FINANCIAL RISK MANAGEMENT THROUGH STATISTICS AND BUSINESS ANALYSIS

Shashank Dhyani
Department of Electronics and Communication Engineering,
DIT University, Dehradun, India
sdhyani008@gmail.com

Abstract: **Identifying the level of client reliability and timely loan repayment is one of the crucial aspects of credit risk assessment. Based on these factors, the credit history of the customer is analyzed and assessed. Credit scoring is one of the widely used methods for determining the risks associated with issuing a loan, or rather, the possibility of its non-repayment. The information submitted in loan applications is used to create the customer score, which determines the associated risk levels. It eliminates human factors, increases objectivity, speeds up, smoothes out, and lowers risk regardless of how it is calculated or what factors are taken into account. Most businesses rely on financial risk management since it reduces losses and increases profits. Because the task heavily relies on data-driven decision making, machine learning is a rich source for new strategies and technologies. The application of machine learning approaches to various risk management tasks has increased recently. However, machine learning researchers may find it challenging to comprehend the vast and complex domain knowledge and the quickly evolving literature. The need for more features leads to a higher dimensional difficulty where the data from the thousands of published companies is insufficient to populate this space with a necessary density. Flexibility computational techniques have a propensity to over-fit to training samples as a result of this "annoyance of dimensionality," which prevents them from generalizing to new data. In this paper , we are using different machine learning algorithms and stochastic Bayesian Inference for exploratory data analysis and prediction of risk associated with credit card and banking loan.**

Keywords-risk mnagement, machine learning, data analysis, simulation, statistics, classification

I. INTRODUCTION

Numerous financial organizations and lenders use statistical credit score assessments to determine a person's or a small business' creditworthiness. Lenders consider a borrower's creditworthiness to determine if they are qualified for extra loans and whether they would be likely to default on their debt obligations. Credit scoring is used by lenders to decide whether to provide credit or not after analyzing the risks associated with approving a loan or credit. The credit score of a person affects their capacity to borrow money for a variety of reasons, including mortgages. Due to their inability to analyze vast amounts of data, traditional banking Systems are totally reliant on credit scores and small values. These significantly limit how well institutions can lower risk.

For companies, financial institutions, and governmental organizations in the modern corporate world, financial risk management is an essential ability. In this thesis, we have examined the interactions between various financial risk management strategies for independent businesses and financial firms. Financial risk management is a component of quality control in finance. It's a general expression that refers to a process of detecting, evaluating, and taking action to minimize or totally remove a person's or organization's vulnerability to loss. It is used in a variety of contexts for various enterprises or things. The practice of risk management uses instruments including insurance, hedges, derivative contracts, auditing, swaps, and other similar instruments, as well as some well-known risk measurement techniques like VaR, to manage a variety of risks. Enterprises are changing due to machine learning. As a result, it's imperative that business managers, banks, investors, and other stakeholders have an understanding and intuition about what potential benefits these algorithms could have for their decision-making. something is presented will examine the limitations and possibilities of business-oriented machine learning techniques focusing on the assessment of credit risk in the financial sector. standard techniques for machine learning artificial intelligence (AI), decision trees, AdaBoost, support vector machines, and logistic regression, The study of Gaussian processes and neural networks How effective these techniques are for the analysis's primary objectives are to express corporate health and improve business performance.

This study examines the potential applications of artificial intelligence and machine learning in established financial systems. It examines the numerous risks presently existing in the banking industry as well as how ML may reduce labor expenses.

Thanks to machine learning algorithms, consumers, investment managers, and entire banking organizations can now gain future insights into how the market will change a lot sooner than they could with conventional banking models. Banks and other financial institutions can significantly lower risk levels by utilizing machine learning algorithms to analyze a variety of data sources. In contrast to past approaches, which usually concentrated primarily on vital data like credit score like vector machines, regression models, and other statistical techniques, ML can analyze enormous volumes of personal information to reduce risk. The further paper is organized in this way 1. Introduction, 2, literature review,3. Method used 4. Result,5. Conclusion and future scope.

