|  |  |
| --- | --- |
|  | CAPSTONE PROJECT  ***Final Report*** |

**Utilizing machine learning to predict loan default.**

***Prepared by:***

Prince Lubisi

202207760

Department of Computer Science & Information Technology

Sol Plaatje University

**Submitted to:**

Dr Silas Verkijika

Lecturer: Capstone Project

Department of Computer Science & Information Technology

Sol Plaatje University

24-26 Scanlan St, New Park, Kimberley, 8301

Date

Table of Contents

[**EXECUTIVE SUMMARY** 2](#_Toc178173907)

[**BUSINESS DOMAIN** 2](#_Toc178173908)

[**1.** **INTRODUCTION** 2](#_Toc178173909)

[**2.** **PROBLEM AND OBJECTIVES** 3](#_Toc178173910)

[2.1. Problem Statement 3](#_Toc178173911)

[2.2. Business Objectives 3](#_Toc178173912)

[**3.** **SCOPE** 3](#_Toc178173913)

[3.1. In-Scope 3](#_Toc178173914)

[3.2. Out-of-Scope 4](#_Toc178173915)

[**4.** **SOLUTION** 4](#_Toc178173916)

[4.1. Overview of Methodology 4](#_Toc178173917)

[4.2. Process Flowchart and Implementation 5](#_Toc178173918)

[a) Data collection 6](#_Toc178173919)

[b) Data Preprocessing and Exploratory Data Analysis (EDA) 6](#_Toc178173920)

[c) Split Preprocessed Data 13](#_Toc178173921)

[d) Applied Machine learning Models 13](#_Toc178173922)

[e) Results 19](#_Toc178173923)

[f) Hyper-Tuning parameters for Random Forest 22](#_Toc178173924)

[g) Save Model 24](#_Toc178173925)

[h) Explainable AI 25](#_Toc178173926)

[i) Deployment 31](#_Toc178173927)

[**5.** **RECOMMENDATIONS** 31](#_Toc178173928)

[5.1. Actionable Insights 31](#_Toc178173929)

[5.2. Potential Risks and Mitigation Strategies 32](#_Toc178173930)

[5.3. Challenges 32](#_Toc178173931)

[**6.** **REFERENCES** 33](#_Toc178173932)

[**7.** **APPENDIX** 34](#_Toc178173933)

# **EXECUTIVE SUMMARY**

This project aimed to develop a robust predictive model for loan default prediction, leveraging machine learning techniques to improve financial decision-making processes. The primary goal was to identify key risk factors associated with loan default and to build a model that could accurately predict default probabilities for new applicants.

Through an extensive evaluation of various machine learning algorithms, including Random Forest, Decision Tree, Logistic Regression, K-Nearest Neighbour, Naive Bayes, Linear Discriminant Analysis, and Gradient Boosting, the Random Forest model emerged as the most effective. After hyperparameter tuning, the Random Forest model achieved a remarkable accuracy of 99%, with macro and weighted average F1-scores also at 99%, indicating excellent model performance. The model's AUC score of 0.99 further validated its predictive power.

Key deliverables from this project include the optimized Random Forest model, which was rigorously tested and validated. Additionally, explainable AI techniques such as feature importance, Partial Dependence Plots (PDPs), Permutation importance and Local Interpretable Mode-Agnostic Explanations (LIME) were employed to provide transparency into how specific features influenced the model's predictions. Additionally, the project involved the development of a comprehensive website that showcases the model's predictions and allows users to interact with the loan default prediction system in a user-friendly manner also Loan default dashboard I was developed using Power BI for customers and financial institutions to get insights on the major contributors of loan default.

The practical impact of this project is significant, as it offers a powerful tool for banks and lending institutions to assess the likelihood of loan defaults more accurately. By implementing this predictive model, lenders can mitigate risks, tailor loan products to individual borrowers' profiles, and enhance the overall efficiency of the lending process.

# **BUSINESS DOMAIN**

The project is directed at financial institutions, including banks and lending companies, who seek to enhance their loan approval processes and mitigate the risk of default.

