World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

[redacted]

Machine learning for credit risk analysis across the United States

Sevidzem Simo Yufenyuy [1,*], Sulaimon Adeniji [2], Emmanuel Elom [1], Somto Kizor-Akaraiwe [3], Abdul-Waliyyu Bello [1], Emmanuel Kanu [1], Oluwadamilola Ogunleye [4], Juliet Ogutu [1], Chinenye Obunadike [5], Valentine Onih [6] and Callistus Obunadike [1]

[1] *Department of Computer Science and Quantitative Methods, Austin Peay State University, Clarksville, USA.*
[2] *Department of Computer Science, University of Lagos, Lagos State Nigeria.*
[3] *School of Law, University of Washington, Seattle, USA.*
[4] *School of Business, George Washington University, Washington DC, USA.*
[5] *Department of Geology, Anambra State University Uli, Anambra State Nigeria.*
[6] *Department of Computer Science, University of Hertfordshire, Hatfield, United Kingdom.*

World Journal of Advanced Research and Reviews, 2024, 22(02), 942–955

Publication history: Received on 03 April 2024; revised on 11 May 2024; accepted on 13 May 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.22.2.1455

Abstract

Money-Lending financial institutions face risk, which necessitates adopting a robust framework to manage it effectively. While traditional methods have been applied across the financial industry, the advent of artificial intelligence offers organizations the opportunity to utilize advanced methods to manage credit risk. This paper focuses on the application of machine learning techniques for credit risk analysis. Secondary data on information related to borrowers was extracted from Kaggle database simulating Credit Bureau data. Two ensemble models in random forest and gradient boosting were adopted for this study. The findings showed that percentage of income for loan repayment, borrower's income, and interest rates on loans are the most important features for determining defaulters. Furthermore, the evaluation results revealed that both the random forest and the gradient boosting algorithms performed well, with F1 scores of 92.9% and 93% respectively. It was recommended that financial institutions should priorities the verification and comprehensiveness of their data, as precise data is essential for developing resilient models.

Keywords: Machine Learning; Algorithm; Credit; Credit Risk; Credit Risk Analysis; Financial Institution

1. Introduction

Artificial intelligence has become an integral part of human lives, which has brought significant changes to how businesses operate globally (Noriega, Rivera, & Herrera, 2023). Deep within the financial industry is the adoption of artificial intelligence for numerous banking activities, which helps to streamline tasks, reducing workloads on employees. As an extension to existing benefits derived from artificial intelligence is the current trend to predict loan eligibility for banks customers, particularly institutions, whose primary service is lending to household and individuals (Bitetto et al., 2023). Lending activities are an important component in the financial industry, especially microfinance and conventional banks. Through credit facilities, financial institutions tends to make the largest percent of their profit, which has aided these organizations to remain in business over many decades (Shi et al., 2022).

Credit risk analysis has been part of the financial industry over the years, which has enabled recovery of loans and identifying customers that are not credit-worthy (Naili & Lahrichi, 2022). However, prior to the advent of artificial intelligence, financial institutions employed different traditional methods, particularly expert judgement to determine the credit worthiness of potential borrowers (Shan & Nilsson, 2018). Analysts in financial institutions conducts credit risks assessments, examines financial ratios, and develop credit scoring models, with the aim to extensively assess the

* Corresponding author: Yufenyuy Simo

capability of the borrowers to repay loan and feasibility of the lender to recoup funds given out, as well as make informed decision (Bitetto et al., 2023). In order to protect financial stability and investor trust, effective credit risk management is crucial, as was demonstrated by the subprime mortgage crisis of 2007 and the global financial crisis that followed in 2008 (Shi et al., 2022). This suggests that it is imperative for financial institutions to decide on individuals and businesses to grant credit facilities.

While credit risks management using traditional methods have been helpful, financial institutions keeps high records of non-performing loans, necessitating adopting other techniques to tackling this growing concern (Suhadoinik et al., 2023). New types of financial intermediation have emerged as a result of recent legislative changes and technological developments, which have increased competition and decreased lending costs. Bhatore et al. (2020) notes that central banks are pushing for initiatives to streamline and modernize the financial system. These initiatives include the introduction of digital currencies issued by the central bank and the expansion of open banking to facilitate the flow of data and services. By fostering change, value creation, stability, and transparency, this environment supports the deployment of credit financial technology (Olufemi et al., 2023).

The use of machine learning techniques has become increasingly important in credit risk assessments due to the transformation of credit judgments brought about by the boom in computer technologies and data volumes (Noriega et al., 2023). Financial institutions cannot operate without using suitable credit risk analysis tools. Nevertheless, worries about machine learning algorithms' effects on monetary stability have been voiced due to their opaque and "black box" characteristics (Wang et al., 2020). Accurate credit risk assessment is critical to the system as a whole. Systemic failures could result from a credit risk estimation error. As a result, lenders invest a lot of resources in estimating customer and company creditworthiness in order to create loan plans that minimize risks. In the past, statistical techniques like logistic regression and linear discriminant analysis have been used in credit risk assessments. Large datasets, however, are difficult to handle with these methods (Naili & Lahrichi, 2022).

Large credit datasets and increased processing capacity opened the door for AI-driven credit risk assessment methods like deep learning and classical machine learning. Statistical approaches are less efficient and less versatile than traditional machine learning techniques, such as k-Nearest Neighbour, Random Forest, and Support Vector Machines (Yuan, 2015). Deep learning techniques, in particular, are a crucial area of machine learning that perform better than previous methods in terms of accuracy and efficiency when applied to massive credit risk data lakes. The financial industry relies on machine learning for business intelligence to analyze massive amounts of data, particularly in order to decrease credit risk—the uncertainty associated with payment behaviour (Addo et al., 2018). The capacity of machine learning algorithms to sift through mountains of mixed data in search of insights and patterns that humans might miss is a major perk. The ability of machine learning models to process high-dimensional data and intricate linkages allows them to pick up on subtle but significant indicators of creditworthiness, as noted by (Jillian et al., 2020).

To maintain equity, openness, and responsibility in credit risk assessment, it is critical to handle the ethical implications of machine learning algorithms, which are becoming more and more embedded in decision-making processes (Breeden, 2020). When it comes to credit risk analysis using machine learning, algorithmic bias is a major ethical concern. Unintentionally biased results, especially against marginalized or underrepresented groups, may result from machine learning models trained on historical data that perpetuate biases already present in the data (Obunadike et al. 2023). For example, according to Shan & Nilsson (2018), inequality in access to credit could be worsened if machine learning algorithms were trained on historical lending data that shows biased decision-making practices like redlining or discriminatory loan approvals.