II. LITERATURE REVIEW

Corporate risk management tools should make it possible for organisations to significantly enhance their risk management. Although in some cases it may be challenging to quantify particular hazards, the process itself can reveal information about the total risk that an organisation faces. Better capital allocation will result as a result of this. better in addition. Capital allocation and the use of derivatives for risk hedging are both very beneficial. Some risks economically cancel one another out, as businesses have learned [1]. At the intersection of finance and deep learning, financial time series forecasting is a popular issue [2], [3] carried performed a detailed analysis of the 2005–2019 literature on financial time series prediction using deep learning. Because their research focused on deep learning, they provided a brief review of machine learning's applications to forecasting financial time series. The authors next looked at specific applications of financial time series prediction, such as stock price and index forecasting, and provided the core ideas and organizational characteristics of eight deep learning models that are frequently used. In [4], also investigated recent advancements in deep learning's applicability for stock market forecasting. Jiang classified several neural network topologies in a non-traditional manner. In a related study area published in [5], the researchers employed distance sum methods, outlier mining, and outlier detection mining to precisely a simulation experiment to identify fraudulent transactions data collection of credit card transactions from a specific commercial bank. One area of data mining called "outlier mining" is mostly utilized in the financial and online industries. Dealing with identifying items that are not connected to the main system, such as the fraudulent transactions. They have adopted traits. depending on the actions of customers and the worth of those they determined the separation between the attributes using

Comparison between the attribute's observed value and its predicted value.

Machine learning has been widely used to handle traditional quantitative finance problems, including return forecasting, risk modeling, and creating the best possible portfolios [6]. In this study, we put a lot of emphasis on how it may be used for financial risk management, which includes risk modeling (the process of evaluating and predicting risk) and risk mitigation [7]. The student has the opportunity to actively engage with the environment in reinforcement learning contexts. To maximize reward across all of its environmental interactions is the learner's goal. Researchers and practitioners have employed reinforcement learning algorithms for risk-optimized dynamic portfolio allocation because they are persistent in analyzing the environment to decide the best course of action [8-10]. Financial institutions (FIs) have collected significantly more data over the past several years due to the expansion of reporting requirements and the production of a substantial volume of high-frequency, unstructured consumer data as a result of the digitization of services. To manage vast amounts of data from many sources and formats while preserving or increasing the granularity of analysis, FIs obviously require more robust analytical tools [11-15]. In figure 1, state-of-the –art method is decocted using diagram.

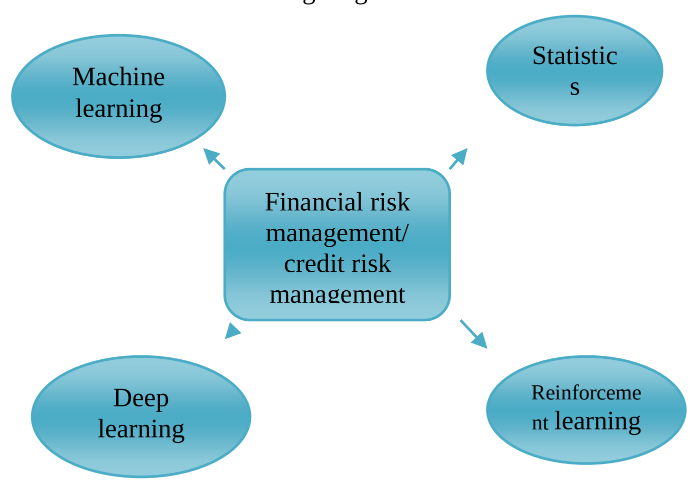

Figure 1: state- of –art- method used.

III. METHODOLOGY

- Before beginning the model generation phase, we initially used JupyterNotebook as a tool for data retrieval and dataset preparation. We next used a variety of categorization approaches.
- The algorithm with the best accuracy was chosen after all classification techniques had been used, and it was implemented on a web application that used the flask framework for server-side scripting.
- A user can interact with the user interface (UI) using flask to provide input values that the model

can process, on the basis of which a prediction is displayed on the UI as shown in block diagram in fig.2.

