance Evaluation. In *2006 IEEE International Conference on Evolutionary Computation* (pp. 2254–2261). https://doi.org/10.1109/CEC.2006.1688586

[176] Raskutti, B., & Kowalczyk, A. (2004). Extreme re-balancing for SVMs. *ACM SIGKDD Explorations Newsletter - Special Issue on Learning from Imbalanced Datasets*, *6*(1), 60–69. https://doi.org/10.1145/1007730.1007739

[177] Ri, S., Lq, H., Y, D. Q. L. X. D. Q. J., Olxvkxdq, P., Qhx, L. V. H., Fq, H. G. X., … Phwkrg, D. K. E. (2008). Application of Genetic Programming in Credit Scoring. *IEEE*, 1106–1110.

[178] Rimmer, J. (2005). Contemporary Changes in Credit Scoring. *Credit Control*, *26*(4), 56–60. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=17602788&site=ehost-live

[179] Roman, D., & Stefano, G. (2016). Towards a reference architecture for trusted data marketplaces: The credit scoring perspective. In *Proceedings - 2016 2nd International Conference on Open and Big Data, OBD 2016* (pp. 95–101). https://doi.org/10.1109/OBD.2016.21

[180] Rosset, S. (2004). Model Selection via the AUC. In *IN PROCEEDINGS OF THE 21ST INTERNATIONAL CONFERENCE ON MACHINE LEARNING* (pp. 1–8). Banff, Canada. Retrieved from https://www.tau.ac.il/~saharon/papers/auc-fixed.pdf

[181] Saunders, A., & Cornett, M. M. (2007). *Financial Institutions Management: A Risk Management Approach*. (B. Gordon & M. Janicek, Eds.), *The McGraw-Hill/Irwin Series in Finance, Insurance, and Real Estate* (SIXTH). Avenue of the Americas, New York, NY, 10020.: McGraw-Hill/Irwin,. Retrieved from http://www.bulentsenver.com/FIN5477/Financial_Institutions_Management_AntonySaunders_TextBook.pdf

[182] Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*. https://doi.org/10.1016/j.neunet.2014.09.003

[183] Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels – support vector machines, regularization, optimization and beyond*. Cambridge: Massachusetts: The MIT Press. https://doi.org/10.1198/jasa.2003.s269

[184] Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., & Anthony, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, *44*(5), 1926–1940. https://doi.org/10.1109/18.705570

[185] Shihab, S., Al-Nuaimy, W., Huang, Y., & Eriksen, A. (2003). A comparison of segmentation techniques for target extraction in ground penetrating radar data. In *2nd International Workshop on Advanced GPR* (pp. 95–100). Netherlands. https://doi.org/10.3997/1873-0604.2003016

[186] Siddiqi, N. (2006). Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. In *John Wiley & Sons, Inc.* (Vol. 1, pp. 1–210). Hoboken, New Jersey: John Wiley & Sons, Inc.

[187] Smaranda, C. (2014). Scoring Functions and Bankruptcy Prediction Models – Case Study for Romanian Companies. *Procedia Economics and Finance*, *10*(14), 217–226. https://doi.org/10.1016/S2212-5671(14)00296-2

[188] Somol, P., Baesens, B., Pudil, P., & Vanthienen, J. (2005). Filter-versus wrapper-based feature selection for credit scoring. *International Journal of Intelligent Systems*, *20*(10 (Oct.)), 985–999. Retrieved from https://lirias.kuleuven.be/bitstream/123456789/85624/1/filterversus.pdf

[189] Sousa, M. R., Gama, J., & Brandão, E. (2016). A new dynamic modeling framework for credit risk assessment. *Expert Systems with Applications*, *45*, 341–351. https://doi.org/10.1016/j.eswa.2015.09.055

[190] Spuchľáková, E., Valašková, K., & Adamko, P. (2015). The Credit Risk and its Measurement, Hedging and Monitoring. *Procedia Economics and Finance*, *24*(July), 675–681. https://doi.org/10.1016/S2212-5671(15)00671-1

[191] Sun, L., & Shenoy, P. P. (2007). Using Bayesian networks for bankruptcy prediction : Some methodological issues. *European Journal of Operational Research*, *180*(302), 738–753. https://doi.org/10.1016/j.ejor.2006.04.019

[192] Suykens, J., & Leuven, K. U. (2003). *A (short) Introduction to Support Vector Machines and Kernelbased Learning*.

Book (SVR).

[193] Swets, J. A. (1996). Signal detection theory and ROC analysis in psychology and diagnostics : collected papers. In *Scientific psychology series* (p. Chp 11). https://doi.org/10.1017/CBO9781107415324.004

[194] Swets, J. A., & Pickett, R. M. (1982). Evaluation of diagnostic systems: methods from signal detection theory. New York: Academic Press. Retrieved from file://catalog.hathitrust.org/Record/000106544

[195] Tang, B., & Qiu, S. (2012). A new Credit Scoring Method Based on Improved Fuzzy Support Vector Machine. *Computer Science and Automation Engineering (CSAE), 2012 IEEE International Conference on*, (4), 0–2.

