A black background with white text

Description automatically generated with low confidence

MSc Data Science Project

7PAM2002-0901-2024

Department of Physics, Astronomy and Mathematics

**Data Science FINAL PROJECT REPORT**

**Project Title:**

EVALUATION OF FINANCIAL RISK FOR LOAN APPROVAL USING ML

**Student Name and SRN:**

VASANTH KUMAR MEKALA, 22058159

Supervisor: Stephen Kane

Date Submitted: 04 JAN 2025

Word Count: 6970

GitHub address: https://github.com/Vasi0796/Vasanth-project

**Declaration**

I VASANTH KUMAR MEKALA declare that the research is done by myself with proper guidance from the professors and supervisors. I would also like to declare that I have given full credit to the authors whose research has been used for getting knowledge.

**Acknowledgement**

I VASANTH KUMAR MEKALA would like to thank my professors and supervisors for their constant guidance and feedback throughout the execution of the project. I would also like to thank my parents for giving me motivation during the coursework as well as projects.

**Abstract**

The purpose of this study was to develop Machine Learning models to predict the likelihood of loan approval and estimate credit risk score of loan applicants. In this context, three different classification models such as Decision Tree, AdaBoost and Stacking Classifier, have been developed using Python programming in Jupyter Notebook to predict likelihood of loan approval. Stacking classifier obtained the highest accuracy (with high recall and precision), reflecting suitability of the model for predicting likelihood of loan approval. Additionally, the estimation of credit risk score of applicants has been predicted using different regression models (such as Random Forest, Gradient Boosting and XGBoost). XGBoost regressor has provided the highest predictive performance, reflecting applicability of the model in the estimation of credit risk scores of loan applicants.

**Keywords*: Machine Learning, Loan Approval, Credit Risk, Credit Score, Stacking Classifier, XGBoost***

**Table of Content**

[Chapter 1: Introduction 7](#_30j0zll)

[1.1 Background 7](#_1fob9te)

[1.2 Research Aim and Objectives 7](#_3znysh7)

[1.2.1 Research Aim 7](#_2et92p0)

[1.2.2 Research Objectives 7](#_tyjcwt)

[1.3 Research Questions 8](#_3dy6vkm)

[1.4 Novelty of the Research 8](#_1t3h5sf)

[1.5 Structure of the Dissertation 9](#_4d34og8)

[Chapter 2: Literature Review 10](#_2s8eyo1)

[2.1 Introduction 10](#_17dp8vu)

[2.2 Selection of Previous Research Studies 10](#_3rdcrjn)

[2.3 Critical Analysis of Previous Research Studies 10](#_26in1rg)

[2.3.1 “Machine Learning for an Enhanced Credit Risk Analysis: A Comparative Study of Loan Approval Prediction Models Integrating Mental Health Data” 10](#_lnxbz9)

[2.3.2 “Accuracy Comparison between Five Machine Learning Algorithms for Financial Risk Evaluation” 11](#_35nkun2)

[2.3.3 “Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach” 11](#_1ksv4uv)

[2.3.4 “Credit Risk Analysis Using Machine and Deep Learning Models” 12](#_44sinio)

[2.3.5 “Enhancing Supervised Model Performance in Credit Risk Classification Using Sampling Strategies and Feature Ranking” 12](#_2jxsxqh)

[2.4 Comparing the Articles 12](#_z337ya)

[2.5 Research Gap 13](#_3j2qqm3)

[2.6 Conceptual Framework 13](#_1y810tw)

[2.7 Chapter Summary 14](#_4i7ojhp)

[Chapter 3: Data 15](#_2xcytpi)

[3.1 Data collection method 15](#_1ci93xb)

[3.2 Description of the Dataset 15](#_3whwml4)

[3.3 Rationale for the selection of the dataset 15](#_2bn6wsx)

[3.4 Data pre-processing technique 15](#_qsh70q)

[3.5 Exploratory Data Analysis 16](#_3as4poj)

[Chapter 4: Ethical Issues 22](#_1pxezwc)

[4.1 Data ethics 22](#_49x2ik5)

[4.2 Ethical Considerations in Research 22](#_2p2csry)

[Chapter 5: Methodology 23](#_3o7alnk)

[5.1 Methodological Architecture and rationale behind the architecture 23](#_23ckvvd)

[5.2 Implementation of Software and Libraries 23](#_ihv636)

[5.3 Machine Learning models applied 24](#_32hioqz)

[Chapter 6: Results 26](#_93xjdfsrj5xi)

[6.1 Introduction 26](#_u6zjz0r6buhi)

[6.2 Selection of features for classification 26](#_1bsmy8ebj1i9)

[6.3 Class balancing using SMOTE 27](#_i7b1tgtt6dwe)

[6.4 Classification models for Loan Approval status 28](#_duk4shtykauq)

[6.4.1 Decision Tree Classifier 28](#_u5r2m92puygl)

[6.4.2 AdaBoost Classifier 30](#_xi04tenrvocm)

[6.4.3 Stacking Classifier 32](#_hga4jifz0wz3)

[6.4.5 Model Comparison 34](#_4xjmgkt908ga)

[6.5 Regression Model for Predicting Credit Risk Score 35](#_6w0q06dn9ago)

[6.5.1 Random Forest Regressor 36](#_wem1dhpgm0c8)

[6.5.2 Gradient Boosting Regressor 38](#_plwwcf160c6s)

[6.5.3 XGBoost Regressor 40](#_ibbyp9s29m9i)

[6.5.4 Model comparison 41](#_jam1ekxrz8qq)

[6.6 Model Deployment for real-world use cases 43](#_5aj373dqtapa)

[6.6.1 Deployment of the Classification Model for Predicting Loan Approval Status 43](#_rjnzjj55rc4m)

[6.6.2 Deployment of the Regression model for predicting Credit Risk Score 44](#_8s233aupqlft)

[6.7 Chapter Summary 46](#_n0p4ek9f11ji)

[Chapter 7: Analysis and Discussion 47](#_lsfyq4rn0fhw)

[7.1 Interpretation of the findings 47](#_59q69t3ehqib)

[7.2 Discussion 47](#_k7dq1ah5rq7a)

[7.3 Limitations 48](#_ko0mv5eafgbq)

[Chapter 8: Conclusion 49](#_7dis9gb13ac)

[8.1 Summary 49](#_y0qvw1774yrd)

[8.2 Linking with Objectives 49](#_hlnq6lbtax6i)

[8.3 Contribution of the research to real-world scenarios 49](#_jh582av7z01u)

[8.4 Future scope 50](#_7p6elbmgaumf)

[References 51](#_1hmsyys)

# 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 act 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 minimising 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 minimise 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 specified in many places for detecting financial risks. 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. It considered 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: Literature Review

## 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 critical 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 pre-processing techniques, and the difficulties in enhancing credit risk prediction in financial environments 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 based on accuracy 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 incorporated by Alagic et al. (2024) is new and brings a fresh dimension to credit risk analysis, a domain 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 about using mental health data for credit evaluation. Regarding this ongoing project, this paper serves as valuable Information for 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 research 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 (Area Under Curve) = 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 since the target variable is not balanced, which might cause the outcomes of the model to be inaccurate and biassed. This study's strength is its comprehensive comparison across accuracy, precision, recall, F1-score, and ROC-AUC. Further, the usage of SMOTE is suitable for dealing with imbalanced data. 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, proper exploration of the neural networks or other complex algorithms could provide an opportunity for future work in understanding factors related to loan approval risks.

### 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 pre-processed with great care before reducing it into 18 key features. For class imbalance, subsampling techniques were applied for the pre-processing 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 mainly designed to estimate the probability of loan defaults based on data from 117019 enterprises by observing credit risk predictions, monitoring, model reliability, and adequate loan processing. The dataset involves financial variables like balance sheets and income statements, with a massive 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. These further question 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 pre-processing 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.

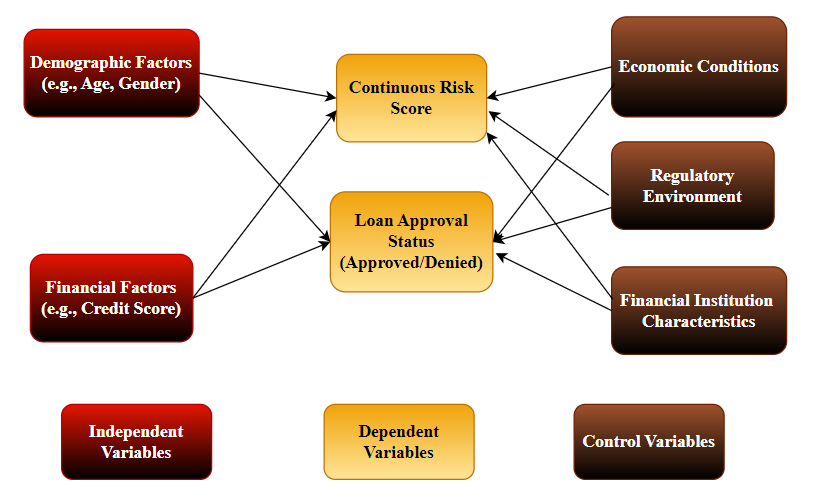
## 2.4 Comparing the Articles

The 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 this area 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 use 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***

The above figure tends to understand the conceptual framework which is a flowchart for this research. With the help of the figure, it can be seen that the dependent variables are loan approval status which would help in understanding whether banks or any other financial institutions can approve loan requests based on factors like Age, Gender, Income status and others. Also, the Continuous risk score is a dependent variable as this helps banks to predict whether the loans asked by a customer are having higher credit score or not which can help in making proper decisions regarding future loan approvals.

## 2.7 Chapter Summary

The chapter reviews previous literature on the application of ML models in assessing financial risk for loan approval. 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.

# Chapter 3: Data

## 3.1 Data collection method

Secondary data aids in the quick access to a humongous amount of Information compared to collecting primary data, considering time and budget constraints. The Research followed the secondary data collection process using a pre-existing dataset from a publicly accessible source named Kaggle (Kaggle, 2024). The dataset is appropriate for use with quantitative data that comes with applicability in statistical analysis and applications in machine learning.