Financial institutions need to have strong data collecting, preprocessing, and model development procedures to reduce algorithmic bias (Oloyede et al., 2023). Implementing post-deployment monitoring systems to detect and resolve biases in real-time, doing comprehensive audits of training data to identify and reduce biases, and using fairness-aware algorithms that prioritize equality and non-discrimination are all part of this (Bhatore et al., 2020). The interpretability and openness of machine learning models is another ethical factor to think about. It is difficult to comprehend the decision-making process behind many machine learning algorithms since they function as "black boxes," such deep neural networks (Obunadike et al., 2023). Stakeholders may not completely understand the elements of credit risk assessments or the grounds for loan rejections due to this opaque process, which can damage confidence and responsibility. Building trust among regulators, consumers, and other stakeholders requires clear and understandable information (Wang et al., 2020).

Given this, it becomes imperative to investigate the application of machine learning techniques in credit risk analysis for financial institutions, aiming to assess their effectiveness, identify potential ethical challenges, and propose strategies to mitigate biases, enhance transparency, and safeguard data privacy.

2. Literature Review

2.1. Credit Risk

The danger that a lender can lose money because a borrower doesn't pay back a loan or doesn't pay their bills on time is called credit risk. Essential for determining the probability of borrower default and controlling related risks, it is a cornerstone of financial management and lending procedures (Tomomewo, Falayi, & Uhuaba, 2023). Many things can affect the level of credit risk, such as the borrower's current financial situation, their credit history, the state of the economy, the direction of their business, and even outside forces. The likelihood of default is typically estimated using conventional methods of credit risk assessment that rely substantially on statistical models, financial ratios, and historical data (Adefabi et al., 2023). Nevertheless, there are certain limitations to these methodologies when it comes to correctly projecting creditworthiness, especially in markets that are always changing and unpredictable (Getahun, Anwen, & Bari, 2015).

Credit risk analysis has been radically altered by recent technological developments, mainly in the area of machine learning, which have provided more complex and data-driven methods (Abedalfattah, 2016). Using supervised and unsupervised learning algorithms, neural networks, ensemble approaches, and other machine learning techniques, banks and other financial organizations may make better use of massive volumes of data to build prediction models for credit risk assessment (Phillips et al., 2023). Overall, these models improve the accuracy and reliability of credit risk assessment by analyzing complicated patterns, identifying non-linear linkages, and adapting to changing market conditions (Emmanuel & Ekwere, 2022).

To provide a more complete picture of the borrower's behavior and creditworthiness, machine learning algorithms can consider a variety of data sources, including structured data, social media, transactional data, and traditional financial metrics (Adegbie & Otitolaiye, 2020). Financial organizations may enhance their decision-making capabilities, credit approval workflow efficiency, and risk management strategy optimization with the help of big data analytics and predictive modeling approaches (Efijemue et al., 2023). It is important to address the challenges that come with using machine learning in credit risk analysis. These include issues with model interpretability, data privacy, algorithmic biases, and regulatory compliance. The financial industry must find a way to responsibly and ethically use these technologies (Oritsegbubemi & Evbayiro-Osagie, 2023).

2.2. Credit Risk Management

An essential part of any financial institution's operations is credit risk management, which seeks to detect, evaluate, lessen, and keep tabs on potential dangers in lending (Johanness, 2022). It includes many plans, regulations, and processes that aim to lessen the impact of losses caused by borrowers not paying their bills on time. The soundness of the financial system as a whole and the continued viability of individual financial institutions depend on efficient methods of managing credit risk (Thakor, 2015).

Strong credit policies and underwriting standards that spell out the requirements for lending and borrower eligibility are one of the main goals of credit risk management (John et al., 2005). Loan terms and conditions, acceptable levels of risk tolerance, and criteria for determining a borrower's creditworthiness are all part of these policies. Financial institutions can reduce the risk of default and keep their loan portfolios in good standing by following these practices (Wilms et al, 2014).

Credit risk management isn't just about establishing guidelines for lending; it also includes thorough risk assessment procedures for determining a borrower's creditworthiness and the dangers that may be involved in a certain loan (Kwabena & Boye, 2014). Analyzing the borrower's repayment capacity, reviewing industrial and economic issues, and doing comprehensive financial studies are common steps in this process. In order to minimize losses and maximize risk-adjusted returns, financial institutions must effectively estimate credit risk before making decisions about loan acceptance, pricing, and structuring (Ibtissem & Bouri, 2013).

Credit risk management also includes preventative actions taken at various points in the credit lifecycle to keep exposures to credit under control (Trueck & Racheev, 2009). Among these measures are techniques for intervening quickly in the event of new hazards, early warning systems to identify when credit quality is declining, and continuous monitoring of the portfolio as a whole. Financial institutions may proactively manage credit risk and reduce losses by keeping a close eye on their exposure to credit and taking the necessary steps to reduce that risk (Bäckman, Giunta, Kalyonge, & Salirwe, 2023).

2.3. Machine Learning Techniques

Machine learning approaches have significantly transformed multiple domains by empowering computers to acquire knowledge from data and provide forecasts or judgments without the need for explicit programming (Obunadike et al., 2023). These techniques utilize algorithms that autonomously identify patterns and connections among datasets, enabling intricate data analysis and prediction tasks (Iorkaa, Barma, & Muazu, 2021). Machine learning is applied in areas such as healthcare for tasks like medical image interpretation, disease diagnosis, medication development, and individualized treatment planning. E-commerce uses recommendation systems that employ machine learning algorithms to propose products or services to users based on their tastes and behavior. Autonomous vehicles utilize machine learning algorithms to analyze sensor data, enabling them to identify objects, traverse surroundings, and make driving judgments (Luo, Desheng, & Dexiang, 2017).

The banking sector has recently utilized machine learning methods as potent instruments for credit risk assessment. Machine learning algorithms are able to autonomously discover correlations and patterns in massive datasets, offering a significant improvement over conventional statistical methods that depend on preconceived notions and assumptions (Lessman, 2015). Traditional models may struggle to adequately represent complicated and nonlinear interactions in data; nevertheless, machine learning approaches are well-suited to such situations. Credit risk assessments and decisions can be improved with the use of machine learning algorithms that can detect complex patterns and interactions among the many factors that affect credit risk (Tap, 2011).