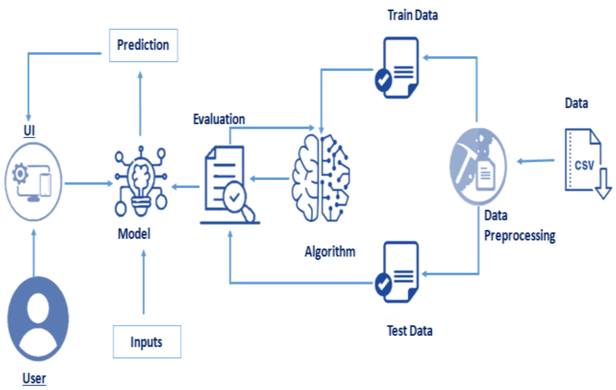

Figure 2: Block diagram

However, the boundaries between these topics are progressively blurring as support vector machines and statistical methods like logistic regression are now covered in almost all machine learning courses. Therefore, for both continuous and discrete outputs, we shall refer to all of these data-driven learning techniques as machine learning methods. The machine learning pipeline, according to Murphy (2012), typically entails the following steps: preprocessing (for example, standardization and centric reduction), dimensionality reduction (for example, principal component analysis), training (for example, learning parameters), model-selection (for example, validation), and testing (e.g., accuracy assessment of the predictions) as shown in figure 3.

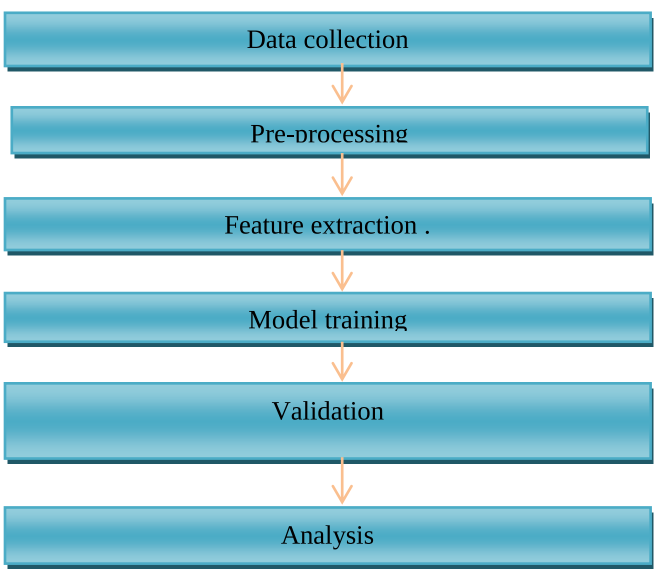

Figure 3: Proposed method

1. Data set The Kaggle dataset used in this investigation. This dataset comprises purchases made over a two-day period by cardholders in September 2013 throughout Europe. In the dataset, there are 31 numerical features. Due to the fact that some of the input variables include financial data, the PCA transformation of these variables was done in order to protect the privacy of the data. For three of them, the specified qualities were left alone. The "Time" feature shows the amount of time that has passed between the dataset's original transaction and each following transaction. The sum of all credit card transactions' monetary values is referred to as the "Amount."

2. Data used for testing: After training with the dataset, testing is carried out.
3. Testing outcomes: The related results for each algorithm will be given, and performance will be displayed in graphs.
4. Exactly right outcomes: Finally, the results of each algorithm are accurately displayed, and the best algorithm is chosen.
5. By taking continuous inputs (such as the Du Pont Ratio, profit margin, efficiency ratio, acid-test ratio, cash ratio, debt ratio, earnings per share, etc.) and producing a continuous output, linear regression can be used to determine the expected profit for the upcoming year. This is already useful for anticipating a corporation's financial catastrophe. If the predicted profit is a big negative number and there isn't enough cash flow to make up the difference, it's probable that the company will go out of business.
6. Naive Bayes: The Bayes theorem calculates the probability that an event will occur given the probability that an earlier event will occur. The naive Bayes algorithm is straightforward and rapid. This method uses relatively little training data and is highly scalable.
7. There are many measures for different algorithms, and these metrics were developed to assess a wide variety of diverse things. It should therefore be a standard for evaluating the many approaches that have been proposed. To assess the accuracy of various strategies, researchers that study credit card fraud detection typically use the words false positive (FP), false negative (FN), true positive (TP), and true negative (TN).
8. The samples used to develop the model are collectively referred to as the "training" data set, while the "test" or "validation" data set is used to assess performance. It's possible that the dataset used to evaluate the effectiveness of the final model has come to be known as the "test set."