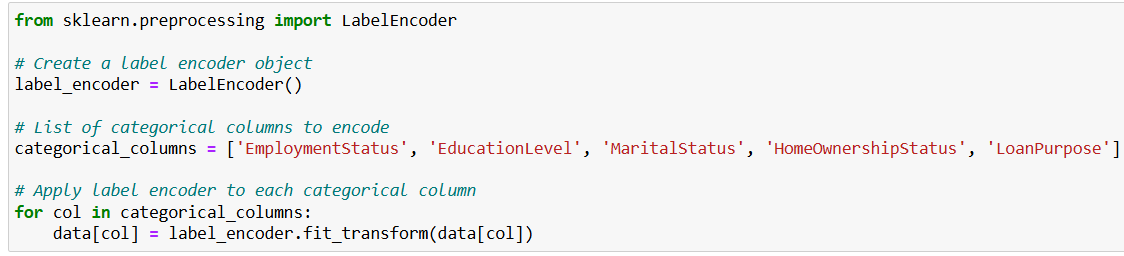
## 3.2 Description of the Dataset

The dataset consists of 20,000 synthetic records corresponding to a model of a loan-granting process with a view toward risk assessment. The key features of the dataset include demographic Information such as *Age, EducationLevel, MaritalStatus, financial metrics* such as *AnnualIncome, DebtToIncomeRatio*, and credit history details such as *BankruptcyHistory, LoanPurpose, CreditScore* (Kaggle, 2024). These are the input variables in the two primary predictive tasks of regression on *RiskScore*, which is a continuous variable with the number indicating the likelihood of defaulting, and the other being a binary classification on *LoanApproved*, signalling whether the loan has been approved or not. This range of features assures a strong foundation for rigorous quantitative loan risk analysis.

## 3.3 Rationale for the selection of the dataset

I have selected this data set because its structure is complete, clearly answering the study's objectives in predicting the risk score and classification into the category of loan approval. According to Mahajan *et al.* (2020), achieving high levels of precision in financial performance forecasting using different financial metrics is very important. The characteristics of the dataset represent a complete view of credit risk due to the extensive involvement of different features such as *CreditScore*, *TotalAssets*, *and TotalLiabilities*. This aligns with the research objective, which includes developing strong machine-learning models for high-accuracy loan risk assessment.

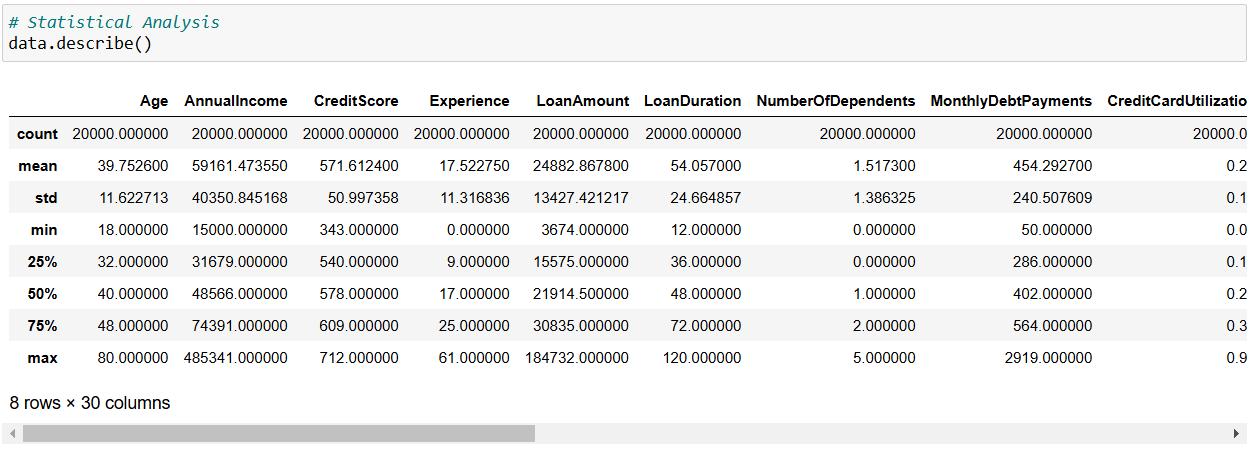
## 3.4 Data pre-processing technique



***Figure 3: Data Preprocessing***

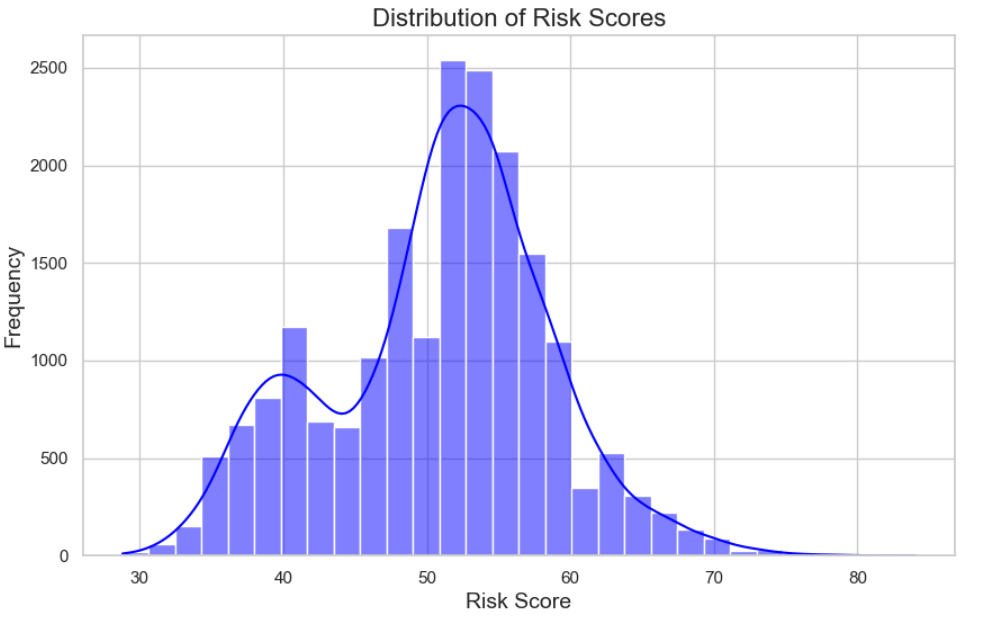
The code in ***Figure 3*** applies label encoding to ‘EmploymentStatus’, ‘EducationLevel’, ‘MaritalStatus’, ‘HomeOwnershipStatus’, and ‘LoanPurpose’, which are categorical. The Label Encoding step converts non-numeric (categorical) data into a numeric format that ML can easily use as input features, as ML models require numerical inputs (Bolikulov et al., 2024). Encoding properly is necessary for proper model training so that it can make predictions on loan approval or financial risk.

## 3.5 Exploratory Data Analysis



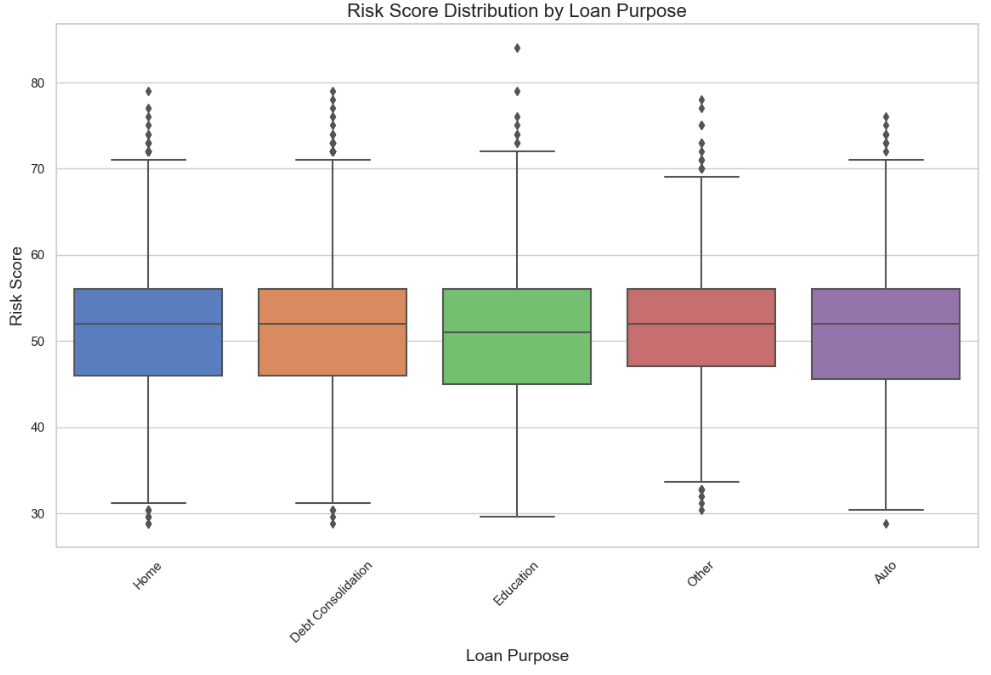
***Figure 4: Statistical Analysis***

The statistical analysis in Figure 5 shows borrower profiles with an average age of 39.75 years and annual income of $59,161 with extreme variation (std $40,351), indicating income disparity among borrowers. This also impacts credit risk assessment precision, and it is ensured that ***Figure 4*** represents the credit score variations between 343 and 712. These help in studying the risk patterns and thus enable the prediction of loan approval based on loan approval models for credit risk evaluation.



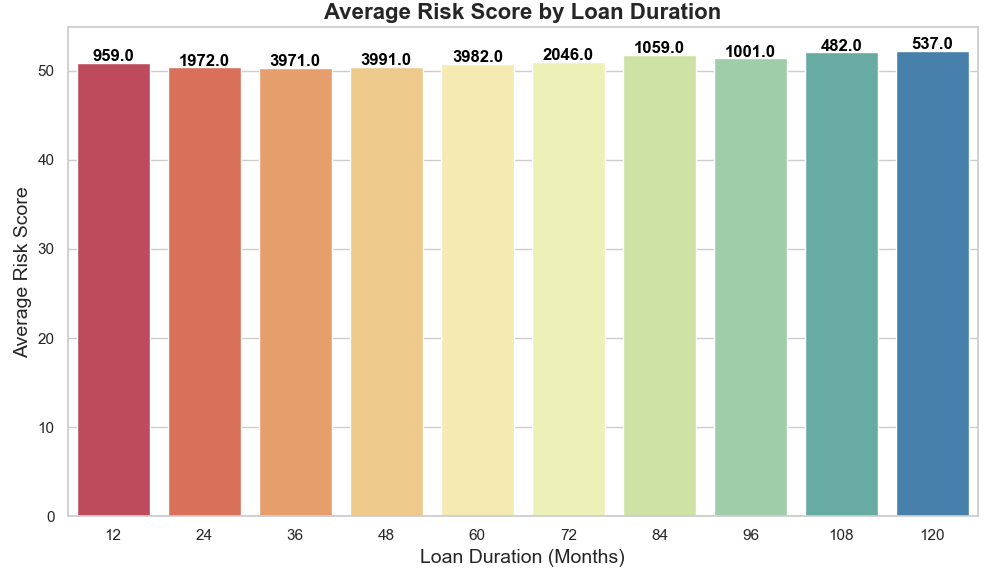
***Figure 5: Risk Score Distribution***

***Figure 5*** shows the risk score distribution that peaks between 50 and 60, illustrating most borrowers fall in the moderate-risk category. This also highlights the balanced lending or limited high-risk applicant approvals by institutions and has followed normal distribution.