Credit risk analysis frequently makes use of supervised learning techniques like support vector machines, decision trees, logistic regression, and random forests to forecast the probability of default or creditworthiness considering past data (Yu & Zhu, 2015). These algorithms can learn from labelled instances of previous credit outcomes and use data about new borrowers' traits and financial situations to place them in one of several risk categories (Callistus et al., 2024). Ensemble approaches, like boosting and bagging, can use many base learners to make predictions more accurate and resilient. As a whole, these methods improve the models' generalizability and decrease the likelihood of overfitting by combining the predictions of different models (Zhao et al., 2015).

Discovering hidden patterns or abnormalities in credit data is possible using both supervised and unsupervised learning algorithms. Unsupervised learning approaches include clustering and anomaly detection (Clement et al., 2024). These methods can help you find clusters of borrowers with similar credit histories or spot suspicious activity that could be a sign of fraud or credit risk (Hens & Tiwari, 2012). Because they develop hierarchical data representations automatically, deep learning models like neural networks also provide superior capabilities for credit risk assessments. The ability of deep learning algorithms to extract complex properties from raw data paves the way for more accurate risk assessments and better prediction capabilities (Taheri & Lundgren, 2023).

2.4. Theoretical Review

This theoretical review examines the application of Signal Detection Theory and Modern Portfolio Theory in the context of machine learning techniques for credit risk analysis. Both theories provide fundamental frameworks for understanding decision-making under uncertainty and optimizing risk-return profiles in portfolio management.

2.4.1. Signal Detection Theory

Signal Detection Theory (SDT) is a theoretical framework that was initially developed in psychology but has since been widely applied in other fields such as economics and finance. SDT offers a methodical approach to comprehending how individuals make decisions when faced with uncertainty. It emphasizes the capacity to distinguish between significant signals, like a reliable credit risk, and insignificant noise, such as an unreliable credit risk. Within the realm of credit risk analysis, SDT (Statistical Decision Tree) can be utilized to augment comprehension of the application of machine learning methods in discerning between individuals or companies that are creditworthy and those who are not. SDT posits that decision-making comprises two fundamental processes: signal detection, which entails identifying the existence of a signal, and decision-making, which involves interpreting the signal and selecting an option.

In the context of credit risk analysis, the term "signal" refers to the accurate measure of an individual or entity's creditworthiness. On the other hand, the term "noise" refers to various elements, such as incomplete or erroneous data, that might hinder the proper assessment of creditworthiness (Olisah et al., 2024). Machine learning algorithms are essential in this process since they analyze past data to identify trends that indicate creditworthiness. SDT additionally presents the notions of sensitivity, which refers to the capacity to accurately detect signals, and bias, which pertains to the inclination to favour one response over another. Sensitivity is a critical factor in credit risk analysis as it impacts the model's capacity to accurately identify persons who are creditworthy and prevent misclassifying them as negative credit risks. However, prejudice can result in errors in judgement, such as exhibiting a preference for conservative lending procedures that reject credit risks that may be good.

2.4.2. Modern Portfolio Theory

Modern Portfolio Theory (MPT) is a fundamental concept in finance that fundamentally transformed the way investors analyze and manage risk and return. Harry Markowitz developed Modern Portfolio Theory (MPT) in the 1950s. MPT is a method for creating investment portfolios that seek to maximize returns based on a specific level of risk, or minimize risk based on a specific level of return. MPT places significant emphasis on diversification as a fundamental method for effectively managing risk. Diversifying assets across many asset classes allows investors to mitigate the influence of any individual asset's performance on the total portfolio. This concept is based on the notion that assets do not uniformly move in the same direction simultaneously. By mixing assets with distinct return patterns, the portfolio as a whole can attain a more consistent return.

The efficient frontier is a fundamental concept in Modern Portfolio Theory (MPT). It refers to a collection of optimal portfolios that provide the highest projected return for a specific level of risk, or the lowest risk for a specific level of return. Portfolios located on the efficient frontier are deemed efficient due to their ability to offer the optimal balance between risk and return. Within the realm of credit risk analysis, Modern Portfolio Theory (MPT) can be utilized to create portfolios of loans or credit instruments that maximize the risk-return balance. Machine learning methods can be employed to evaluate the creditworthiness of individual borrowers. Subsequently, Modern Portfolio Theory (MPT) can be utilized to merge these individual evaluations into a diversified portfolio that optimizes returns based on a certain level of credit risk, or minimize credit risk for a given level of return.

2.5. Empirical Review

In order to assess credit risk for small- and mid-sized businesses (SMBs), Bitertto et al. (2023) compare two methods: a non-parametric method that uses a historical random forest (HRF) model calibrated using machine learning techniques, and a traditional parametric method that uses an ordered probit model. Based on a distinct dataset of detailed quarterly firm-level data from 464 Italian SMBs, the analysis was conducted between 2015 and 2017 and was gathered from an international insurance business and a European investment bank. The HRF methodology performs better than the conventional ordered probit model, according to the results, proving the usefulness of sophisticated estimating techniques utilising machine learning for predicting SMB credit risk, particularly in situations with significant information asymmetries. To further enhance the predictive power of the models, the study also uses Shapley values to evaluate the significance of each variable in forecasting SMB credit risk.

Noriega et al. (2023) highlighted the increasing importance for financial institutions to adopt Artificial Intelligence (AI) and Machine Learning (ML) for credit risk assessment, leveraging vast amounts of data. The study addresses key research questions regarding ML algorithms, metrics, datasets, and variables in credit risk prediction, focusing on the microfinance industry. Challenges such as the use of black box models, the need for explanatory AI, feature selection, multicollinearity, and data imbalance are discussed. The review identifies the Boosted Category as the most researched ML model family, with evaluation metrics such as AUC, ACC, Recall, F1, and Precision being commonly used. While public datasets are predominantly used for model comparison, private datasets are utilized for real-world application.

