7PAM2002-0901-2024 Data Science Project

EVALUATION OF FINANCIAL RISK FOR LOAN APPROVAL USING ML

Student Name: VASANTH KUMAR MEKALA

Student ID: 22058159

Course: MSC Data Science with Advanced Research

Module: Data Science Project

**Table of Content**

[Chapter 1: Introduction](#_2vhqs6vi1fni) 2

[1.1 Background](#_nnxktjn5nh1q) 3

[1.2 Research Aim and Objectives](#_7y5v6351yeja) 3

[1.2.1 Research Aim](#_j5oeq8il2yet) 3

[1.2.2 Research Objectives](#_4qq0966drr1i) 3

[1.3 Research Questions](#_fylr8eyjf5nj) 4

[1.4 Novelty of the Research](#_9zm02f8l3ngv) 4

[1.5 Structure of the Dissertation](#_mfl0zorm4zs3) 5

[Chapter 2: Background](#_d2sqolqiqila) 5

[2.1 Introduction](#_53a9v8l0ax2u) 5

[2.2 Selection of Previous Research Studies](#_bv3gzwj5z8n7) 5

[2.3 Critical Analysis of Previous Research Studies](#_n7svi9xb1ii5) 6

[2.3.1 “Machine Learning for an Enhanced Credit Risk Analysis: A Comparative Study of Loan Approval Prediction Models Integrating Mental Health Data”](#_g04bcg6f77k0) 6

[2.3.2 “Accuracy Comparison between Five Machine Learning Algorithms for Financial Risk Evaluation”](#_pcjg7fi44g8n) 6

[2.3.3 “Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach”](#_10pzhauz9nf8) 7

[2.3.4 “Credit Risk Analysis Using Machine and Deep Learning Models”](#_l5cyyxk78ub) 7

[2.3.5 “Enhancing Supervised Model Performance in Credit Risk Classification Using Sampling Strategies and Feature Ranking”](#_etxe4lwvrqf1) 8

[2.4 Comparing the Articles](#_7zs3h96ppued) 8

[2.5 Research Gap](#_1up4ljdue5pz) 9

[2.6 Conceptual Framework](#_tydtq7290wrm) 9

[2.7 Chapter Summary](#_15shpprvjzbh) 9

[References](#_atit3ioxc2pq) 10

# Chapter 1: Introduction

## 

## 1.1 Background

One of the most critical tasks a bank or financial institution has in terms of loan approval is evaluating the economic risk of an application. The activity is a determining factor in whether the applicant has a chance of repaying his loan or not. According to Orlova (2021), traditionally, checks were done regarding credit score, income, and employment status for approving loans. These methodologies, however, seem highly restricted in their ability to capture the complexity of financial behaviours. Thus, machine learning (ML) as a tool acts quite effectively in increasing both the accuracy and efficiency of such risk evaluations as far as institutions are concerned to deal with huge data volumes or identify patterns that may go unnoticed (Kalogiannidis et al., 2024). According to Orji *et al.* (2022), loan defaults have been reduced, and the approval rate for low-risk applicants has been enhanced using ML algorithms compared to traditional methods. Banks have also employed ML in their evaluation, which allows them to process transactions faster and more effectively.

Recent studies have discussed that data models have reduced credit risks and losses in lower-income segments by approximately 20-50% and doubled their application approval rates (Baer, Goland, and Schiff, 2023). This helps facilitate rapid loan and credit approvals that customers can receive, thereby enhancing customer experience (Viswanatha *et al.*, 2023). For instance, a firm like ZestFinance has illustrated that it is possible to improve the returns for a loan portfolio by as much as 10% by using ML underwriting tools and minimizing the number of defaulting customers (Merrill and Zestfinance, 2019). However, the application of ML in financial risk analysis also includes issues such as complying with regulations and biases in algorithms. Therefore, these issues mainly promote transparency and ethical values as essential to confirm trustworthiness. Hence, this Research helps evaluate the financial risks for loan approval using regression and classification techniques to ensure that ML-based decisions are trustworthy and fair.

## 

## 1.2 Research Aim and Objectives

### 1.2.1 Research Aim

This study aims to develop Machine Learning models for predicting borrowers' credit risks and classifying loan status.

### 1.2.2 Research Objectives

* To evaluate the demographic and financial factors influencing credit risks in the context of the loan approval process.
* To develop Regression models (state-of-the-art Machine Learning models such as Random Forest regressor, Support Vector Regression, and Neural Networks) for predicting continuous risk scores associated with the likelihood of each loan default or financial instability.
* To develop classification models for predicting the loan approval status of applicants, indicating whether an applicant is likely to be approved or denied a loan.
* To recommend effective strategies for financial institutes to minimise credit risks.