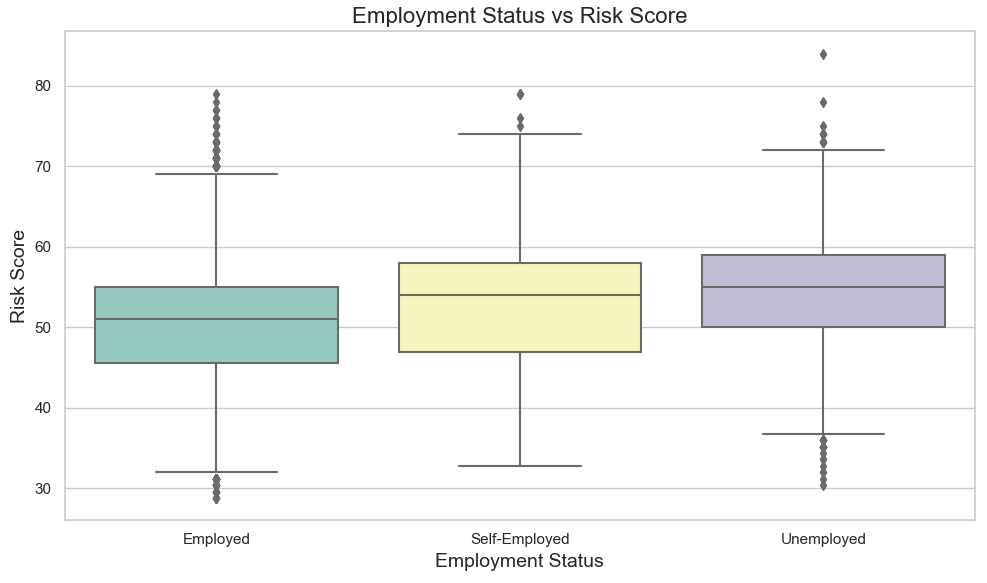
***Figure 6: Risk Score by Loan Purpose***

Boxplots in ***Figure 6*** represent the median risk scores of about 52 for “Home”, “Debt Consolidation”, “Other”, and “Auto” loans, while “Education” has 51 average scores. This indicates financial instability impacts loan type, aiding target risk assessment strategies.



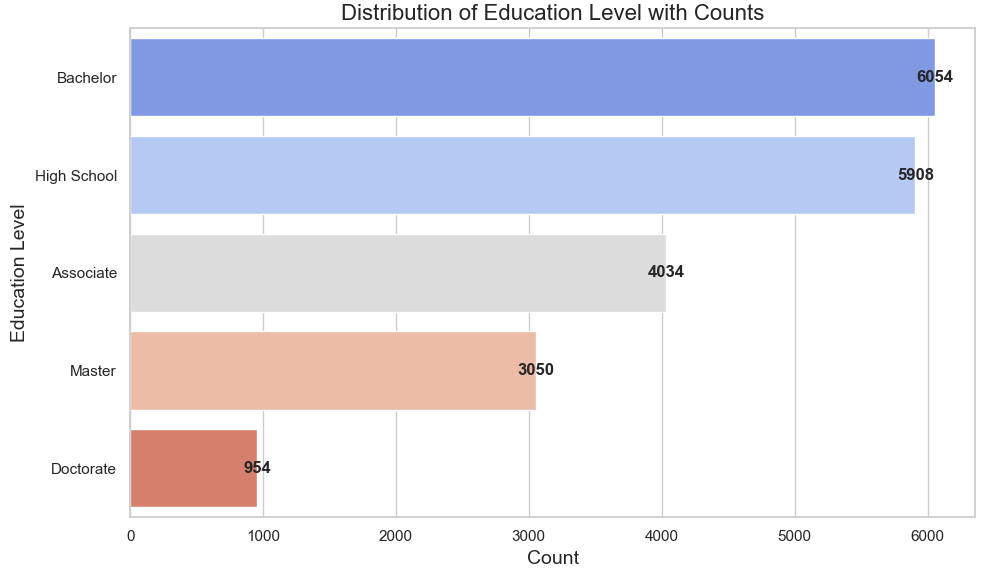
***Figure 7: Average Risk Score by Loan Duration***

***Figure 7*** emphasises that 84-month loans have the highest average risk scores among 1059 borrowers and 96-month loans among 1001 borrowers. This indicates more extended periods, which correlate with higher borrower risks in this context.



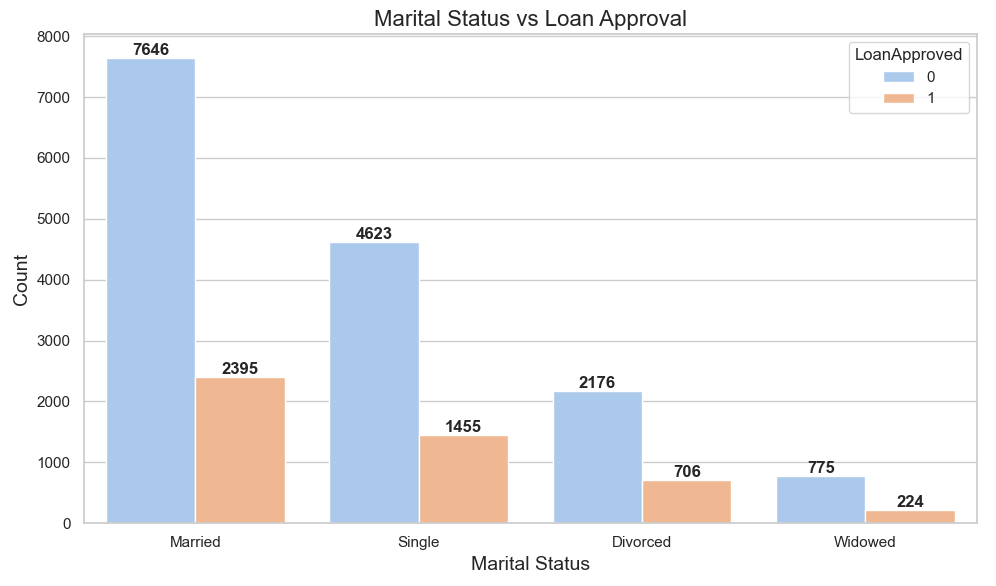
***Figure 8: Employment Status vs Risk Score***

The box plots in ***Figure 8*** represent that median risk scores are 51 for employed people, 55 for the self-employed, and 58 for the unemployed. This suggests unemployment increases perceived financial risk, affecting the probability of loan approvals.



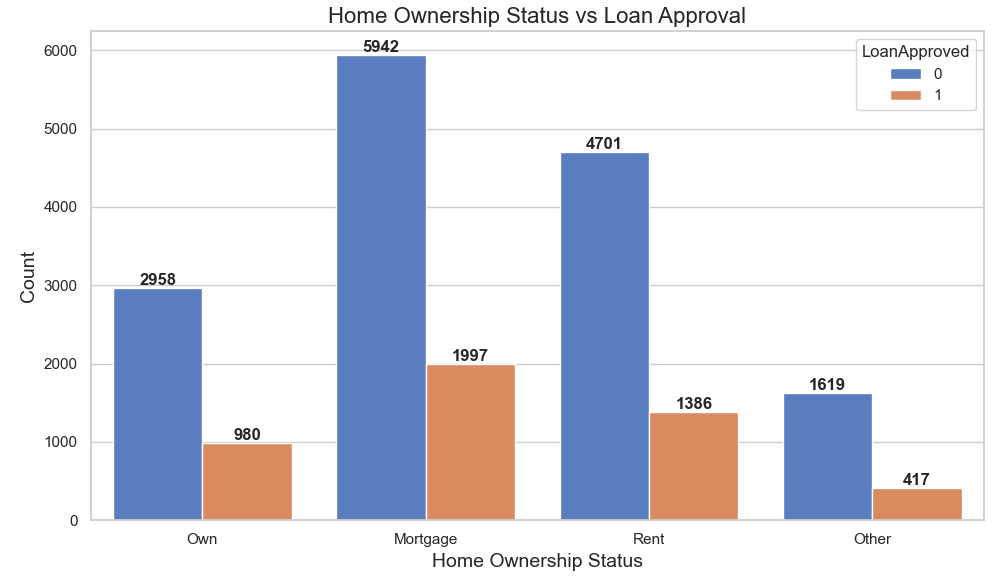
***Figure 9: Distribution of Education Level***

***Figure 9*** demonstrates that most borrowers hold a bachelor's degree (6054) or high school degree (5908), while few hold Master’s degrees (3050) and Doctorates (954). Education influences creditworthiness as it reflects the earning ability and stability of borrowers.