Shi et al. (2022) evaluated recent advancements in financial artificial intelligence, focusing on machine learning (ML) techniques for credit risk assessment. With a thorough analysis of 76 papers published over the past eight years, the study categorizes ML-driven credit risk algorithms and ranks their performance using public datasets. The findings highlight the superiority of deep learning models over traditional statistical and machine learning approaches in credit risk estimation. Additionally, ensemble methods are identified as providing higher accuracy compared to single models. The review also addresses key challenges such as data imbalance, dataset inconsistency, model transparency, and underutilization of deep learning models in credit risk analysis.

Shan & Nilsson (2018) evaluated the effectiveness of various machine learning algorithms for credit risk analysis using a large peer-to-peer lending dataset. The study compares the classification performance of logistic regression, support vector machine (SVM), decision tree, multilayer perceptron neural network (MLP), probabilistic neural network (PNN), and deep learning models. Results indicate that SVM performs the best, followed by decision tree, logistic regression, MLP, PNN, and deep learning, with deep learning showing the poorest performance. These findings contribute to the ongoing debate on the efficacy of different machine learning algorithms in credit risk analysis, highlighting the importance of selecting the appropriate model for specific datasets.

2.6. Research Gap

The empirical reviews provide valuable insights into credit risk assessment methodologies for small- and mid-sized businesses (SMBs), highlighting the superiority of machine learning techniques over traditional approaches. However, the literature lacks exploration into alternative non-parametric methods beyond the historical random forest (HRF) model for SMB credit risk assessment. There is also a need for broader application of AI and ML in credit risk assessment beyond the microfinance industry.

Additionally, the studies underscore the importance of interpretability and explainability of ML models, especially in the context of black box models. Further research is needed to delve into the selection and significance of variables in credit risk prediction models, particularly in addressing multicollinearity and data imbalance. Moreover, there is a gap in research regarding the application of ensemble methods in conjunction with deep learning models to enhance credit risk assessment accuracy. These areas present opportunities for future research to advance the field of credit risk assessment using machine learning.

3. Methodology

The datasets employed in this study is derived from Kaggle, simulating data in the database of Credit Bureau. The following are the variables within the dataset Age, Annual Income, Home ownership, Employment length (in years), Loan intent, Loan grade, Loan amount, Interest rate, Loan status (0 is non-default 1 is default), Percent income, Historical default, Credit history length. Each of these indicators were described based on their respective data type, which was followed by correlation analysis to determine the relationship among the variable s. Furthermore, the data was explored before adopting Random Forest classifier to predict loan defaulters.

3.1. Missing Values

The variables with missing values were handled using the average values for integer and float data types, while the mode was used where appropriate for variables with strings data types. Employment length (31686) and Interest rate (29465) were the indicators with missing values out of 32581 observations. The missing values were filled using Pandas module in Python, more specifically, the fill na method was utilized to make up for black roles to avoid loss of relevant data.

Table 1 Variables Details

Variables	Non-Null Count	Missing Values	Dtype
person_age	32581 non-null	0	int64
person_income	32581 non-null	0	int64
person_home_ownership	32581 non-null	0	object
person_emp_length	31686 non-null	895	float64
loan_intent	32581 non-null	0	object
loan_grade	32581 non-null	0	object
loan_amnt	32581 non-null	0	int64
loan_int_rate	29465 non-null	3116	float64
loan_status	32581 non-null	0	int64
loan_percent_income	32581 non-null	0	float64
cb_person_default_on_file	32581 non-null	0	object
cb_person_cred_hist_length	32581 non-null	0	int64

3.1.1. Ensemble Methods

Ensemble methods combine multiple machine learning models to improve predictive performance. One popular ensemble method is the Random Forest algorithm. It works by creating a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The equation for the Random Forest classifier can be written as:

$$y(x) \ = \ mode\{T(x, \Theta k)\}_{k=1}^{K}$$

Where;

y(x) is the predicted class for input x

T(x,Θ) is the predictor of the k-th tree in the forest

K is the total number of trees in the forest

4. Results and Discussion

The majority, 50.48%, are renters, suggesting a significant portion of the population may have more fluid housing arrangements. On the other hand, 41.26% have a mortgage, indicating a substantial number of individuals who have committed to homeownership. This group is likely to be more financially stable, as they have secured a mortgage, which could reflect positively on their credit risk. A smaller percentage, 7.93%, own their homes outright, indicating a financially secure subset of the population. The remaining 0.33% fall into the "Other" category, which likely includes individuals with unconventional housing situations.

Table 2 Descriptive Analysis of Categorical Variables

Home Ownership	**Frequency**	**Percentages**
RENT	16446	50.48%
MORTGAGE	13444	41.26%
OWN	2584	7.93%
OTHER	107	0.33%
Loan Intent	Frequency	Percentages
DEBTCONSOLIDATION	5212	16.00%
EDUCATION	6453	19.81%
HOMEIMPROVEMENT	3605	11.06%
MEDICAL	6071	18.63%
PERSONAL	5521	16.95%
VENTURE	5719	17.55%
Loan Grade	Frequency	Percentages
A	10777	33.08%
B	10451	32.08%
C	6458	19.82%
D	3626	11.13%
E	964	2.96%
F	241	0.74%
G	64	0.20%
Loan Status		
NON-DEFAULT	25473	78.18%

DEFAULT	7108	21.81%
Historical default	Frequency	Percentage
No	26836	82.37%
Yes	5745	17.63%

Debt consolidation is the most common intent, representing 16.00% of the dataset. This suggests a significant number of individuals are managing multiple debts and are seeking to consolidate them, potentially to simplify their finances and improve their credit management. Education is another prevalent category, representing 19.81% of the dataset. This indicates a substantial portion of individuals are seeking loans for educational purposes, which could indicate a desire to invest in their future earning potential but could also indicate higher debt levels. Home improvement represents 11.06% of the dataset, indicating a portion of individuals are investing in their properties, which could reflect positively on their credit risk. Medical expenses represent 18.63% of the dataset, suggesting a significant number of individuals are facing unexpected medical costs, which could impact their financial stability and credit risk. Personal loans represent 16.95% of the dataset, indicating a diverse range of personal financial needs among the individuals. Venture loans represent 17.55% of the dataset, suggesting a portion of individuals are seeking loans for entrepreneurial or investment purposes, which could indicate a higher risk appetite and potential financial volatility.