IV. RESULT

A. *Statistical result and machine learning result*

The parameter estimates are often comparable to Maximum Likelihood (ML) point estimates, and frequently no attempt is made to capture the estimation uncertainty and explicitly incorporate it into the resulting risk measures; doing so would involve additional, time-consuming financial engineering. Instead, parameter estimates are usually modified heuristically until the results of "back-testing" risk measurements on historical data become "acceptable." This Python notebook demonstrates how Bayesian Inference and Probabilistic Programming (using PYMC3) can be applied for utilizing whatever prior quantitative or qualitative knowledge on market pricing that is available, calculating the parameters of a stochastic process, and naturally accounting for parameter uncertainty in risk assessments.

By applying the empirical mean and standard deviation of the daily logarithmic (or geometric) returns, the parameters are typically computed. The justification for this can be understood by rearranging the equation for $tildeS t$ in the previous sentence as follows:
Log left (frac S t S t-1), then left (mu - frac sigma 2), then right. Delta t with sigma tilde, Delta W t That is implied by $$.
$$log left(frac S t S t-1, right), sim text, normal, left

Delta t, sigma 2 Delta t, right (left(mu - frac)sigma 22) $$ \sWhere, Delta t = frac1365 in dollars $$ as shown in figure 4. Using Probabilistic Programming and Bayesian Inference for Parameter Estimation, Similar to statistical data analysis in general, the main objective of Bayesian data analysis (BDA) is to infer unknown parameters for models of observed data in order to assess hypotheses about the underlying physical processes that produced the observations. Bayesian data analysis differs from classical statistics in a practical sense when it comes to the explicit integration of prior knowledge regarding the uncertainty of the model parameters into the statistical inference process and overall analytic workflow. To accomplish this, BDA focuses on the posterior distribution in figure 5.

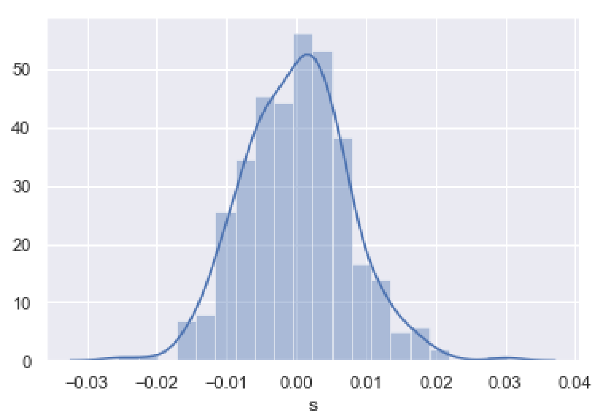

Figure 4: Maximum Likelihood (ML) point estimates

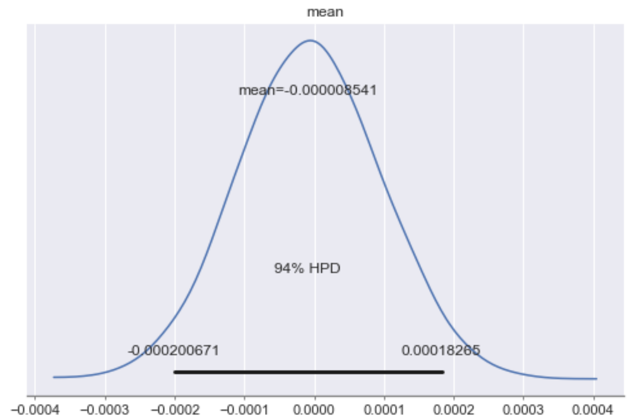

Figure 5: parameter estimation bayesian plot

By applying EDA the risk of several items are shown in figure 6.