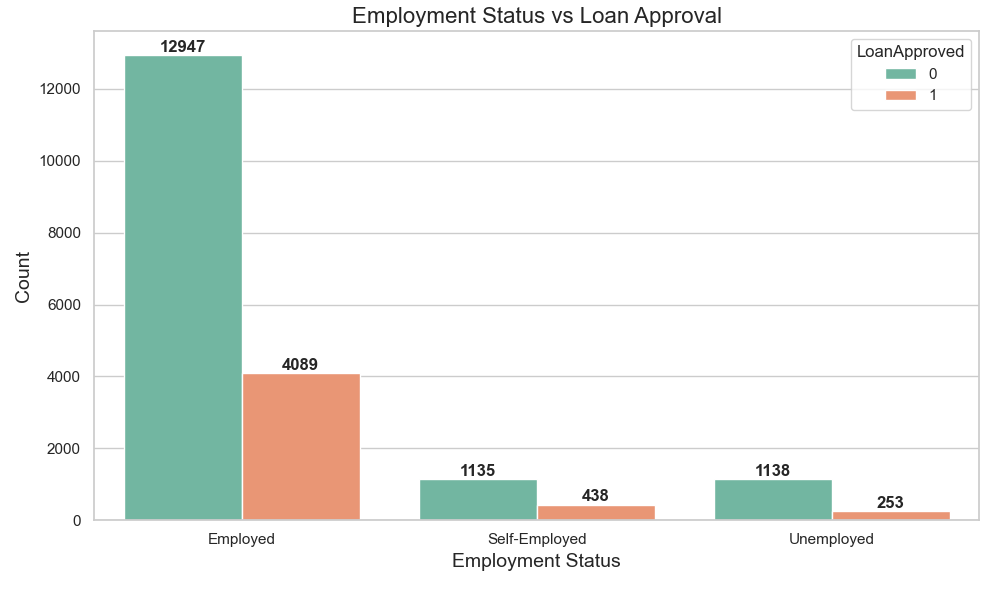
***Figure 10: Marital Status vs Loan Approval***

Loan approval rates by marital status in ***Figure 10*** show that married (7646 ejected, 2395 accepted) and single (4623 rejected, 1455 accepted) account for most loan applications. This indicates financial stability and income factors that can influence lender acceptance.



***Figure 11: Home Ownership vs Loan Approval***

Homeowners with mortgages (5942 denied, 1997 approved), and renters (4701 denied, 1386 approved) face different levels of approval. Mortgage holders may be perceived as stable, while renters may appear to be at a greater risk.



***Figure 12: Employment Status vs Loan Approval***

***Figure 12*** implies the approved loan has the highest rate in employed persons, with 12947 denied and 4089 approved, and the lowest is in unemployed, with 1138 denied and 253 approved, showing the extensive impact of job stability on loan approvals.

# Chapter 4: Ethical Issues

## 4.1 Data ethics

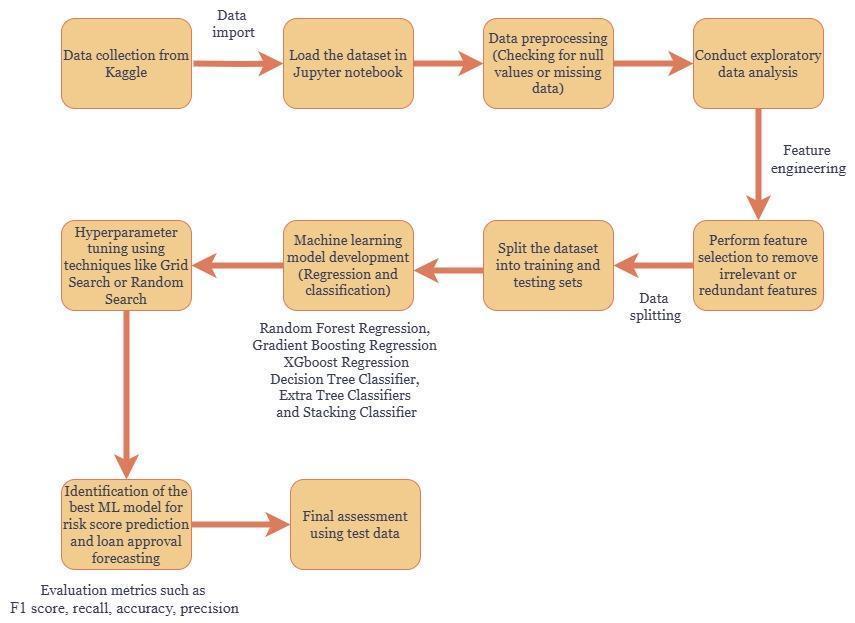
I have considered the ethical implications of using this dataset from Kaggle in evaluating data ethics. This synthetic dataset does not contain personally identifiable data, so privacy and General Data Protection Regulations do not concern me here. I also ensured I had the right to use the data. Kaggle datasets are authentic open-source data sites and provide open access to education and research. I have also checked on using synthetic data in my Research and established that it does not need university ethics because it does not contain accurate personal data.

## 4.2 Ethical Considerations in Research

Respecting rights of intellect related to the dataset and materials I have referenced; in performing this Research, I have stayed within the bounds of the Copyright, Designs and Patents Act by the Government of the UK (1988). All sources are appropriately credited to avoid plagiarism, and I have only used data for which I have obtained explicit permission. Furthermore, synthesised data has been ethical since the dataset is publicly available. I approach with transparency, respect for intellectual property, and respect for legal and ethical standards.

# Chapter 5: Methodology

## 5.1 Methodological Architecture and rationale behind the architecture



***Figure 13: Methodological architecture***

The methodology architecture for this project is depicted in ***Figure 13*** as a structured workflow, from data collection to the final model evaluation of best practices in predictive modelling. Dealing with null values during data pre-processing can enhance the data quality, which is crucial to the accuracy of the model (Palanivinayagam and Damaševičius, 2023). Exploratory data analysis (EDA) and feature engineering have also been employed to explore insight and create new variables, contributing to robust model development. According to Poian *et al.* (2023), feature selection narrows down the dataset to eliminate irrelevant variables, which improves model performance, and data splitting helps avoid biased model assessment through a training set and a testing set. Moreover, the model's accuracy is optimised through hyperparameter tuning using Grid Search or Random Search. Lastly, the models are compared using the metrics of F1-score, accuracy, recall, and precision. These standard procedures ensure the reliability of the results from the data analysis and machine learning model build-up.

## 5.2 Implementation of Software and Libraries

In this Research, I used Python programming as the primary software for data analysis and machine learning using Jupyter Notebook's interface. Jupyter Notebook is extensively valued for its ability to combine code, text, and output into one document, which is an ideal environment for interactive model development and analysis (Pimentel *et al.*, 2021). This interface allows it to use data exploration, documentation of code, and visualisation through a unified workflow, which is especially helpful in building machine learning models and documenting the performance of such models.

The significant libraries applied in this study are Numpy, Pandas, Matplotlib, Seaborn, and sci-kit-learn (sk-learn). Pandas and NumPy performed significant activities in managing the numerical and tabular data and shared the responsibilities of sufficient data preprocessing and feature engineering. For most of the visualisation requirements, Matplotlib and Seaborn have been needed to create effective graphs to identify patterns or trends in the datasets (Hafeez and Sial, 2021). The data preprocessing to encode categorical variables and splitting of data set along with model building have been initiated mostly from the scikit-learn framework. The consolidation of multiple libraries on this platform has been helpful and has contributed to the enhancement of model development and refinement, performance evaluation, and driving analytical goals in the project.

## 5.3 Machine Learning models applied

|  |  |
| --- | --- |
| **Target variable: Risk Score** | |
| **Machine learning models** | **Justification** |
| Random Forest Regression | They were selected for their ensemble approach, which reduces variance and can thus improve the prediction accuracy of continuous risk scores by using weak trees. |
| Gradient Boosting Regression | It has been selected for its iterative boosting capabilities to reduce prediction errors since the risk patterns involved are the most complex. |
| XGBoost Regression | This model is highly suitable due to its faster measurement and efficiency in handling large datasets, which enhances accuracy by prioritising misclassified data. |
| **Target variable: LoanApproved** | |
| Decision Tree Classifier | Suitable for splitting data and providing interpretability to improve accuracy when classifying loan approval outcomes. |
| Extra Tree Classifier | This model has been chosen for its random splits, which maximises diversity variability and prevents overfitting when predicting loan approval outcomes by creating forests of weaker trees and providing outcomes by maximising the accuracy of each tree. |
| Stacking Classifier | It combines multiple classifiers into one to make predictions more accurate and robust about the class label under which the loan application falls. |

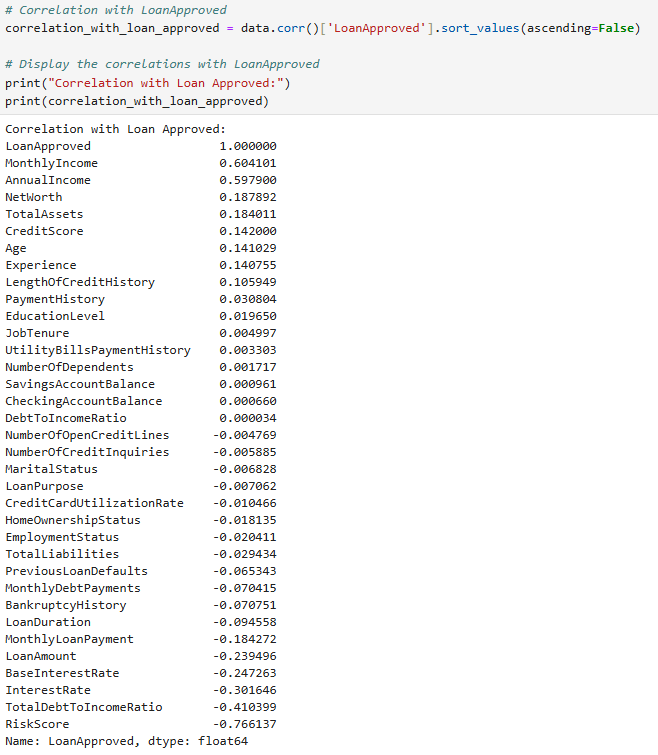
***Table 1: Choice of ML Models***

# Chapter 6: Results

## 6.1 Introduction

Demographic attributes such as age, gender, marital status, and income level as well as financial attributes such as previous loan status, credit history as well as loan repaying capacity, can substantially influence customer loan approval status. Additionally, these factors can have a significant influence on determination of the credit risk score of loan applicants. Therefore, prediction of loan approval status and credit risk score can be beneficial for financial organisations and loan-providing companies to approve loan applications based on their loan repaying capacity and credit risk score. Discussion in this chapter is based on the development and evaluation of classification (for prediction of loan approval status) and regression model (for estimating credit risk score) models. Additionally, deployment of the models for making predictions based on real-world scenarios has also been performed in this chapter.