The distribution of loan grades in the dataset provides insights into the creditworthiness of the borrowers. Grade A loans, representing 33.08% of the dataset, are typically associated with the lowest risk and highest creditworthiness. Grade B loans, at 32.08%, also indicate a relatively low credit risk. Grades C, D, E, F, and G represent increasingly higher levels of credit risk, with fewer borrowers falling into these categories. The low percentages for grades E, F, and G suggest that a small portion of borrowers are considered high-risk.

The majority, 78.18%, are classified as non-default, indicating that the borrowers have successfully repaid their loans. On the other hand, 21.81% are classified as default, indicating that these borrowers have failed to meet their repayment obligations. This distribution provides important information about the performance of the loans in the dataset and can be used to assess the effectiveness of the credit risk assessment process. A higher percentage of defaults may indicate weaknesses in the assessment process or economic conditions that affect borrower ability to repay.

Table 3 Summary Statistics of the Numerical Data

Variables	**Mean**	**Std. Dev**	**Min**	**Max**
Age	27.73	6.34	20	80
Annual Income	6.61	6.20	4.00	6.00
Employment length	4.78	4.14	0	35
Loan amount	9589.37	6322.08	500	35000
Interest Rates	0.21	0.413	0	1
Percent Income	0.17	0.106	0	0.83
Credit history length	5.80	4.055	2	30

The summary statistics provide valuable insights into the characteristics of borrowers and loans in the dataset, shedding light on their ages, incomes, employment lengths, loan amounts, interest rates, percentage of income used for loan repayment, and credit history lengths. The average age of borrowers is 27.73 years, indicating a relatively young borrower population, with a diverse age range from 20 to 80 years. The average annual income is 6.61 thousand dollars, suggesting a moderate-income level among borrowers, with some variability as indicated by the standard deviation of 6.20. The average length of employment is 4.78 years, with a wide range from 0 to 35 years. The maximum employment length seems to be an outlier and warrants further investigation. The average loan amount is 9589.37 dollars, with a standard deviation of 6322.08, indicating variability in loan amounts. The interest rates also vary, with an average of 0.21 and a standard deviation of 0.413. Borrowers use, on average, 17% of their income to repay loans, with some borrowers using up to 83% of their income for repayment. Finally, the average length of credit history is 5.80 years, suggesting that borrowers have some credit history, which is important for assessing creditworthiness.

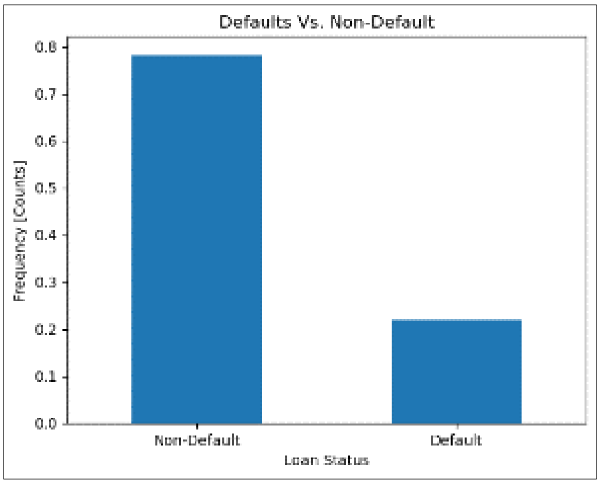

Figure 1 Distribution of Loan Status

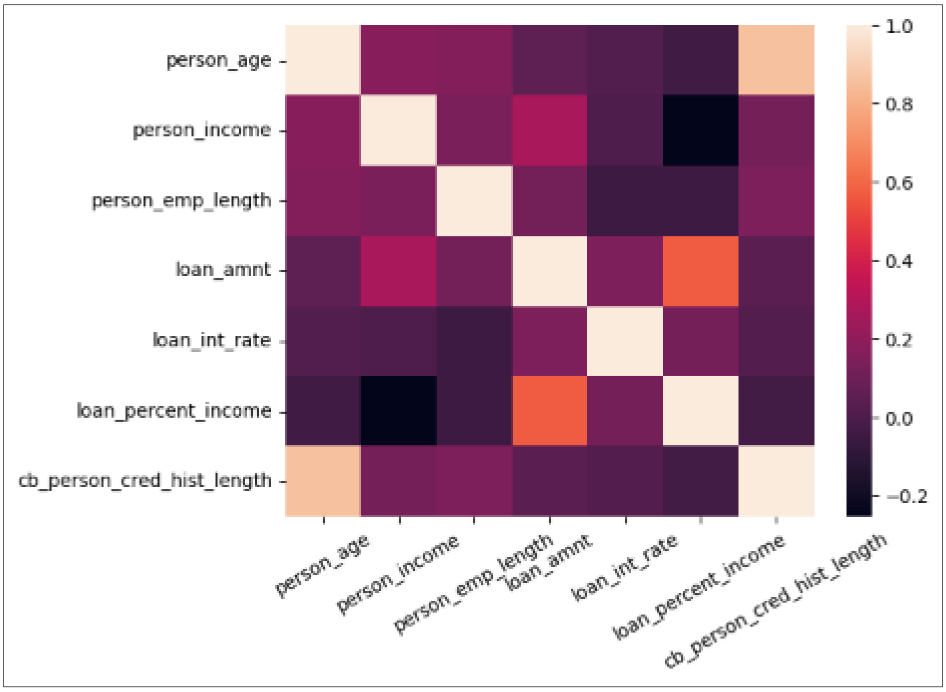

Figure 2 Correlation Analysis Using Heatmap

The heatmap showed the correlation analysis, indicating the relationship between a pair of variables. It was revealed that there is a high correlation between age and credit history length. However, the correlation coefficients across other pairs were low, indicating weak relationship. Regarding the variables with high correlation coefficient, it becomes imperative to drop one variable when building the predictive model to improve performance and avoid obtaining spurious results.

4.1. Machine Learning Analysis

The metrics employed to test the performance of the models and their respective scores are presented below in Table 4. The study evaluated the models using the F1 score, recall values, accuracy, and precision. Accuracy is defined as the proportion of model variables that are correctly predicted relative to the whole dataset. This statistic alone is insufficient to evaluate the model. Thus, the F1 score, accuracy, and recall metrics must be considered as well. Table 4 displays the results of the models that were used.