# **INTRODUCTION**

In today’s dynamic financial environment, accurately assessing and mitigating the risk of default on loans is crucial to the sustainability and success of lending institutions (Zhang et al., 2023). This project aims to validate a model tailored for the financial sector, focusing specifically on personal unsecured loans, with a clear emphasis on efficiency in terms of minimizing the risk associated with such loans. I seek to use machine learning models to provide robust risk assessment tools, considering the inherent uncertainty and variability in lending decisions. The goal is to equip financial institutions with practical insights to optimize lending strategies, minimize risk exposure, and improve overall portfolio performance. In doing so, the project will contribute to the advancement of risk management practices, promoting stability and resilience in the face of market dynamics and changes in economic uncertainty.

# **PROBLEM AND OBJECTIVES**

## Problem Statement

The problem to be addressed in this study is the challenge of accurately predicting loans in default, a crucial problem in the financial industry. Online personal loans have grown in popularity over the years due to the continuous evolution of technology. Users of such platforms find it easier to borrow money; however, interest rates on delayed repayments and processing fees have increased, increasing the risk of non-payment (Zhu et al., 2023). With increasing economic volatility and uncertainty, financial institutions face significant challenges in accurately assessing credit risk. Loan default presents a tangible problem for both lenders and borrowers, impacting financial stability, credit availability, and overall economic health (Xianyu & Hai, 2023).

Falling to address the challenges of accurately predicting loan defaults can lead to severe negative consequences that can put a strain on financial institution, which may lead to liquidity issues, reduced profitability, and even bankruptcy in extreme case. For borrowers, defaulting on a loan can result in severe financial distress, including damage to the credit score, loss of assets, and limited access to future credit. Another issue can be economic stability; as we know, widespread loan defaults can destabilise economies, causing ripple effects across various sectors, including reduced consumer spending, decreased investment, and increased unemployment (Jean Ross, 2023).

To improve the predictive model for loan default, the study will delve deeper into secondary data and apply advanced machine learning techniques to improve predictive accuracy. It will also conduct comprehensive analyses to identify and evaluate various risk factors that contribute to loan defaults, including economic indicators and borrower characteristics. Additionally, the study will validate the performance of the model through rigorous testing and comparison with industry standards to ensure reliability.

## Business Objectives

* Analyse historical loan data to identify key predictors of loan default within the financial industry, specifically on personal unsecured loans.
* Develop and implement machine learning algorithms, including ensemble methods and feature engineering, to create a model that can predict loan default.
* Evaluation of the performance of the developed model through rigorous testing and validation against industry standards, ensuring reliability and applicability within the financial services sector.

# **SCOPE**

## In-Scope

* Considering feature engineering and selection methods to enhance model effectiveness.
* Dashboard to get insight into the dataset and key feature contributors to loan default.
* Implementation of supervised machine learning methods like Decision Trees (DT), Logistic Regression (LR), K-Nearest Neighbour (KNN), Linear Discriminant Analysis (LDA), Gradient Boosting (GB), Random Forest (RF), and Naïve Bayes (NB).
* Website for real time predictions of loan default.
* Explainability and Interpretability of the final model using Explainable AI.

## Out-of-Scope

* Deep learning techniques will not be considered due to the complexity and resource requirements.
* Integration with external third-party APIs for credit scoring will not be explored.
* Ethical and legal considerations beyond basic model fairness will not be addressed in depth.

# **SOLUTION**

## Overview of Methodology

This project employs the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which is widely recognized and comprehensive, making it particularly suitable for predictive modeling tasks like loan default prediction (Figure 1). The iterative nature of CRISP-DM allows for continuous refinement of the analysis as more insights are gained from the data.

The first phase, Business Understanding, involves clarifying the financial and regulatory objectives of the credit institution. This ensures that the model aligns with the institution's goals of minimizing losses from loan defaults while maximizing loan approvals for creditworthy customers. Additionally, regulatory considerations are integrated to ensure the model supports sustainable lending practices. In the Data Understanding phase, the dataset is explored to understand its structure, identify missing values, and gain initial insights into key variables such as credit score, income, and loan amount. Exploratory Data Analysis (EDA) techniques, including data visualization and correlation analysis, will be employed to uncover patterns and relationships within the data.