## 6.2 Selection of features for classification

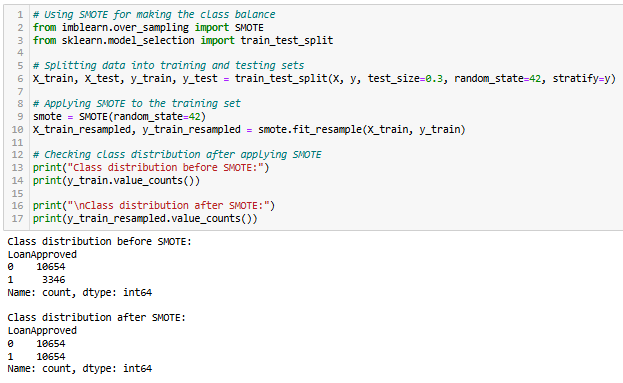


******

***Figure 14: Selection of features***

Selection of features for the classification models has been performed by calculating the correlation between features and target variable (LoanApproved). The features which are between the range of +/- 0.70 are considered in developing the classification models. From the above, it can be seen that only the risk score has a correlation of -0.77 with Loan Approve which shows multicollinearity problems for which this variable has been dropped. (***Refer to Figure 14***).

## 6.3 Class balancing using SMOTE

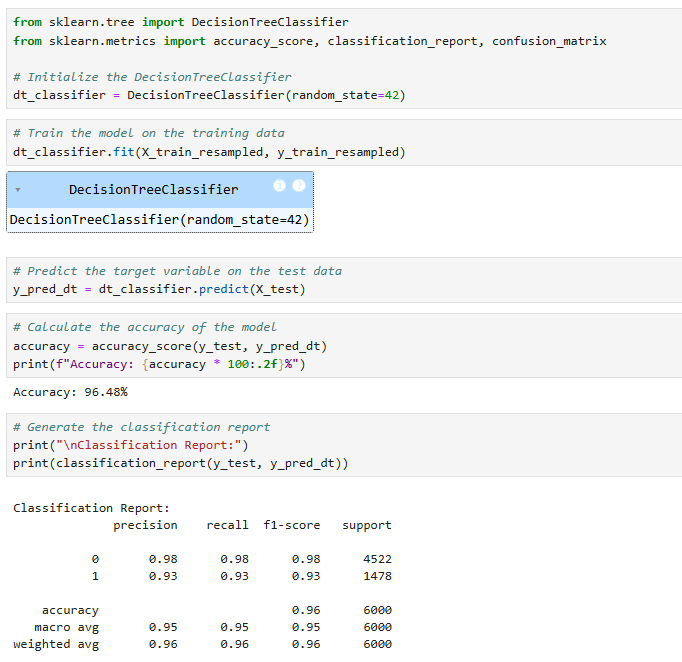
******

***Figure 15: Class balancing using the SMOTE technique***

Class balancing has been performed for the target variable (Loan Approved) as severe class imbalance has been present in the dataset (Class 0: 10654, Class 1: 3364) (***Refer to Figure 15***). The severe class imbalance can lead to inadequate training of classification models (Decision Tree, AdaBoost and Stacking), leading to model overfitting due to a lack of generalisation on unforeseen test data. As per the viewpoint of Joloudari *et al.* (2023) and Meng and Li (2022), implication of the SMOTE technique introduces a resampling strategy that creates minority class samples by generating synthetic samples in the minority class, ensuring class balancing. For this purpose, class balancing has been performed using the SMOTE technique, through which synthetic samples have been generated for minority class (Class 1). This has ensured training of the classification models on equal numbers of samples for Class 0 (10654) and Class 1 (10654), leading to better generalisation and prevention of model overfitting.

## 6.4 Classification models for Loan Approval status

### 6.4.1 Decision Tree Classifier



***Figure 16: Decision Tree Classifier***

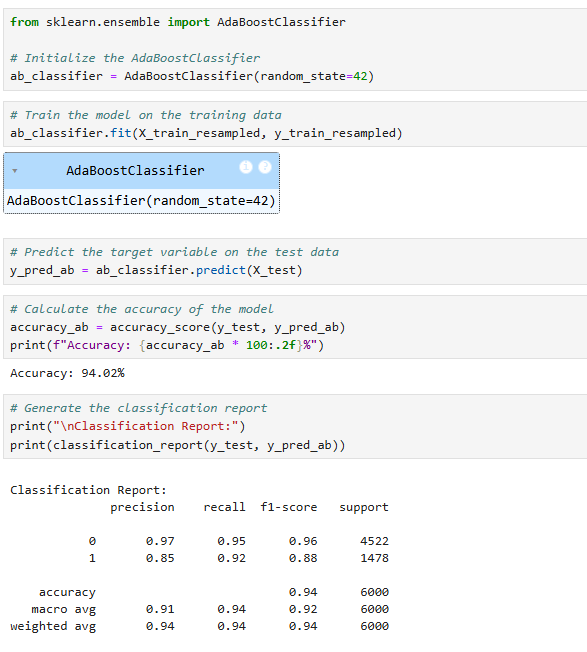
***Figure 16*** shows the model architecture and predictive performance of the decision tree classifier. The Decision Tree Classified model can be included in binary classification due to its simpler architecture and capacity for multiple features (Liu and Yang, 2022). The obtained accuracy of the decision tree classifier is 96.48%, indicating that the model has accurately classified 96.48% of instances in target classes (Loan Approval). The precision (Class 0 (0.98) and Class 1 (0.93)) and recall (Class 0 (0.98) and Class 1 (0.93)) are considerably high, indicating high predictive performance of the model in predicting Class 0 as well as Class 1 instances in loan approval.

******

***Figure 17: Confusion Matrix for Decision Tree Classifier***

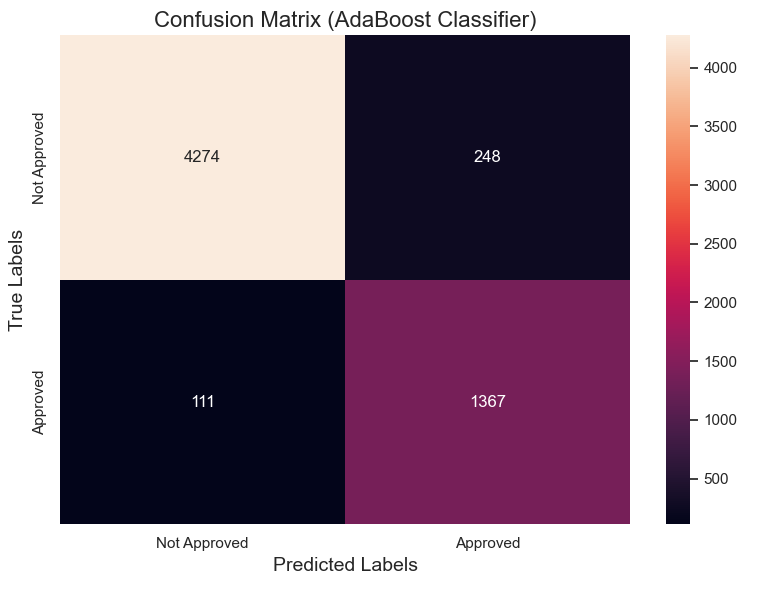
The correctly classified (True Positives = 4412 and True Negatives =1377) are considerably higher than the misclassified instances (False Positives = 110 and False Negatives = 101) (***Refer to Figure 17***). This reflects that the predictive performance of the model is exceptionally good, indicating the reliability of the model in predicting loan approval status based on financial attributes in real-world scenarios.

### 6.4.2 AdaBoost Classifier

******

***Figure 18: Model and predictive performance of AdaBoost Classifier***

The AdaBoost classifier is initialised from the ‘AdaBoostClassifier’ object from the ensemble module of Scikit-learn framework (***Refer to Figure 18***). The ensemble learning approach in AdaBoost Classifier has justified the selection as it can ensure the combination of multiple base learners to improve overall predictive performance (Ramakrishna *et al.*, 2023). The obtained accuracy of the AdaBoost classifier is 94.02% with a precision of 0.97 (Class 0) and 0.85 (Class 1) (***Refer to Figure 18***). This reflects that the model has also shown high predictive performance in classifying Class 1 instances, indicating strength of the model in predicting loan approval status in real-world scenarios in comparison to other models but less in comparison to Decision Tree.

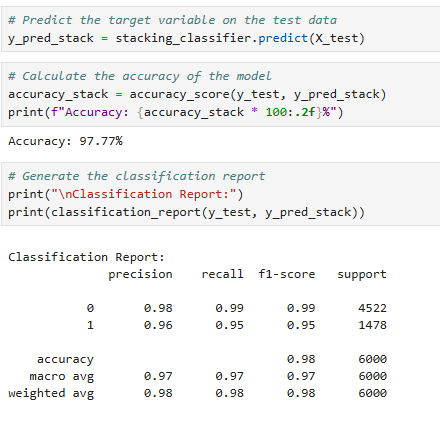
******

***Figure 19: Confusion Matrix for AdaBoost Classifier***

The high number of False Positives (248) and False Negatives (111) in the AdaBoost Classifier indicates a comparatively higher level of misclassification for the AdaBoost Classifier (***Refer to Figure 19***). This indicates a high level of performance for the AdaBoost Classifier model in predicting the loan approval status of loan applicants but less than Decision Tree.