Table 4 Evaluation Metrics for the Random Forest Model

Evaluation Metrics	Random Forest	Gradient Boosting
Accuracy	0.932	0.934
Precision	0.935	0.935
Recall	0.932	0.934
F1 Score	0.929	0.930

The Random Forest shows an accuracy score of 93.2%, while the gradient boosting shows an accuracy score of 93.4%. Additionally, random forest and gradient boosting both have a precision score 93.5%, indicating that the true positives predicted in both models are 93.5%. Both models also demonstrated a recall score of 93.2% and 93.4% respectively, indicating that the actual positives cases correctly identified by the random forest and gradient boosting is 93.2% and 93.4%. Lastly, the F1-scores in the random forest and gradient boosting are 92.9% and 93% respectively.

4.1.1. The Confusion Matrix

The confusion matrix in Figure 3 presents the random forest model's performance based on how the targets are classified. There are two categories predicted in the confusion matrix, which include default and non-default, relating to the ability of the ensemble model to predict credit risk. The matrix displays the category distribution based on the prediction between the true label and the predicted label. It was revealed that 1555 instances of "Default" class were correctly classified, while 53 instances of the "Non-Default" class were incorrectly classified as "Default". Furthermore, 607 instances were incorrectly classified as "Non-Default", while 7560 instances of the "Non-Default" class were correctly classified. The evaluation of the model revealed that the model performed well as significant cases of defaults and non-defaults were correctly predicted. The F1-score of 92.9% indicated an overall excellent performance of the model.

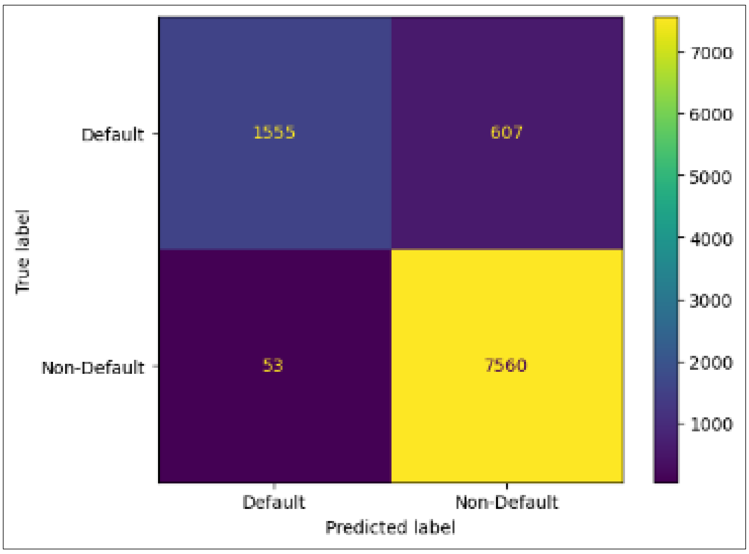

Figure 3 Random Forest Confusion Matrix

The confusion matrix in Figure 4 presents the gradient boosting model's performance based on how the targets are classified. There are two categories predicted in the confusion matrix, which include default and non-default, relating to the ability of the ensemble model to predict credit risk. The matrix displays the category distribution based on the prediction between the true label and the predicted label. It was revealed that 1588 instances of "Default" class were correctly classified, while 74 instances of the "Non-Default" class were incorrectly classified as "Default". Furthermore, 574 instances were incorrectly classified as "Non-Default", while 7539 instances of the "Non-Default" class were correctly classified.

The evaluation of the model revealed that the model performed well as significant cases of defaults and non-defaults were correctly predicted. The F1-score of 93% indicated an overall excellent performance of the model.