14]:	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	67	male	2	own	little	little	1169	6	radio/TV	good
1	22	female	2	own	little	moderate	5951	48	radio/TV	bad
2	49	male	1	own	little	NaN	2096	12	education	good
3	45	male	2	free	little	little	7882	42	furniture/equipment	good
4	53	male	2	free	little	little	4870	24	car	bad
...	...	...	...	...	...	...	...	...	...	...
995	31	female	1	own	little	NaN	1736	12	furniture/equipment	good
996	40	male	3	own	little	little	3857	30	car	good
997	38	male	2	own	little	NaN	804	12	radio/TV	good
998	23	male	2	free	little	little	1845	45	radio/TV	bad
999	27	male	2	own	moderate	moderate	4576	45	car	good

Figure 6: Risk on few products from dataset

Green bars that make up the histogram's "left tail" are visible. In addition, these represent the lowest 5 percent of daily returns (since the returns are ordered from left to right, the worst be always the "left tail"). However, the red bars range from daily losses of 4% to 8%. We can declare with 95% certainty that the worst daily loss will not exceed 4% simply because they are the five percent of daily returns that are the worst. Alternatively, we anticipate that our gain will be greater than -4 percent with a 95 percent confidence level.

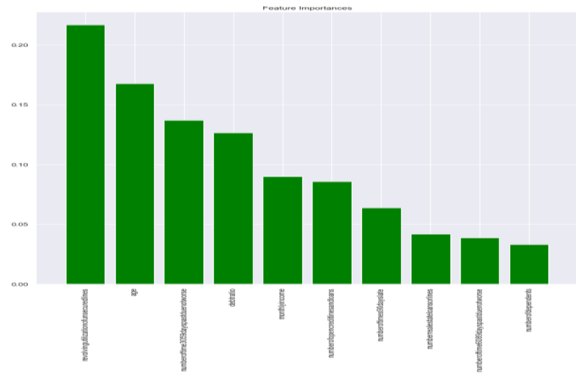

Figure 7: Important features

After examining additional quality control factors, it turns out that less complex machine learning algorithms, such as K-D trees and logistic regression, are actually better at generalising to new data because of the sparse data and numerous features. This is true even though these classifiers have the lowest AUC of all the classifiers we tested. On the test data, they both performed about equally well, with logistic regression performing a little better and requiring much less storage to produce predictions more quickly. As a result, it is the most effective machine learning model for this dataset.