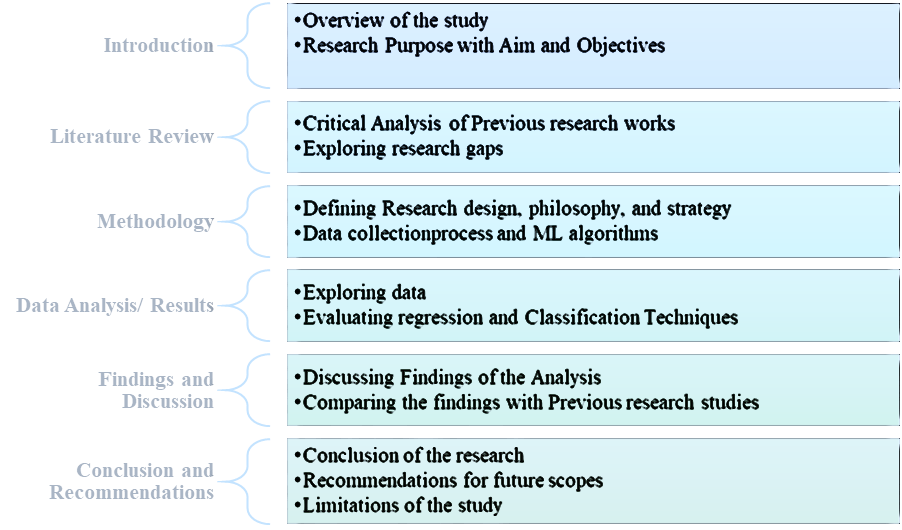
## 1.3 Research Questions

* What are the demographic and financial factors influencing credit risks in the context of the loan approval process?
* How do we develop regression models (state-of-the-art machine learning models such as random forest regressor, support vector regression, and neural networks) to predict continuous risk scores associated with the likelihood of each loan default or financial instability?
* What is the process of developing classification models for predicting the loan approval status of applicants, indicating whether an applicant is likely to be approved or denied a loan?
* What are the recommended effective strategies for financial institutes to minimize credit risks?

## 1.4 Novelty of the Research

The current industry practices regarding the financial risk evaluation to approve loans are still mainly based upon traditional credit scoring models that involve income-to-debt ratios and employment history. However, these methods, apart from their apparent strengths, often need to take into account more subtle, complicated interplays between some specific financial behaviours or trends within the broader economy (Gu *et al.*, 2024). Some organisations have started embracing machine learning algorithms, but their usage still needs to be narrowed in many of these places. For instance, structured financial data has yet to be adopted for most unstructured data, such as social media behaviour or alternative financial indicators (Boukherouaa *et al.*, 2021). Thus, this Research applies machine learning (classification and regression models) to analysing structured and unstructured data, rendering the financial risk assessment more holistic. This study also stands apart from most previous research studies, which were singular models in nature by considering several advanced ML techniques, comprising neural networks, Random Forest regressor, Support Vector Regression, and classification models. The novelty of this study lies in integrating significant data resources in constructing a more robust, fair, and explainable loan approval system.

## 1.5 Structure of the Dissertation



***Figure 1: Structure of the Dissertation***

# Chapter 2: Background

## 2.1 Introduction

The chapter mainly outlines the critical analysis of existing work on financial risk assessment based on machine learning models for loan approval. This is initiated by narrowing the selection of key studies relevant to this project by focusing only on works that apply high-level machine learning algorithms. Therefore, the data, methods, and conclusions of the mentioned studies are critically analysed to explore the gaps and inform the developed conceptual framework for this study

## 2.2 Selection of Previous Research Studies

The articles are selected for critical analysis based on the relevance of machine learning applications in credit risk evaluation. Comparisons of multiple algorithms, exploration of different data, and studies addressing specific challenges such as model accuracy, feature selection, and incorporation of alternative data involving mental health information have been chosen. Additionally, only articles published in peer-reviewed journals are selected to ensure credibility and the usage of robust methodologies. In addition, including selected studies would provide insights toward models of supervised machine learning, its preprocessing techniques, and the difficulties in enhancing credit risk prediction in financial environments that are identified as facets of interest

## 2.3 Critical Analysis of Previous Research Studies

### 

### 2.3.1 “Machine Learning for an Enhanced Credit Risk Analysis: A Comparative Study of Loan Approval Prediction Models Integrating Mental Health Data”

The research study mainly discusses the possibility of including mental health in loan approval predictions. It used two datasets, one containing all mental health data and another involving traditional loan approval data covering credit history and income (Alagic *et al.*, 2024). It compares different machine learning models such as XGBoost, Random Forest, Decision Trees, and KNN, showing that the XGBoost (84%) and Random Forest (85%) had the highest mental health and loan approval data accuracy, respectively.