[196] Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149–172. https://doi.org/10.1016/S0169-2070(00)00034-0

[197] Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit Scoring and Its Applications*. (M. Holmes, Ed.). Philadelphia: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898718317

[198] Thomas, L. C., Ho, J., & Scherer, W. T. (2001). Time will tell: Behavioural scoring and the dynamics of consumer credit assessment. *IMA Journal of Management Mathematics*, *12*(1), 89–103. https://doi.org/10.1093/imaman/12.1.89

[199] Tomczak, J. M., & Zieba, M. (2015). Classification Restricted Boltzmann Machine for comprehensible credit scoring model. *Expert Systems with Applications*, *42*(4), 1789–1796. https://doi.org/10.1016/j.eswa.2014.10.016

[200] Tong, E. N. C., Mues, C., & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, *218*(1), 132–139. https://doi.org/10.1016/j.ejor.2011.10.007

[201] Tsai, C. F. C. C. F. C.-F., & Wu, J. W. J.-W. J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *34*(4), 2639–2649. https://doi.org/10.1016/j.eswa.2007.05.019

[202] Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *34*(4), 2639–2649. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417407001558

[203] Van Gestel, I. T., Baesens, B., Garcia, I. J., & Van Dijcke, P. (2003). A support vector machine approach to credit scoring. *Forum Financier-Revue Bancaire Et Financiaire Bank En Financiewezen-*, (1), 73–82. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.6492&rep=rep1&type=pdf

[204] Van Rijsbergen, C. J. . (1979). INFORMATION RETRIEVAL. In *Information Retrieval* (second ed., pp. 112–140). https://doi.org/10.1007/SpringerReference_16360

[205] Vedala, R., & Kumar, B. R. (2012). An application of Naive Bayes classification for credit scoring in e-lending platform. *Proceedings - 2012 International Conference on Data Science and Engineering, ICDSE 2012*, 81–84. https://doi.org/10.1109/ICDSE.2012.6282321

[206] VENKAT, S., & KIM, Y. H. (1987). Credit Granting - Comparative analysis of classification procedures.pdf. *The Journal of Finance*, *42*(3), 665–681.

[207] Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, *238*(2), 505–513. https://doi.org/10.1016/j.ejor.2014.04.001

[208] Vicente, G., Marqués, A. I., & Sánchez, J. S. (2014). An insight into the experimental design for credit risk and corporate bankruptcy prediction systems. *Journal of Intelligent Information Systems*, *44*(1), 159–189. https://doi.org/10.1007/s10844-014-0333-4

[209] Walter, S. D. (2005). The partial area under the summary ROC curve. *Statistics in Medicine*, *24*(13), 2025–2040. https://doi.org/10.1002/sim.2103

[210] Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223–230. https://doi.org/10.1016/j.eswa.2010.06.048

[211] Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, *26*, 61–68. https://doi.org/10.1016/j.knosys.2011.06.020

[212] Wang, Q., Lai, K. K., & Niu, D. (2011). Green credit scoring system and its risk assessemt model with support vector machine. *Proceedings - 4th International Joint Conference on Computational Sciences and Optimization, CSO 2011*, 284–287. https://doi.org/10.1109/CSO.2011.143

[213] Wang, S., Yin, S., & Jiang, M. (2008). Neural networks based on evolutional algorithm for residential loan. *Chinese Control and Decision Conference, 2008, CCDC 2008*, 2516–2520. https://doi.org/10.1109/CCDC.2008.4597778

[214] Wang, X., Tian, X., & Cheng, Y. (2007). Value approximation with least squares support vector machine in reinforcement learning system. *Journal of Computational and Theoretical Nanoscience*, *4*(7–8), 1290–1294.

[215] Wei, G., & Mingshu, C. (2013). A new dynamic credit scoring model based on clustering ensemble. *Proceedings of 2013 3rd International Conference on Computer Science and Network Technology*, 421–425. https://doi.org/10.1109/ICCSNT.2013.6967144

[216] Wei, L., Wei, L., Li, J., Li, J., Chen, Z., & Chen, Z. (2007). *Credit Risk Evaluation Using Support Vector Machine with Mixture of Kernel*. *Lecture Notes in Computer Science*. Retrieved from http://www.springerlink.com/index/n9442431w75u2u2g.pdf

[217] Weiss, G. M. (2004). Mining with Rarity: A Unifying Framework. *SIGKDD Explorations*, *6*(1), 7–19. https://doi.org/10.1145/1007730.1007734

[218] Wendel, C., & Harvey, M. (2003). Credit Scoring: Best Practices and Approaches. *Commercial Lending Review*, *18*(3), 4. https://doi.org/10.1080/17516230902734536

[219] Weng, C. G., & Poon, J. (2006). A data complexity analysis on imbalanced datasets and an alternative imbalance recovering strategy. *Proceedings - 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings), WI'06*, 270–276. https://doi.org/10.1109/WI.2006.9

[220] West, D. (2000). Neural network credit scoring models. *Computers and Operations Research*, *27*(11–12), 1131–1152. https://doi.org/10.1016/S0305-0548(99)00149-5

[221] West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers and Operations Research*, *32*(10), 2543–2559. https://doi.org/10.1016/j.cor.2004.03.017

[222] Wiginton, J. C. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *The*

Journal of Financial and Quantitative Analysis, *15*(3), 757–770. https://doi.org/10.2307/2330408.