The Data Preparation phase involves cleaning and transforming the dataset to make it suitable for modelling. This includes handling missing values, encoding categorical variables, and normalizing features. The goal is to create a clean and relevant dataset that is optimized for predictive modelling. Next, in the Modelling phase, various modelling techniques, such as Navie Bayes, Gradient Boosting, Linear Discriminant Analysis, Decision Trees, Logistic Regression, and Random Forest, will be applied to build predictive models. Experiments will be conducted to determine the most effective model based on factors like accuracy, interpretability, and computational efficiency.

The Evaluation phase involves assessing the performance of the predictive models using metrics like precision, recall, accuracy, AUC-score and F1 score. Back testing with historical data will be performed to evaluate the model's predictive accuracy and robustness. Finally, in the Deployment phase, the validated predictive model will be integrated into the institution's operations through a web-based interface, allowing for real-time predictions on new customer data. This integration directly influences loan approval decisions and risk management strategies. The entire process will be documented in a comprehensive report, and continuous monitoring of the deployed model will ensure its ongoing effectiveness.

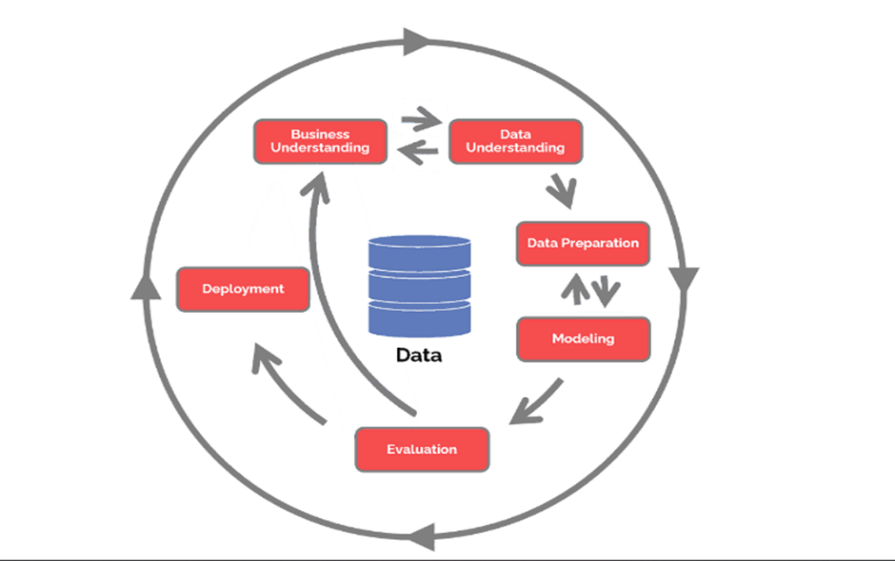


Figure 1 illustrates the six phases of CRISP-DM (Data Science Process Alliance, 2021)

## Process Flowchart and Implementation

The process flowchart below (Figure 2) illustrates the high-level steps involved in the development of the Loan Default Predictor. This flowchart is designed to provide a clear overview of the entire process, from data preprocessing to model deployment

A diagram of a software development process

Description automatically generated

Figure 2 Diagrammatic flow chart of the proposed methodology.

### Data collection

The dataset is secondary data and can be found on the Kaggle platform. The name of the dataset is called “[Loan Default Prediction Dataset](https://www.kaggle.com/datasets/nikhil1e9/loan-default)”. It contains 255347 rows and 18 columns in total, which include both numerical and categorical features (Figure 3).



Figure 3 Shows different features in the dataset which was used.

### Data Preprocessing and Exploratory Data Analysis (EDA)

#### Data Preprocessing

Data preprocessing includes, dropping non required column, handling missing values, encoding categorical features, Balancing the Target (Default) column and checking for outliers and removing them. In the dataset the column called LoanID was dropped due to that it is insignificant for the loan default prediction. When checking for missing values in the dataset, the dataset show that it does not have missing values (Figure 4).

A screenshot of a computer

Description automatically generated

Figure 4 Code snippet for checking and Handling Missing Values.

In this project, the categorical features such as Education, Employment Type, Marital Status, Loan Purpose, Has Mortgage, has Dependents, and Has Cosigner were label encoded using Label Encoder (Figure 5).