The introduction of mental health data is new and brings a fresh dimension to credit risk analysis, a domain that has been based on financial indicators thus far. Experimenting with multiple machine learning algorithms, the Research demonstrates how mental health can have predictive value for financial risk, indicating that it may well be worth having comprehensive datasets to improve loan approval accuracy.

Although the results have much potential, one limitation is that specific datasets have been used; the data may only sometimes reflect larger populations or portray all vital risk factors (financial indicators, demographic data, and customer behaviour). The study does not fully address ethical concerns by using mental health data for credit evaluation. Regarding this ongoing project, this paper serves as valuable information regarding comparing various machine learning models on credit risk. However, its limitations in data diversity and ethical consideration are considered crucial issues to pursue.

### 

### 2.3.2 “Accuracy Comparison between Five Machine Learning Algorithms for Financial Risk Evaluation”

This paper attempts to evaluate a non-parametric approach using five machine learning algorithms for credit risk purposes: KNN, Naive Bayes, Decision Tree, Logistic Regression, and Random Forest classifiers. The experiment used German credit data from 1000 observations and Taiwan credit data with 30000 records, divided into an 80:20 train-test split (Dong, Liu, and Tham, 2024). Data preprocessing techniques, such as SMOTE to handle imbalanced data, feature selection, and hyperparameter tuning, were included in the study to optimise model performance. In the present experiment, the results show that Random Forest (AUC= 0.98) was a better learner than others for all experiments, followed by Logistic Regression (AUC=0.71) and Naive Bayes (AUC=0.709). Decision Trees overfit small datasets and underperform with large datasets, whereas Logistic Regression and Naive Bayes were insensitive to hyperparameter tuning and data imbalances.

The methodology adopted in this paper involves different preprocessing steps and classifier evaluation, which are relevant to this ongoing project. However, this ongoing study also involves machine learning and imbalanced datasets. This study’s strength is in it’s comprehensive comparison across accuracy, precision, recall, F1-score, and ROC-AUC. Further, the usage of SMOTE is suitable for dealing with imbalanced data. Still, on the other hand, it needs to be improved to only go a little into the details of feature engineering, where selecting the relevant predictors forms a core idea for this ongoing study. Furthermore, not much exploration of the neural networks or other complex algorithms could provide an opportunity for future work.

### 2.3.3 “Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach”

This Research explores model development for credit risk using machine learning algorithms on a vast dataset from Lending Club. Ten models were tested (Logistic Regression, Decision Trees, XGBoost, KNN, SVM, ANN, Random Forest, Extra Trees, AdaBoost, and Gradient Boosting), which are compared with AUC, precision, and recall (Suhadolnik, Ueyama, and Da Silva, 2023). The dataset involved over 2.5 million loans with 151 variables and was preprocessed with great care before reducing it into 18 key features. For class imbalance, subsampling techniques were applied for the preprocessing step when robust models were under construction. The paper concludes that boosting models, especially XGBoost (accuracy 0.6563), are more accurate and processed faster than the state-of-the-art benchmark models.

This extensive dataset test has conducted a substantive empirical analysis to compare model performance in credit risk assessment. Therefore, their exclusion of macroeconomic variables limits the model from being flexible enough to adapt to many real-world financial scenarios. Even though their approach fits perfectly well with this ongoing project using machine learning in credit risk assessment, this study can be further improved by implementing external factors like economic trends that help improve the robustness of the predictions. Thus, it has been demonstrated that applying ML with credit scoring is promising. However, further analysis concerning the impact of additional features outside borrower-specific characteristics is required.