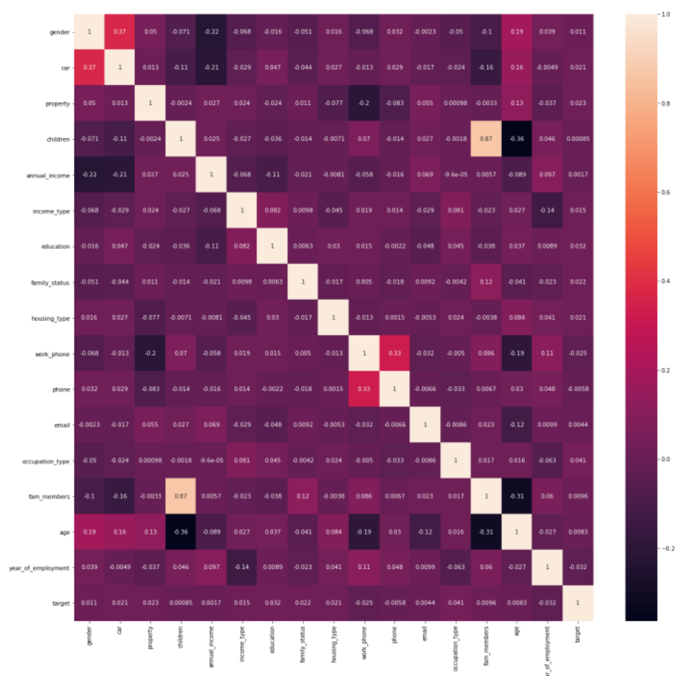

Figure 8: Confusion matrix of correctly predicted labels

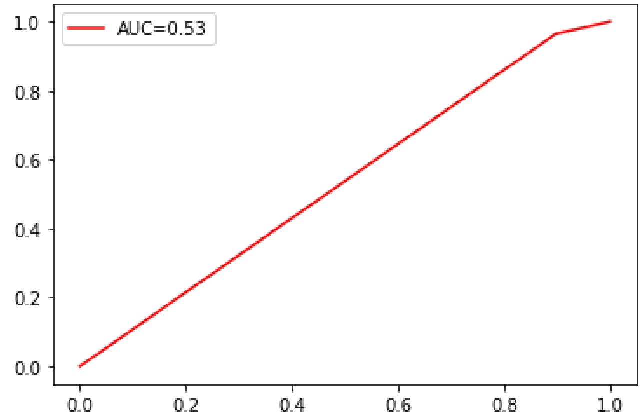

Figure 9: AUC curve of 0.53 correctly predicted labels

In Table 1, because risk analysis have an uneven distribution of data, only a small portion of everyday transactions are fraudulent. In order to address over fitting brought on by uneven data, Chen et al. added data using the SMOTE technique . In order to exclude the accuracy is shown on arias parameters in table 1. e safety samples, they used a classifier based on the kNN algorithm and LSTM to address the issue that SMOTE method contributed noise data to alter the decision of classification border. After confirming the viability and usefulness of the new model, they went on to further suggest the kNN-SMOTE-LSTM credit card fraud detection network model.

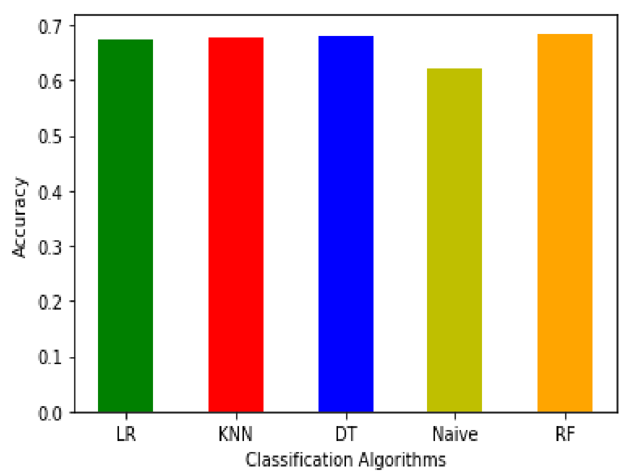

Figure 10 : result of different classifier

Table 1 : Classification summary

Measure	Value	Formula
Sensitivity	0.5349	TPR = TP / (TP + FN)
Specificity	0.4154	SPC = TN / (FP + TN)
Positive Predictive Value (Precision)	0.2323	PPV = TP / (TP + FP)
Negative Predictive Value	0.7297	NPV = TN / (TN + FN)
False Positive Rate	0.5846	FPR = FP / (FP + TN)
False Discovery Rate	0.7677	FDR = FP / (FP + TP)
False Negative Rate	0.4651	FNR = FN / (FN + TP)
Accuracy	0.4451	ACC = (TP + TN) / (TP + TN + FP + FN)
F1 Score	0.3239	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	-0.0434	MCC = (TP x TN – FP x FN) / (sqrt((TP + FP) x (TP + FN) x (TN + FP) x (TN + FN)))

V. CONCLUSION

Financial firms can use our approach to assess the creditworthiness of future consumers. A system could determine how much credit should be offered to a certain consumer by looking at their historical spending habits and tendencies. Millennial and other new customers or those without a long credit history would benefit most from the technology. Banks can improve their overall credit and risk scoring models by mass-automating the processes for credit and risk scoring. The main objective of this research was to create machine learning algorithms that might identify potential future defaulters and so reduce corporate loss. The ideal model would be able to minimize false negatives, identifying every defaulter within the client base, and false positives, preventing clients from being mistakenly classified as defaulters. Since there is a trade-off between recall and precision, increasing the value of one of these metrics typically degrades the value of the other, it can be difficult to achieve these objectives. Because minimizing corporate loss is so important, we decided to put more of an emphasis on removing false positives while searching for the best hyper parameters that may increase the recall rate.