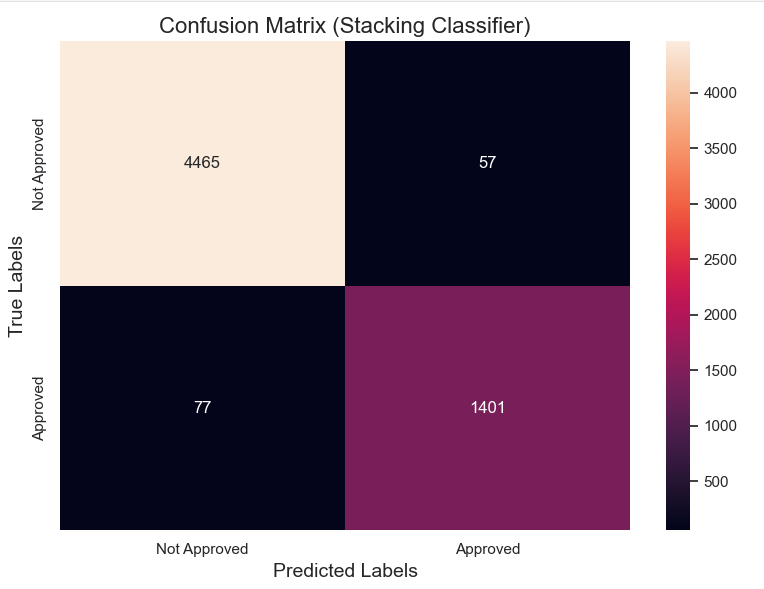
### 6.4.3 Stacking Classifier



******

***Figure 20: Model architecture and predictive performance of the Stacking classifier***

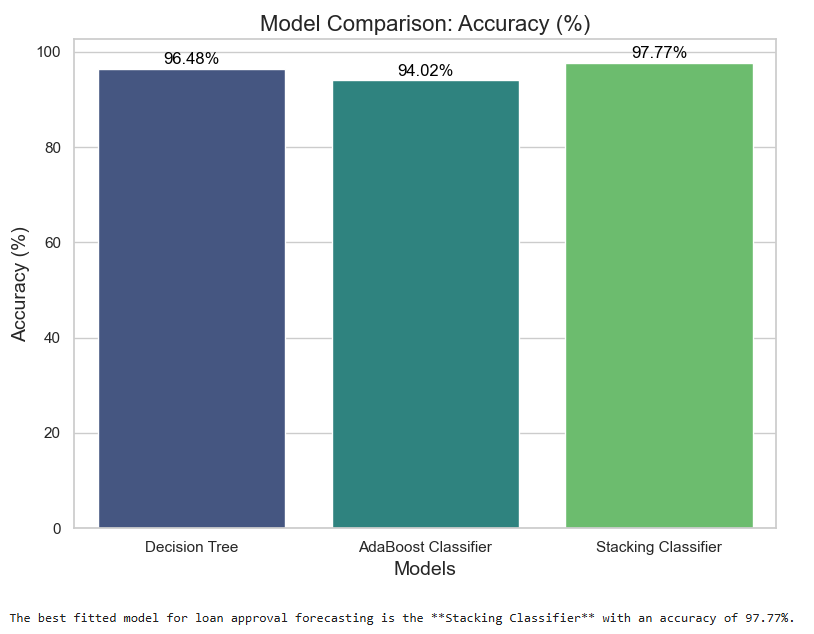
***Figure 20*** shows the model architecture as well as the predictive performance of the Stacking Classifier model. The Stacking classifier has considered decision tree, extra tree and random forest classifiers as the base estimators due to their capabilities to handle multiple features. Logistic regression has been considered as the final estimator in the stacking classifier architecture due to the binary nature of the target variable (Class 0 (Not Approved) and Class 1 (Approved)). The obtained accuracy of the Stacking classifier model is 97.77%, indicating that the model has accurately classified 97.77% of instances in the target variable (loan approval). The precision (Class 0: 0.98 and Class 1: 0.96) and recall (Class 0: 0.99 and Class 1: 0.95) are significantly higher for both Class 0 and Class 1 instances. This showcases the high predictive performance of the Stacking classifier model in predicting loan approval status based on the financial attributes of loan applicants in real-world scenarios.

******

***Figure 21: Confusion matrix for Stacking Classifier***

The number of accurately classified instances (True Positives = 4465 and True Negatives =1401) is significantly higher than misclassified instances (False Positives = 57 and False Negatives = 77) (***Refer to Figure 21***). Additionally, the True positives and negative instances are comparatively higher than other models (Decision Tree and AdaBoost. This reflects a higher level of predictive performance of the Stacking classifier model in predicting loan approval status based on unforeseen test data.

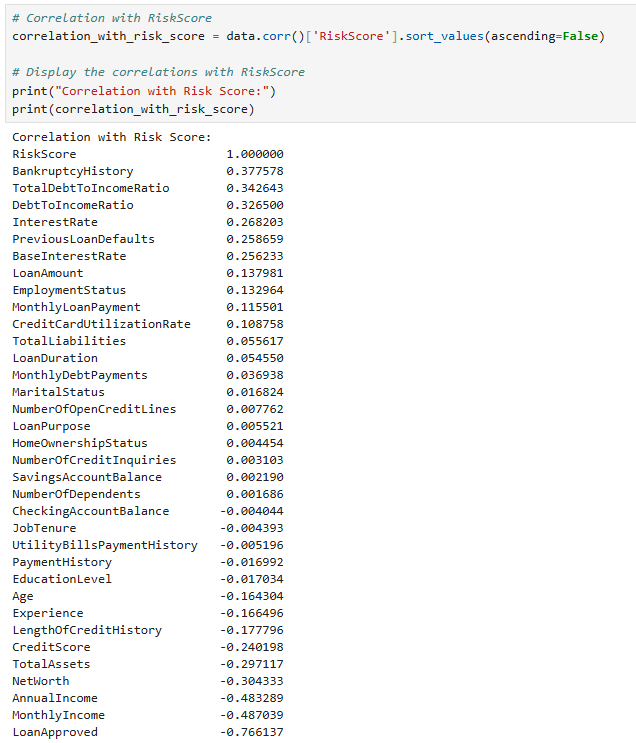
### 6.4.5 Model Comparison

******

***Figure 22: Comparison of the classifier models in terms of accuracy***

***Figure 22*** shows the comparison of predictive performance of the classifier models in terms of accuracy. The obtained accuracy of the Stacking classifier (97.77%) is considerably higher than the Decision Tree classifier (96.48%) and AdaBoost Classifier (94.02%). This shows that the integration of multiple ensemble classifiers (decision tree, extra tree and random forest) as base estimators and compilation of these base estimators using Logistic Regression as the final estimator has led to a better generalisation of unforeseen test data. This infers that the Stacking classifier is the best-fitted model for predicting loan approval status of customers. Thus, this model API can be further used in model deployment to develop a text-based user interface to allow users to predict loan approval status of applicants based on their financial attributes.

## 6.5 Regression Model for Predicting Credit Risk Score



******

***Figure 23: Features selection based on correlation***

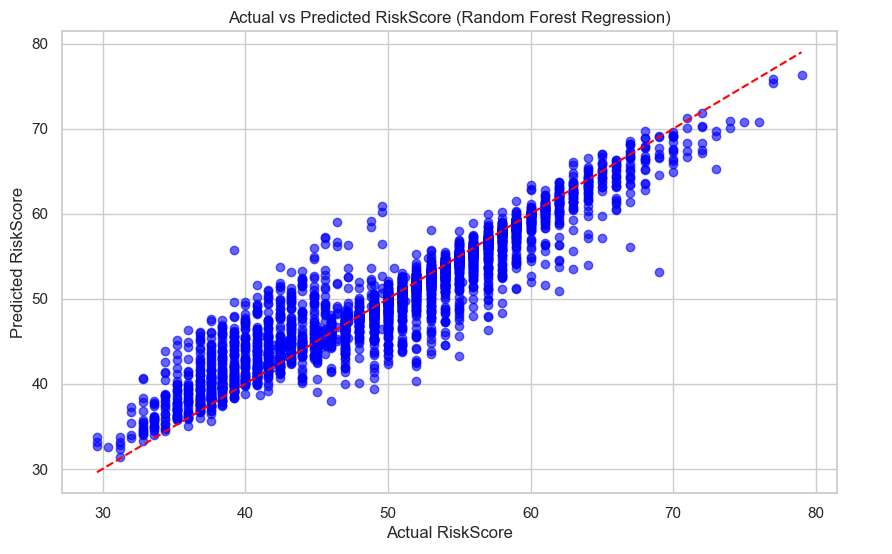
The important features of the regression model (for predicting credit risk score) have been selected using the correlation. The features with a correlation coefficient of more than between the range of +/- 0.70 are considered in training the regression models (Random Forest, Gradient Boosting and XGBoost Regressor). Through this method, all the variables except LoanApproved is dropped since the correlation value is -0.77 with Risk Score showing the multicollinearity problem. (***Refer to Figure 23***).

### 6.5.1 Random Forest Regressor

******

***Figure 24: Predictive performance of the Random Forest Regressor Model***

The Random Forest Regressor has been initialised from the ‘RandomForestRegressor’ object from the ensemble module of Scikit-learn framework. The model has been initialised with a parameter ‘n\_estimators’ =100 (default value of the parameter), which has ensured trade-off between model complexity and a sufficient level of training by considering a sufficient number of trees within the forest. The parameter of the Random Forest Regressor model ‘random\_state = 42’ has ensured reproducibility of model performance across multiple executions of the model (***Refer to Figure 24***). The obtained R-square value of the Random Forest regressor model is 0.8785, indicating the model has the capability to explain 87.85% of the variability in credit risk score, which is moderate. The Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are respectively 7.54, 2.75 and 1.71, which are high (***Refer to Figure 24***). This reflects that the predictive performance of the Random Forest regressor is high in the context of prediction of credit risk score of customers.

******

***Figure 25: Actual versus Predicted Curve for Random Forest Regressor model***

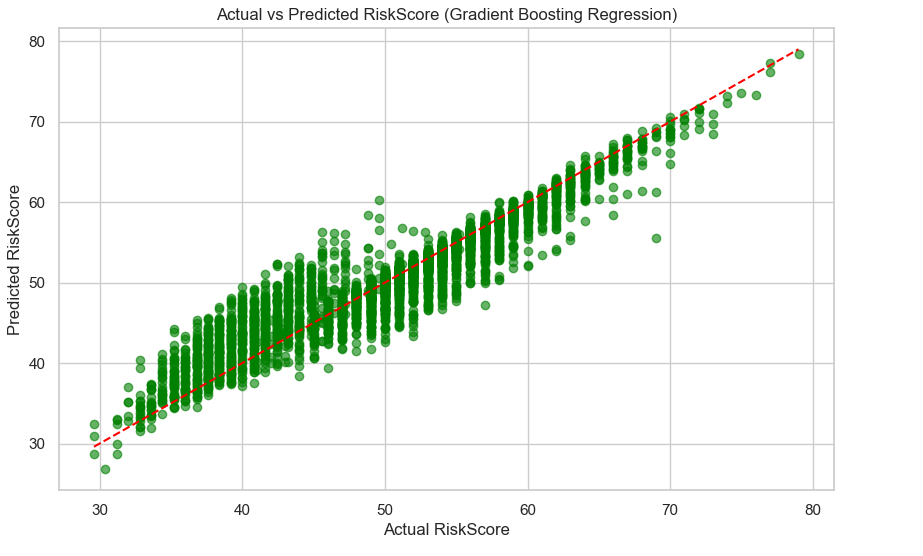
The actual versus predicted curve for Random Forest Regressor indicates that the model has passed through the actual values, indicating capability of the model in capturing the linear trends in credit risk score of customers (***Refer to Figure 25***). The capability of the model in capturing some non-linear patterns in credit risk score is good in accurately estimating the credit risk score of customers based on the financial attributes of users.