### 

### 2.3.4 “Credit Risk Analysis Using Machine and Deep Learning Models”

The paper aims to analyse credit risk by developing binary classifiers that are mainly designed to estimate the probability of loan defaults based on data from 117019 enterprises by observing credit risk predictions, monitoring, model reliability, and effective loan processing. The dataset involves financial variables like balance sheets and income statements, with a huge imbalance between good health companies at 98.5% and those defaulting at 1.5%. All datasets are first adjusted by applying SMOTE before applying any models tested- namely, Logistic Regression, Random Forest, Gradient Boosting, and four Deep Learning models (Addo, Guegan, and Hassani, 2018). Results show that the tree-based models, specifically Random Forests and Gradient Boosts, outperformed the deep learning models in terms of AUC (0.99, 0.99) and RMSE (0.09, 0.04) even with only using the top 10 most important variables. This further questions how stable and good-performance deep learning models could be in real-world applications.

The paper relates directly to this ongoing study in which machine learning models evaluate risk factors. However, the need for more exploration of model combinations (hybrids) and even less comparison in deep learning models narrows the scope of conclusions for their study. Additionally, the reliance on financial data ignores all other factors that may impact credit risk, a limitation that the authors narrow down in the possible implications of their results.

### 2.3.5 “Enhancing Supervised Model Performance in Credit Risk Classification Using Sampling Strategies and Feature Ranking”

The paper aims to enhance supervised model performance in credit risk classification using sampling strategies and feature ranking. It evaluated a massive dataset from LendingClub, involving almost 2.93 million records and 141 features between 2007 and 2020, to improve credit risk classification performance with sampling strategies and feature ranking. Using three kinds of supervised machine learning techniques, which involve Logistic Regression, Random Forest, and Gradient Boosting, the authors intended to classify loan statuses as Good or Risk (Wattanakitrungroj *et al.*, 2024). For managing class imbalance, they used under-sampling, over-sampling, and combined sampling, and the best performance result was derived through Gradient Boosting with excellent accuracy, precision, recall, and F1-scores over 99%.

Although this study represents an approach to the prediction of credit risks, there are some limitations. Mutual information can focus only on features with considerable mutual information with the target and potentially ignore other significant relations between variables that may remain unseen, leading to less intense models. Furthermore, the limitation of the temporal-bound dataset may prevent applying the results in the present lending environment. Thus, this study fits well with the credit risk analysis project theme, as it strengthens the claims of preprocessing and how the appropriate features are selected in machine learning. The paper delivers relevant insights but recognises further research areas, specifically advanced feature selection techniques and optimisation strategies for better models.

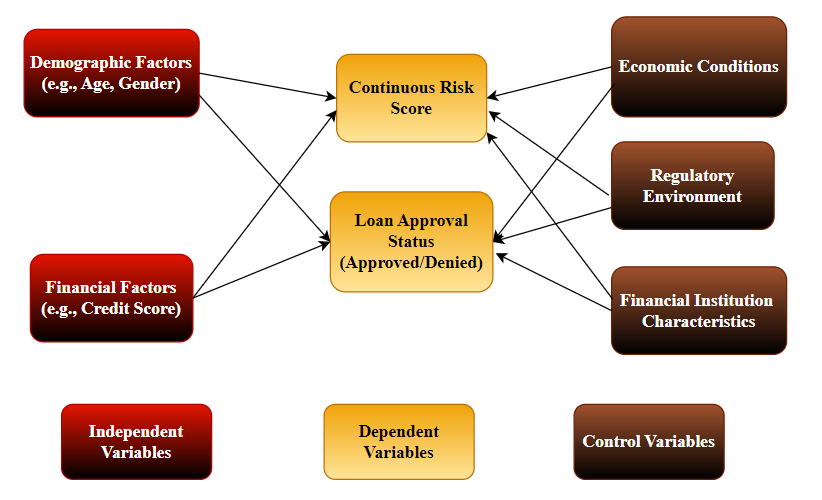
## 2.4 Comparing the Articles

The examined papers offer several opportunities for the further extension of credit risk analysis using machine and deep learning algorithms. Alagic *et al.* (2024) mainly focused on presenting mental health data to underpin its predictiveness of credit loan approval. In contrast, Dong, Liu, and Tham (2024) compared several algorithms by focusing on better efficiency of Random Forest. Similarly, Suhadolnik, Ueyama, and Da Silva (2023) studied a huge LendingClub dataset, where boosting models like XGBoost proved effective but restricted their flexibility by excluding macroeconomic variables. Addo, Guegan, and Hassani (2018) found that tree-based models could do well beyond the deep learning methods with financial variables excluding essential risk factors. Wattanakitrungroj *et al.* (2024) provided a very significant overview concerning feature selection and sampling strategies to be drawn from this analysis. Thus, comparing the studies, Addo, Guegan, and Hassani (2018) provided the best analysis and helped update the current credit risk analysis project by highlighting the requirement for multiple kinds of data and robust model evaluations.