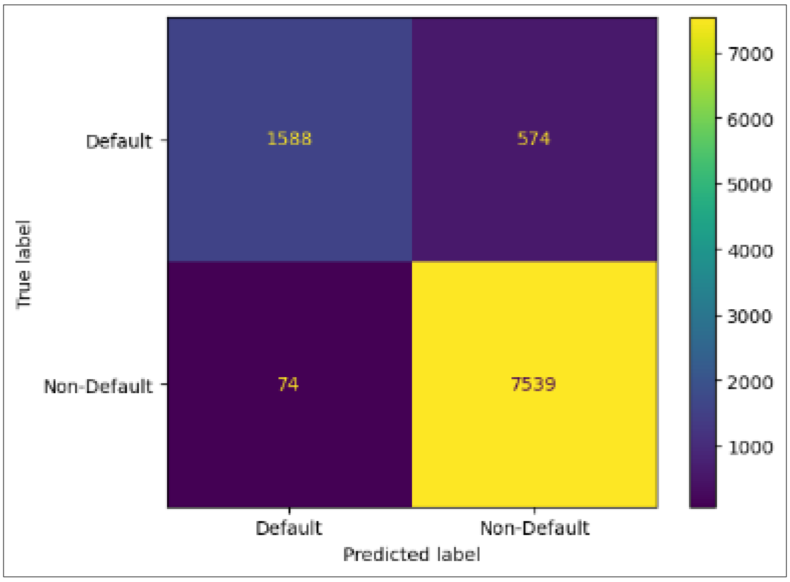

Figure 4 Gradient Boosting Confusion Matrix

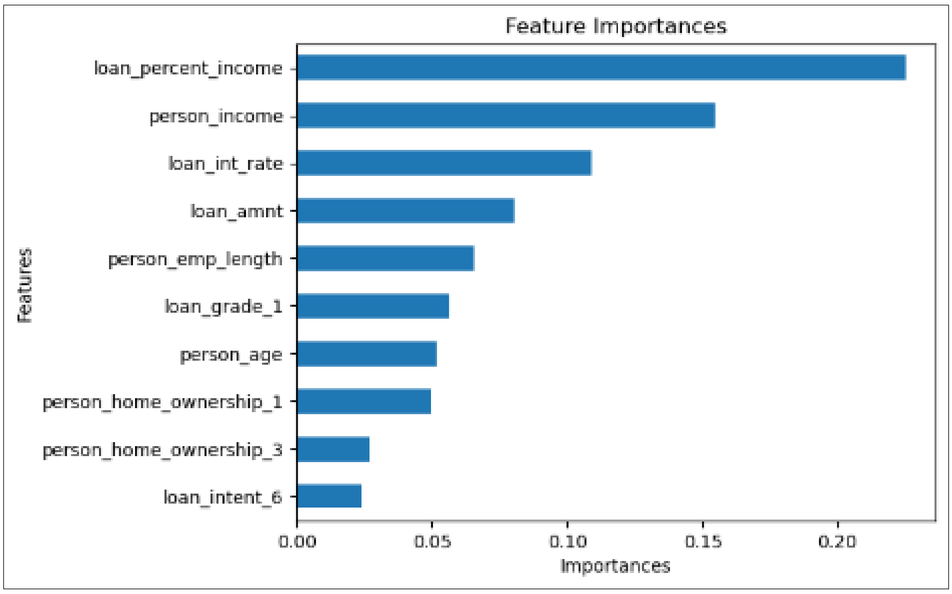

Figure 5 Importance Feature

Based on the outcome of the ensemble models, it could be seen that credit risk can be managed effectively in the financial industry with machine learning techniques. Among the most important features are percentage of income for loan repayment, annual income of borrower, interest rate, loan amount, and how long the borrower has been in employment. Hence, it was concluded that the indicators identified as being important can be leveraged by financial institutions to predict defaulters, thereby making credit risk management effective.

5. Conclusion and Recommendation

Credit risk has continued to be a menace in the financial industry, which could lead to bankruptcy of an organization if not properly managed. Managing the risks associated with credit facilities cannot be overemphasized, as it is necessary for the survival of deposit money banks and microfinance institutions, among other categories of businesses in the financial industry. It is on this backdrop that the paper investigated the application of machine learning techniques to credit risk management. Utilizing credit bureau data simulated on Kaggle with 32581 observations. Each indicator in the datasets was described, and correlation analysis was conducted to determine the relationship between a pair of variables. Ensemble models of random forest and gradient boosting were employed as the machine learning techniques for the credit risk analysis. Both techniques adopted performed excellently in predicting borrowers that are likely to default and those that are not defaulters. Given this, it could be said that financial organizations can adopt these models in order to improve their respective credit risk management.

In order to optimize the efficiency of these methods, financial institutions should priorities the verification and comprehensiveness of their data, as precise data is essential for developing resilient models. Institutions should also consider the utilization of ensemble approaches, such as Random Forests or Gradient Boosting Machines, to merge the advantages of numerous models and enhance predictive performance. Consistent model monitoring and updating are crucial to maintain the effectiveness and accuracy of the models in response to evolving risk profiles. Moreover, it is crucial to priorities interpretability and transparency, as these factors are essential for adhering to regulations and gaining the trust of stakeholders. Utilizing machine learning methods can offer useful insights and enhance the precision of credit risk evaluation. However, it necessitates meticulous evaluation of data quality, selection of appropriate models, and continuous monitoring.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

Statement of informed consent

Informed consent was obtained from all individual participants included in the study.

References

[1] References

[2] Abedalfattah, Z. (2016). Factors Affecting Credit Risk: An Empirical Study of the Jordanian Commercial Banks. European Scientific Journal December 2016 edition vol.12, No.34, 01-15.

[3] Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. Economics Research Paper, 8(1), 1-32.

[4] Adegbie, F., & Otitolaiye, E. (2020). Credit Risk and Financial Performance: An Empirical Study Of Deposit Money Banks In Nigeria. European Journal of Accounting, Auditing and Finance Research Vol.8, No.2, 38-58.

[5] Adefabi, A., Olisah, S., Obunadike, C., Oyetubo, O., Taiwo, E., and Tella, E. (2023) Predicting Accident Severity: An analysis of factors affecting Accident Severity using Random Forest Model. International Journal on Cybernetics & Informatics (IJCI) pp. 107-121 Vol.12, No.6, December 2023 https://doi.org/10.5121/ijci.2023.120609

[6] Bäckman, E., Giunta, V., Kalyonge, J., & Salirwe, M. (2023). Risky Business: The Intersection of Sustainability and Credit Risk Assessment – a Strategic Perspective. 01-76.

[7] Bhatore, S., Mohan, L., & Reddy, R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology, 4(1). doi:http://dx.doi.org/10.1007/s42786-020-00020-3

[8] Bitetto, A., Cerchiello, P., Filameni, S., Tanda, A., & Tarantino, B. (2023). Machine learning and credit risk: Empirical evidence from small- and mid-sized businesses. Journal of Socio-Economic Planning Sciences, 90, 101746.

[9] Breeden, J. L. (2020). Survey of Machine Learning in Credit Risk. 1-60.

[10] Clement, T., Obunadike, C., Ekweli, D C., Ejiofor, O E., Ogunleye, O., Yufenyuy, S S., Nnaji, C I., Obunadike, C J. (2024) Cyber Analytics: Modelling the Factors Behind Healthcare Data Breaches for Smarter Security Solutions. International Journal of Advance Research, Ideas and Innovations in Technology, 10(1), Paper ID V10I1-1197, 49-75.https://IJARIIT.com/V10I1-1197

[11] Efijemue O. P., Obunadike, C., Olisah S., Taiwo E., Kizor-Akaraiwe, S., Odooh, C., and Ejimofor, I., (2023): CyberSecurity Strategies for Safeguarding Customer's Data and Preventing Financial Fraud in the United States

Financial Sectors. International Journal on Soft Computing (IJSC). 14(.3) August 2023 https://doi.org/10.5121/ijsc.2023.14301

[12] Emmanuel, I., & Ekwere, R. (2022). Credit risk and performance of banks in Nigeria. International Journal of Research in Finance and Management 2022; 5(1): 16-24.

[13] Getahun, T., Anwen, L., & Bari, S. (2015). Credit Risk Management and Its Impact on Performance of Commercial Banks: In of Case Ethiopia. Research Journal of Finance and Accounting, 01-12.

[14] Hens, A., & Tiwari, M. (2012). Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. Expert Systems with Applications, 6774-6781.

[15] Hibbeln, M., Kopp, R., & Urban, N. (2021). Credit Risk Modeling in the Age of Machine Learning. 1-76.

[16] Ibtissem, B., & Bouri, A. (2013). Credit Risk Management In Microfinance: The Conceptual Framework. ACRN Journal of Finance and Risk Perspectives Vol. 2, Issue 1, 9-24.

[17] Iorkaa, A., Barma, M., & Muazu, G. H. (2021). Machine Learning Techniques, methods and Algorithms: Conceptual and Practical Insights. International Journal of Engineering Research and Applications 11(8), 55-64.