### 6.5.2 Gradient Boosting Regressor

******

***Figure 26: Model architecture and performance of the Gradient Boosting Classifier***

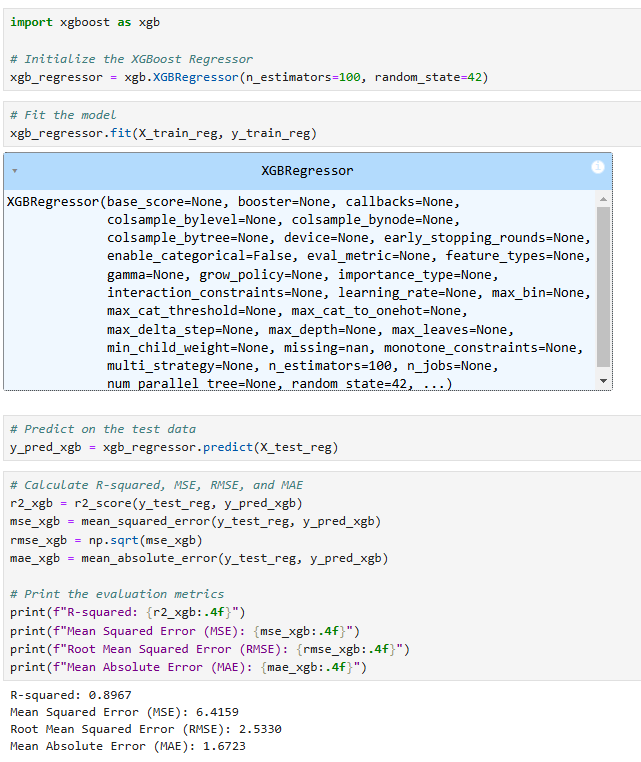
***Figure 26*** shows the model architecture and predictive performance of the Gradient Boosting Regressor model. The selection of the parameters n\_estimators = 100 (which is the default value for n\_estimators) (indicating the number of trees) has ensured low model complexity. While the parameter ‘random\_state = 42’ has ensured reproducibility of the model across multiple executions. The obtained R-square value of the Gradient Boosting Regressor model is 0.8822, which reflects that the model has explained 88.22% of variability in target variable (credit risk score) (***Refer to Figure 26***). The MSE, RMSE and MAE of the model are respectively 7.31, 2.71 and 1.85, indicating a comparatively better performance of the Gradient Boosting regressor model compared to Random Forest regressor in regard to predicting credit risk scores.

******

***Figure 27: Actual versus predicted curve for the Gradient Boosting model***

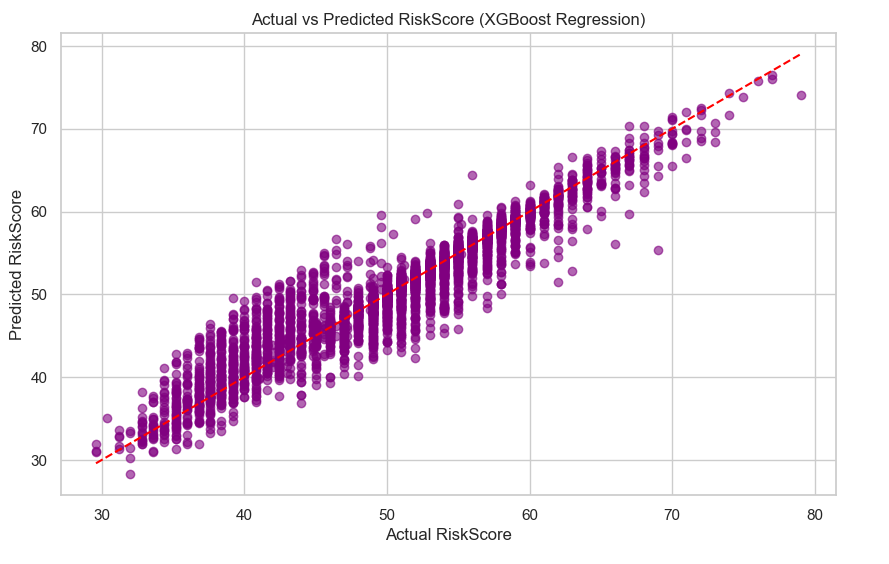
***Figure 27*** shows that the model has the capability to capture the linear relationships between features and credit risk scores. This infers that the model can be utilised in real-world scenarios for predicting the credit risk score of customers. This leads to the inference that while the model can capture linear patterns in credit risk scores of loan applicants in real-world scenarios, the model's capability in handling non-linear patterns can limit the model in real-world applications.

### 6.5.3 XGBoost Regressor



***Figure 28: Model architecture and performance of the XGBoost regressor model***

The XGBoost regressor has been initialised with the default parameters to predict the credit risk score of loan applicants (***Refer to Figure 28***). The obtained R-square value of the XGBoost regressor model is 0.8967, reflecting that the model is capable of explaining 89.47% of variability in credit risk score. The obtained MSE, RMSE and MAE of the XGBoost regressor model are 6.42, 2.53 and 1.67, indicating a very strong predictive performance of the XGBoost regressor in comparison to the other models.

******

***Figure 29: Actual versus predicted curve for XGBoost regressor***

From ***Figure 29***, it can be observed that the actual versus predicted curve has followed a linear pattern, indicating capability of XGBoost regressor model to capture linear trends. The dispersion of the predicted values from actual values of credit risk score of customers indicates the strength of the model in capturing non-linear patterns. Thus, it can be inferred that the model has provided a moderate level of predictive performance for estimating credit risk score of customers based on financial attributes (such as previous loan history, loan amount and loan repaying capacity).

### 6.5.4 Model comparison

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

***Figure 30: Comparison of regression models***

***Figure 20*** shows the comparison of predictive performance of the regression models (such as Random Forest, Gradient Boosting and XGBoost regressor) in terms of R2 score, MSE, RMSE and MAE. The R2 score of the XGBoost model (0.8967) is considerably higher than the Random Forest (0.8785) and Gradient Boosting regressor (0.8822) model. Additionally, the MSE (6.42), RMSE (2.53) and MAE (1.67) of the XGBoost model are lower than RF and Gradient regressors (***Refer to Figure 30***). This leads to the inference that the XGBoost model is the best-fitted model for predicting credit risk scores. Thus, this model can be utilised in deploying into a text-based UI to allow end users to predict credit risk score of loan applicants.

## 6.6 Model Deployment for real-world use cases

### 6.6.1 Deployment of the Classification Model for Predicting Loan Approval Status

|  |
| --- |
|  |

***Figure 31: Text-based UI for classification of loan approval status of customers***

The above figure has shown the UI for the ‘Loan Approval System’, which has been developed by deploying the stacking classifier model as a pickle object. The best-fitted model has been first serialised as a pickle object by dump method, which has been further deserialised within the prediction pipeline. From the above test cases (based on unforeseen data), it can be observed that the system has accurately classified ‘Approved’ and ‘Not Approved’ instances, justifying the applicability of the system in real-world loan approval scenarios.

### 6.6.2 Deployment of the Regression model for predicting Credit Risk Score

|  |
| --- |
|  |

***Figure 32: Test cases of the deployed Regression model in the context of prediction of credit risk score***

The best-fitted model (gradient\_boosting regressor) model API has been serialised using the Pickle library through the use of the dump method (***Refer to Figure 32***). Additionally, deserialisation of the model API has been performed within the prediction pipeline to utilise the already trained regression model for predicting the credit risk score of new applicants. For developing the UI, TKinter library has been used, through which a text-based UI system has been developed for end-users to predict the credit risk score of loan applicants based on their previous loan history and financial attributes. The above figure shows that the model has predicted the credit risk score of applicants based on new test data related to financial attributes and previous loan history of applicants. This shows the applicability of the model in real-world scenarios for predicting credit risk score of customers, which can allow financial institutions to make informed decisions within the loan approval process.

## 6.7 Chapter Summary

The implication of the Stacking classifier model has obtained the highest accuracy for predicting the loan approval status of applicants. The utilisation of the model can be highly reliable in classifying loan approval status based on the financial attributes of loan applicants. However, the Gradient Boosting Regressor model has provided the highest predictive performance for predicting credit risk score of loan applicants. The best-fitted model (Gradient Boosting regressor) model has been used for developing the text-based user interface to allow users (specifically loan approval teams of financial institutions) to estimate credit risk score of loan applicants.

# Chapter 7: Analysis and Discussion

## 7.1 Interpretation of the findings

The results obtained from exploratory data analysis infer that the majority of customers fall in the moderate-risk category, reflecting the existence of a moderate level of lending risks. Additionally, financial stability (in terms of income, savings, and assets) shows a positive influence on the minimising of credit risk scores of loan applicants, leading to approval of loan applications. Results obtained from classification models reveal that factors like monthly income, net worth, total assets, credit score, age, work experience and length of credit history contributed towards the prediction of loan approval status. The obtained accuracy of the Stacking classifier is considerably higher than the Decision Tree and AdaBoost classifier, inferring that the Stacking classifier model can be utilised for predicting loan approval status of loan applicants in real-world scenarios.

The factors that have contributed towards the prediction of credit risk score are, respectively bankruptcy history, debt-to-income ratio, interest rate, previous loan defaults, base interest rates and loan amount. This infers that the loan repaying capacity of loan applicants can fundamentally influence minimisation of credit risk score of customers. The obtained predictive performance of the XGBoost regressor model is slightly higher than the Random Forest and Gradient Boosting regressor model. This infers applicability of XGBoost regressor model in predicting credit risk score of loan applicants.