[223] Williams, G. (2008). *Use R ! Data Mining with Rattle and R, The Art of Excavating Data for Knowledge Discovery*. (R. Gentleman, K. Hornik, & G. G. Parmigiani, Eds.), *Springer Science+Business Media*. Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA. https://doi.org/10.1007/978-0-387-78171-6

[224] Wilson, T. C. (1998). *Portfolio Credit Risk*. *FEDERAL RESERVE BANK of NEW YORK ECONOMIC POLICY REVIEW*. https://doi.org/10.2139/ssrn.1028756

[225] Xiao, H., Xiao, Z., & Wang, Y. (2016). Ensemble classification based on supervised clustering for credit scoring. *Applied Soft Computing Journal*, *43*, 73–86. https://doi.org/10.1016/j.asoc.2016.02.022

[226] Yao, P. (2009). Comparative study on class imbalance learning for credit scoring. In *Proceedings - 2009 9th International Conference on Hybrid Intelligent Systems, HIS 2009* (Vol. 2, pp. 105–107). https://doi.org/10.1109/HIS.2009.133

[227] Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications*, *34*(2), 1434–1444. https://doi.org/10.1016/j.eswa.2007.01.009

[228] Zekic-Susac, M., Sarlija, N., & Bensic, M. (2004). Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models. *26th International Conference on Information Technology Interfaces*, *1*, 265–270. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1372413

[229] Zhang, D., Huang, H., Chen, Q., & Jiang, Y. (2007). A comparison study of credit scoring models. *Proceedings - Third International Conference on Natural Computation, ICNC 2007*, *1*(Icnc), 15–18. https://doi.org/10.1109/ICNC.2007.15

[230] Zhang, D., & Xu, W. (2013). A Data-Distribution-Based Imbalanced Data Classification Method for Credit Scoring Using Neural Networks. In *2013 Sixth International Conference on Business Intelligence and Financial Engineering* (pp. 557–561). https://doi.org/10.1109/BIFE.2013.116

[231] Zhang, D. Z. D., Hifi, M., Chen, Q. C. Q., & Ye, W. Y. W. (2008). A Hybrid Credit Scoring Model Based on Genetic Programming and Support Vector Machines. *2008 Fourth International Conference on Natural Computation*, *7*, 8–12. https://doi.org/10.1109/ICNC.2008.205

[232] Zhang, D., Zhou, X., Leung, S. C. H., & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, *37*(12), 7838–7843. https://doi.org/10.1016/j.eswa.2010.04.054

[233] Zhang, G. P. (2000). Neural networks for classification: a survey. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, *30*(4), 451–462. https://doi.org/10.1109/5326.897072

[234] Zhang, Y., Jia, H., Diao, Y., Hai, M., & Li, H. (2016). Research on Credit Scoring by Fusing Social Media Information in Online Peer-to-Peer Lending. *Procedia Computer Science*, *91*(Itqm), 168–174. https://doi.org/10.1016/j.procs.2016.07.055

[235] Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perception neural networks for credit scoring. *Expert Systems with Applications*, *42*(7), 3508–3516. https://doi.org/10.1016/j.eswa.2014.12.006

[236] Zheng, Y., & Heagerty, P. J. (2004). Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics*, *5*(4), 615–632. https://doi.org/10.1093/biostatistics/kxh013

[237] Zheng, Z., Wu, X., & Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 80. https://doi.org/10.1145/1007730.1007741

[238] Zhong, H., Miao, C., Shen, Z., & Feng, Y. (2014). Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings. *Neurocomputing*, *128*, 285–295. https://doi.org/10.1016/j.neucom.2013.02.054

[239] Zhou, H., Wang, J., Wu, J., Zhang, L., Lei, P., & Chen, X. (2013). Application of the hybrid SVM-KNN model for credit scoring. *Proceedings - 9th International Conference on Computational Intelligence and Security, CIS 2013*, 174–177. https://doi.org/10.1109/CIS.2013.43

[240] Zhou, L., & Lai, K. K. (2009). Weighted LS-SVM credit scoring models with AUC maximization by direct search. *Proceedings of the 2009 International Joint Conference on Computational Sciences and Optimization, CSO 2009*, *2*, 7–11. https://doi.org/10.1109/CSO.2009.333

[241] Zhou, S., & Gan, J. Q. (2004). Mercer Kernel Fuzzy C-Means Algorithm and Prototypes of Clusters. In *in: Proc. of Conf. on Internat. Data Engineering and Automated Learning* (Vol. 3177, pp. 613–618).

[242] Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2011). Statistical Methods in Diagnostic Medicine. *Statistical Methods in Diagnostic Medicine*, 1–545. https://doi.org/10.1002/9780470906514

[243] Zhuang, Y., Xu, Z., & Tang, Y. (2015). A Credit Scoring Model Based on Bayesian Network and Mutual Information. In *2015 12th Web Information Sys*