[18] Jillian, M. C., Di, X., Nooshin, Y., & Dmitry, E. (2020). Sequential Deep Learning for Credit Risk Monitoring with Tabular Financial Data. Journal of Risk Management, 15330.

[19] Johanness, R. (2022). Machine Learning in Credit Risk Management. 1-142.

[20] John, B., Kwak, S., & Smith, B. (2005). The Real Output Losses Associated with Modern Banking Crises. Journal of Money, Credit and Banking 37, 977-99.

[21] Kwabena, M., & Boye, A. (2014). Credit Risk Management in Financial Institutions: A Case Study. Research Journal of Finance and Accounting Vol.5, No.23, 67-86.

[22] Lessman, S. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 2473-2480.

[23] Luo, C., Desheng, W., & Dexiang, W. (2017). A deep learning approach for credit scoring using credit default swap. Engineering Applications of Artificial Intelligence, 65, 465-470.

[24] Naili, M., & Lahrichi, Y. (2022). Banks' credit risk, systematic determinants and specific factors: recent evidence from emerging markets. Heliyon, 8(2), e08960.

[25] Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine Learning for Credit Risk Prediction: A Systematic Literature Review. Journal of Data, 8(11), 169.

[26] Nyangena, B. O. (2019). Consumer credit risk modelling using machine learning algorithms: A comparative approach [Thesis, Strathmore University]. http://suplus.strathmore.edu/handle/11071/6789. 1-56.

[27] Obunadike, C., Adefabi, A., Olisah, S., Abimbola, D. and Oloyede, K. (2023) Application of Regularized Logistic Regression and Artificial Neural Network model for Ozone Classification across El Paso County, Texas, United States. Journal of Data Analysis and Information Processing, 11(3) 217-239 August 2023. https://doi.org/10.4236/jdaip.2023.113012

[28] Obunadike, S., Obunadike, C., Oloyede, K., Phillips, A., Olisah, S., Taiwo, E. and Yusuf, B. (2023): Geographic Information System as a Tool for Effective Counter Terrorism across the United States. International Journal of Information Technology (IJIT). 2(3), August 2023 https://doi.org/10.5121/ijit.2023.10401

[29] Olisah, S., Odooh, C., Efijemue, O., Obunadike, E., Onwuchekwa, J., Owolabi, O., Akintayo, S. and Obunadike, C. (2024) Unveiling Global Human Trafficking Trends: A Comprehensive Analysis. Journal of Data Analysis and Information Processing, 12, 49-75.https://doi.org/10.4236/jdaip.2024.121004.

[30] Oloyede, K., Ajibade, I., Obunadike, C., Phillips, A., Shittu, O., Taiwo, E., and Kizor-Akaraiwe, S., (2023): A Review of Cybersecurity as an Effective Tool for Fighting Identity Theft across United States. International Journal on Cybernetics & Informatics (IJCI). 12(5), https://doi.org/10.5121/ijci.2023.120504

[31] Olufemi, C. Obunadike, A. Adefabi and D. Abimbola (2023): Application of Logistic Regression Model in Prediction of Early Diabetes Across United States. International Journal of Scientific and Management Research. 6(05)34-48. http://doi.org/10.37502/IJSMR.2023.6502.

[32] Oritsegbubemi, K., & Evbayiro-Osagie, E. (2023). Credit Risk Management and the Financial Performance of Deposit Money Banks: Some New EvidenceOritsegbubemi. Journal of Risk and Financial Management 16(7):302, 01-23.

[33] Ouaiid, A., & Moumoun, L. (2023). Application of Machine Learning Techniques for Credit Risk Management: A survey. 1-11.

[34] Phillips, A., Ojalade, I., Taiwo, E., Obunadike, C. and Oloyede, K. (2023): Cyber-Security Tactics in Mitigating Cyber-Crimes: A Review and Proposal. International Journal on Cryptography and Information Security (IJCIS). 13(2/3), https://doi.org/10.5121/ijcis.2023.1320

[35] RAAB, J. (2022). Machine Learning in Credit Risk Management: Advanced Default Risk and Credit Ratings Modeling. Dissertation, 1-142.

[36] Shan, Q., & Nilsson, M. (2018). Credit risk analysis with machine learning techniques in peer-to-peer lending market. International Journal of Risk Management, 4(2), 1-43.

[37] Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine Learning-Driven Credit Risk: A Systematic Review. Neural Computing and Applications, 34, 14327-14339.

[38] Suhadoinik, N., Ueyama, J., & Da Silva, S. (2023). Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach. Journal Risk Financial Management, 16(12), 496.

[39] Taheri, S., & Lundgren, J. (2023). Performance Benchmarking and Cost Analysis of Machine Learning Techniques: An Investigation into Traditional and State-Of-The-Art Models in Business Operations. 01-13.

[40] Tap, B. O. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Systems with Applications, 13274-13283.

[41] Thakor, A. (2015). The Highs and the Lows: A Theory of Credit Risk Assessment and Pricing through the Business Cycle. Journal of Financial Intermediation, 15-29.

[42] Tomomewo, A., Falayi, I., & Uhuaba, O. (2023). Credit Risk Management as a Determinant of Non-Performance Loan of Deposit Money Banks in Nigeria. Journal of Finance and Accounting. Vol. 11, No. 6, 179-188.

[43] Trueck, S., & Racheev, S. (2009). Rating Based Modeling of Credit Risk: Theory and Application of Migration Matrices. Academic Press Advanced Finance Series, Elsevier Inc, 23-44.

[44] Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning ——a case study of bank loan data. Procedia Computer Science, 174, 141-149.

[45] Wilms, P., Swank, J., & Haan, J. (2014). Determinants of the Real Impact of Banking Crises: A Review and New Evidence. Working Paper No. 437, De Nederlandsche Bank NV, 01-27.

[46] Yu, J., & Zhu, Y. (2015). A data-driven approach to predict default risk of loan for online Peer-to-Peer(P2P) lending. 2015 Fifth International Conference on Communication Systems and Network Technologies, 3508-3516.

[47] Yuan, D. (2015). Application of Machine Learning: Consumer Credit Risk Analysis. Journal of Engineering and Computer Science, 2(3), 1-56.

[48] Zhao, Z., Xu, S., Kang, B., Kabir, J., & Liu, Y. W. (2015). Investigation and improvement ofmulti-layer perceptron neural networks for credit scoring. Expert Systems with Applications , 3508-3516