## 7.2 Discussion

Demographic factors such as age and marital status are found to have a positive influence on loan approval of customers, where customers in the middle-aged group exhibited a higher chance of getting a loan approval. On the contrary, previous studies have utilised the Random Forest classifier model to emphasise the negative influence of family dependence on the loan repaying capacity of applicants (Li and Hua, 2023 and Jin *et al.*, 2024). In terms of financial attributes, findings of this study outlined those factors like monthly income, total assets, credit score, and length of credit history have a substantial positive influence on overall approval of applicants. Additionally, financial factors such as bankruptcy score, total debt-to-income ratio, interest rate, base interest rates, monthly loan payments and credit utilisation rates have exhibited substantial positive influence on credit risk score of customers. Thus, it can be inferred that demographic attributes (such as age, marital status, family dependence, education, and experience) and financial factors (such as income, assets, credit score, loan amount, loan history) substantially influence credit risks of customers (***addressing research objective 1 and question 1***).

The purpose of developing a regression model is to predict the credit score of loan applicants based on financial attributes. In this context, three different regression models (such as Random Forest, Gradient Boosting and XGboost regressor) have been developed. The obtained R2 score of the XGBoost regressor model (0.8967) is considerably higher than Random Forest (0.8785) and Gradient Boosting regressor (0.8822). This shows XGBoost is the best-fitted model for predicting credit risk score of loan applicants. Previous studies have emphasised the high accuracy of XGBoost model (with accuracy of more than 90%), indicating relevance of XGBoost model in estimating credit risk scores (Li, 2019). This shows high applicability of Regression models predicting continuous risk scores associated with the likelihood of loan default and financial instability (***addressing research objective 2 and question 2***).

In the context of the prediction of loan approval status of loan applicants, three different classification models have been developed (such as Decision Tree, AdaBoost and Stacking Classifier). The obtained accuracy of the stacking classifier (97.77%) model is considerably higher than Decision Tree (96.48%) and AdaBoost Classifier (94.02%). This reflects high applicability and reliability of the Stacking classifier model in predicting the loan approval status of customers in real-world scenarios. On the other hand, previous studies have achieved a low accuracy of 65.63% from the XGBoost model, potentially due to inappropriate handling of class imbalance (Suhadolnik, Ueyama, and Da Silva, 2023). Thereby, it can be inferred that the Stacking classifier model has shown better predictive performance compared to previous studies, inferring the reliability of stacking classifier model in predicting loan approval of customers in real-world scenarios (***addressing research objective 3 and question 3***).

The results of this study outlined that minimisation of credit risks can be critical for financial institutions to maintain profitability and ensure long-term sustainability. Therefore, effective strategies include the implementation of advanced credit scoring models using machine learning algorithms (such as Stacking) to evaluate creditworthiness of the borrowers accurately. Previous studies have outlined that the incorporation of alternative data, such as social behaviour and payment history, enhances the prediction of credit risks effectively (Sangwan *et al.*, 2021). Therefore, consideration of social behaviour, payment history, previous loan history can be effective strategies before approving a loan as this can infer loan repaying capability of applicants (***addressing research objective 4 and question 4***).

## 7.3 Limitations

The non-consideration of advanced deep learning models such as Artificial Neural Networks (ANN) and LSTM as well as transformed-based architecture such as Auto-Encoder, can limit this study in terms of accuracy and scalability in real-world scenarios. This can possibly limit this study in estimating likelihood of loan approval and credit risks estimation. Additionally, the non-inclusion of Flask-based Python applications for loan approval and credit risk estimation systems can limit the applicability of the models in real-world scenarios due to the non-availability of web applications.

# Chapter 8: Conclusion

## 8.1 Summary

Based on the discussion, it can be summarised that financial factors such as monthly income, total assets, credit score, net worth and length of credit have a positive influence on the likelihood of loan approval. On the other hand, factors like loan amount, loan duration, monthly loan amount and total debt-to-income ratio have a negative influence on likelihood of loan approval. In the context of loan approval status prediction, the Stacking classifier provided the highest accuracy with considerably high precision and recall across classes (Class 0 and Class 1). This infers high applicability of the Stacking classifier model in predicting loan approval status in real-world scenarios. In addition to that, the XGBoost regression model has provided the highest predictive performance for estimating credit risk score of loan applicants.

## 8.2 Linking with Objectives

The discussion of this study revealed that demographic attributes like age, marital status, education level and working experience are found to have a positive influence on loan approval of customers. The customers in the middle-aged group exhibited a higher chance of getting loan approval. On the contrary, in terms of financial attributes, findings of this study outlined those factors like monthly income, total assets, credit score, and length of credit history have a substantial positive influence on likelihood of loan approval. On the contrary, financial factors like loan amount, monthly loan amount, base interest rate and total debt-to-income ratio have a substantial negative influence on likelihood of loan approval (***fulfilling research objective 1***). XGBoost regressor model has provided a comparatively higher R2 score compared to Regression models such as Random Forest and Gradient Boosting Regressor model. This shows high reliability of the XGBoost regressor model in predicting credit risk scores based on financial attributes such as loan amount, previous loan history, interest rates, monthly income, net worth and credit score (***fulfilling research objective 2***).

The Stacking Classifier model has provided a comparatively higher predictive accuracy (with high precision and recall across classes) as compared to Decision Tree and AdaBoost Classifier. This highlights the high reliability and applicability of the Stacking classifier in real-world scenarios for predicting the likelihood of loan approval of applicants based on attributes such as monthly income, net worth, total assets, credit score, loan amount, monthly loan payment, bankruptcy history and total debt-to-income ratio (***fulfilling research objective 3***). Effective strategies that financial institutions can employ in the loan approval process are consideration of previous loan history, loan repaying behaviour, family income and patent history. The consideration of these factors can allow loan approval teams of financial institutions to evaluate the credit risk of customers (***fulfilling research objective 4***).

## 8.3 Contribution of the research to real-world scenarios

The implication of the best-fitted classification model (Stacking classifier) can allow financial institutions as well as loan providing organisations to predict likelihood of loan approval based on demographic as well as financial attributes. This can allow these organisations to approve loan applications of customers where the risk of loan repayment is low, which can allow them to minimise financial losses. The implication of the best-fitted regression model (XGBoost) can allow these organisations in estimating the credit risk score of customers in real-world scenarios. This can possibly help them to identify customers that can be safe in terms of loan repayment.

## 8.4 Future scope

Future studies can focus on integration of advanced deep learning architectures such as ANN and LSTM and transformed-based models such as Auto-Encoder to capture the non-linear relationship between financial attributes and the likelihood of loan approval. This can possibly enhance the overall applicability of the loan approval system in real-world scenarios. Additionally, future studies can focus on the integration of macroeconomic factors like employment rate, inflation rate, and GDP growth rate as these factors can have indirect effects on individual income and expenditure.

# 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.

Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F. and Cho, Y.-I. (2024). Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms. *Mathematics*, [online] 12(16), pp.2553–2553. doi: https://doi.org/10.3390/math12162553.

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.

Government of the UK (1988) *Copyright, Designs and Patents Act 1988*, *Legislation.gov.uk*. www.legislation.gov.uk. Available at: https://www.legislation.gov.uk/ukpga/1988/48/contents (Accessed: 12 November 2024).

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.

Hafeez, A. and Sial, A. H. (2021). 'Comparative Analysis of Data Visualisation Libraries Matplotlib and Seaborn in Python', *International Journal of Advanced Trends in Computer Science and Engineering*, 10(1), pp. 277–281. doi https://doi.org/10.30534/ijatcse/2021/391012021.

Jin, C. *et al.* (2024) ‘Factors Influencing Household Debt Structure under Various Credit Constraints’, *Journal of Management Science and Engineering*, pp. 1–31. doi: https://doi.org/10.1016/j.jmse.2024.08.003.

Joloudari, J. H. *et al.* (2023) ‘Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks’, *Applied Sciences*, 13(6), p. 4006. doi: https://doi.org/10.3390/app13064006.

Kaggle (2024) *Financial Risk for Loan Approval*, *Kaggle.com*. Available at: https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval?resource=download (Accessed: 12 November 2024).

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.

Li, W. and Hua, X. (2023) ‘The value of family social capital in informal financial markets: Evidence from China’, *Pacific-Basin Finance Journal*, 77, p. 101922. doi: https://doi.org/10.1016/j.pacfin.2022.101922.

Li, Y. (2019) ‘Credit Risk Prediction Based on Machine Learning Methods’, *2019 14th International Conference on Computer Science & Education (ICCSE)*. doi: https://doi.org/10.1109/iccse.2019.8845444.

Liu, Y. and Yang, S. (2022) ‘Application of Decision Tree-Based Classification Algorithm on Content Marketing’, *Journal of Mathematics*, 2022, p. e6469054. doi: https://doi.org/10.1155/2022/6469054.

Mahajan, P. D. *et al.* (2020). 'Optimising predictive precision in imbalanced datasets for actionable revenue change prediction', *European Journal of Operational Research*, 285(3), pp. 1095–1113. doi: https://doi.org/10.1016/j.ejor.2020.02.036.

Meng, D. and Li, Y. (2022) ‘An imbalanced learning method by combining SMOTE with Center Offset Factor’, *Applied Soft Computing*, 120, p. 108618. doi: https://doi.org/10.1016/j.asoc.2022.108618.

Merrill, D. & 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.

Palanivinayagam, A. and Damaševičius, R. (2023). 'Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods', *Information*, 14(2), p. 92. doi: https://doi.org/10.3390/info14020092.

Pimentel, J. F. *et al.* (2021). 'Understanding and improving the quality and reproducibility of Jupyter notebooks’, *Empirical Software Engineering*, 26(4). doi: https://doi.org/10.1007/s10664-021-09961-9.

Poian, V. D. *et al.* (2023). 'Exploratory data analysis (EDA) machine learning approaches for ocean world analogue mass spectrometry', 10. doi: https://doi.org/10.3389/fspas.2023.1134141.

Ramakrishna, M. T. *et al.* (2023) ‘Homogeneous Adaboost Ensemble Machine Learning Algorithms with Reduced Entropy on Balanced Data’, *Entropy*, 25(2), p. 245. doi: https://doi.org/10.3390/e25020245.

Sangwan, S. *et al.* (2021) ‘Borrowers’ credit risk factors, perception towards repayment interventions and moral hazard in loan delinquency: an investigation of Indian microfinance institutions’, *Applied Economics*, 53(56), pp. 6554–6569. doi: https://doi.org/10.1080/00036846.2021.1946478.

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>.