REFERENCES

[1] K. Peng, G. Yan, and School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 510006, China, "A survey on deep learning for financial risk prediction," *QFE*, vol. 5, no. 4, pp. 716–737, 2021, doi: 10.3934/QFE.2021032.

[2] M. Schroeder and S. Lodemann, "A Systematic Investigation of the Integration of Machine Learning into Supply Chain Risk Management," *Logistics*, vol. 5, no. 3, p. 62, Sep. 2021, doi: 10.3390/logistics5030062.

[3] J. C. K. Chow, "Analysis of Financial Credit Risk Using Machine Learning," p. 73.

[4] S P Maniraj, Aditya Saini, Shadab Ahmed, Swarna Deep Sarkar, and SRM Institute of Science and Technology , INDIA, "Credit Card Fraud Detection using Machine Learning and Data Science," *IJERT*, vol. 08, no. 09, p. IJERTV8IS090031, Sep. 2019, doi: 10.17577/IJERTV8IS090031.

[5] J. Galindo, "Credit Risk Assessment using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications," p. 38.

[6] E. V. Orlova, "Decision-Making Techniques for Credit Resource Management Using Machine Learning and Optimization," *Information*, vol. 11, no. 3, p. 144, Mar. 2020, doi: 10.3390/info11030144.

[7] T. Lynn, J. G. Mooney, P. Rosati, and M. Cummins, Eds., *Disrupting Finance: FinTech and Strategy in the 21st Century*. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-02330-0.

[8] D. Mhlanga, "Financial Inclusion in Emerging Economies: The Application of Machine Learning and Artificial Intelligence in Credit Risk Assessment," *IJFS*, vol. 9, no. 3, p. 39, Jul. 2021, doi: 10.3390/ijfs9030039.

[9] D. Mhlanga, "Financial Inclusion in Emerging Economies: The Application of Machine Learning and Artificial Intelligence in Credit Risk Assessment," *IJFS*, vol. 9, no. 3, p. 39, Jul. 2021, doi: 10.3390/ijfs9030039.

[10] A. Mashrur, W. Luo, N. A. Zaidi, and A. Robles-Kelly, "Machine Learning for Financial Risk Management: A Survey," *IEEE Access*, vol. 8, pp. 203203–203223, 2020, doi: 10.1109/ACCESS.2020.3036322.

[11] A. Mashrur, W. Luo, N. A. Zaidi, and A. Robles-Kelly, "Machine Learning for Financial Risk Management: A Survey," *IEEE Access*, vol. 8, pp. 203203–203223, 2020, doi: 10.1109/ACCESS.2020.3036322.

[12] M. Leo, S. Sharma, and K. Maddulety, "Machine Learning in Banking Risk Management: A Literature Review," *Risks*, vol. 7, no. 1, p. 29, Mar. 2019, doi: 10.3390/risks7010029.

[13] M. Leo, S. Sharma, and K. Maddulety, "Machine Learning in Banking Risk Management: A Literature Review," *Risks*, vol. 7, no. 1, p. 29, Mar. 2019, doi: 10.3390/risks7010029.

[14] G. Kou, X. Chao, Y. Peng, F. E. Alsaadi, and E. Herrera-Viedma, "MACHINE LEARNING METHODS FOR SYSTEMIC RISK ANALYSIS IN FINANCIAL SECTORS," *Technological and Economic Development of Economy*, vol. 25, no. 5, pp. 716–742, May 2019, doi: 10.3846/tede.2019.8740.

[15] Z. Shahbazi and Y.-C. Byun, "Machine Learning-Based Analysis of Cryptocurrency Market Financial Risk Management," *IEEE Access*, vol. 10, pp. 37848–37856, 2022, doi: 10.1109/ACCESS.2022.3162858.