## 2.5 Research Gap

Apart from the progress in machine learning methods in credit risk analysis, significant gaps still exist in the areas of Research. Alagic et al. (2024) and Suhadolnik et al. (2023) focused on specific datasets and failed to address the population at large, which resulted in decreased generalisability. Most significantly, the ethics involving the data are not considered explicitly about mental health. In addition, reliance on traditional financial variables, as recognised by Addo et al. (2018), needs to be a factor that can take the form of external effects such as macroeconomic trends. This could also make a point of using more diverse datasets, robust feature selection methods, and deeper exploration into ethical considerations in credit risk appraisal.

## 2.6 Conceptual Framework



***Figure 2: Conceptual Framework***

## 2.7 Chapter Summary

The chapter reviews previous literature on the application of ML models in assessing financial risk for the approval of loans. Here, it compares key studies, points out gaps in terms of diversity in data and ethical considerations, and claims the need for stronger methodologies along with alternative sources of data to improve credit risk evaluation.

# References

Addo, P., Guegan, D. and Hassani, B. (2018) ‘Credit Risk Analysis Using Machine and Deep Learning Models’, *Risks*, 6(2), p. 38. doi: https://doi.org/10.3390/risks6020038.

Alagic, A. *et al.* (2024) ‘Machine Learning for an Enhanced Credit Risk Analysis: A Comparative Study of Loan Approval Prediction Models Integrating Mental Health Data’, *Machine Learning and Knowledge Extraction*, 6(1), pp. 53–77. doi: https://doi.org/10.3390/make6010004.

Baer, T., Goland, T. and Schiff, R. (2023) *New credit-risk models for the unbanked | McKinsey*, *www.mckinsey.com*. Available at: https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/new-credit-risk-models-for-the-unbanked.

Boukherouaa, E. B. *et al.* (2021) ‘Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance’, *Departmental Papers*, 2021(024). Available at: https://www.elibrary.imf.org/view/journals/087/2021/024/article-A001-en.xml.

Dong, H., Liu, R. and Tham, A. W. (2024) ‘Accuracy Comparison between Five Machine Learning Algorithms for Financial Risk Evaluation’, *Journal of Risk and Financial Management*, 17(2), p. 50. doi: https://doi.org/10.3390/jrfm17020050.

Gu, Z. *et al.* (2024) 'Credit risk assessment of small and micro-enterprise based on machine learning', *Heliyon*, 10(5), pp. e27096–e27096. doi https://doi.org/10.1016/j.heliyon.2024.e27096.

Kalogiannidis, S. *et al.* (2024) ‘The Role of Artificial Intelligence Technology in Predictive Risk Assessment for Business Continuity: A Case Study of Greece’, *Risks*, 12(2), pp. 19–19. doi: https://doi.org/10.3390/risks12020019.

Merrill, D. and Zestfinance, C. (2019) *Testimony to the House Committee on Financial Services AI task force available at*: https://www.congress.gov/116/meeting/house/109735/witnesses/HHRG-116-BA00-Wstate-MerrillPhDD-20190626.pdf.

Orji, Ugochukwu. E. *et al.* (2022) ‘Machine Learning Models for Predicting Bank Loan Eligibility’, *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*. doi https://doi.org/10.1109/nigercon54645.2022.9803172.

Orlova, E. V. (2021) ‘Methodology and Models for Individuals’ Creditworthiness Management Using Digital Footprint Data and Machine Learning Methods’, *Mathematics*, 9(15), p. 1820. doi: https://doi.org/10.3390/math9151820.

Suhadolnik, N., Ueyama, J. and Da Silva, S. (2023) ‘Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach’, *Journal of Risk and Financial Management*, 16(12), p. 496. doi: https://doi.org/10.3390/jrfm16120496.

Viswanatha, V. *et al.* (2023) ‘Prediction of Loan Approval in Banks using Machine Learning Approach’, *International Journal of Engineering and Management Research*, 13(4), pp. 7–19. doi: https://doi.org/10.31033/ijemr.13.4.2.

Wattanakitrungroj, N. *et al.* (2024) ‘Enhancing Supervised Model Performance in Credit Risk Classification Using Sampling Strategies and Feature Ranking’, *Big data and cognitive computing*, 8(3), pp. 28–28. doi https://doi.org/10.3390/bdcc8030028.