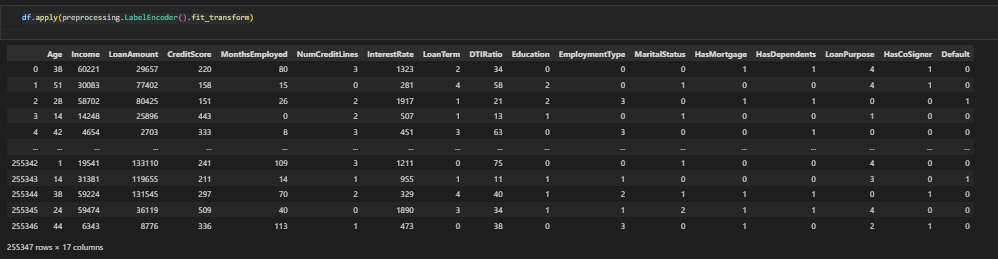


Figure 5 Code snippet for Label encoding categorical features.

To check if the dataset is balanced especially the target (Default) column, a bar plot wat plotted. The code snippet (Figure 6) performs oversampling on a dataset to balance the classes in the Default column. It starts by counting the occurrences of each class in the Default column and identifies the majority class, which is the class with the highest count. The number of samples in this majority class is then determined. Next, the code oversamples the majority class by grouping the dataset by the Default column and applying a lambda function that samples the majority class with replacement until it matches the number of samples in the majority class. The resulting oversampled dataset is then reset to remove the group by index.

A screen shot of a computer

Description automatically generated

Figure 6 Code snippet for Balancing the Data.

Finally, the shapes of the original and oversampled datasets are printed to show the change in the dataset size after oversampling. Table 1 shows Bar plot of Default before and After Balancing.

Table 1 Show Bar plot of Default before and After Balancing.

|  |  |
| --- | --- |
| Original Dataset | Original Data after Balancing |
| A graph with blue rectangular bars  Description automatically generated | A graph with blue rectangular bars  Description automatically generated |

To check and handle outliers in the code we use this code snippet (Figure 7). The code is designed to check and adjust the skewness of specific columns in a Data Frame by removing outliers. The code prints the skewness of each column before removing outliers. For each column, it calculates the first quartile (Q1), third quartile (Q3), and the interquartile range (IQR). Using these values, it determines the upper and lower limits for outliers. Any values above the upper limit are replaced with the upper limit, and any values below the lower limit are replaced with the lower limit.



Figure 7 Code snippet that check and remove outliers.

#### Exploratory Data Analysis (EDA)

During the Exploratory Data Analysis (EDA) phase, I performed several key tasks to understand and prepare the dataset. I started by calculating various summary statistics to gain insights into the central tendency, dispersion, and overall distribution of the data (Figure 8). From the observations in Figure 8, the youngest person in the dataset is 18 years old, while the oldest is 69 years, with a mean age of 43 years. The minimum income is 15,000, and the maximum is 149,999, with a mean income of 82,499.30. Credit scores range from 300 to 849, with a mean score of 574.26. There are individuals with 0 months of employment history, while the highest employment history is approximately 10 years, with a mean of 5 years. Loan amounts range from 5,000 to 249,999, with a mean of 127,578.87. Interest rates vary from 2% to 25%, with a mean rate of 13%. Loan terms range from 12 to 60 months, with a mean term of approximately 36 months

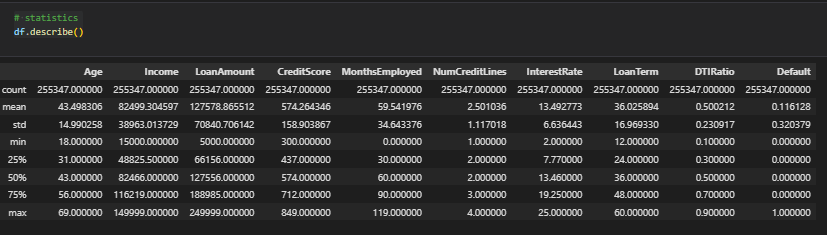


Figure 8 Code snippet for statistical Distribution of the Data

Table 2 depict the statistical distribution using Histogram and Box whisker plot along with some observations or insight found on the distributions.

Table 2 Show statistical distribution using Histogram and Box whisker plot

|  |  |  |
| --- | --- | --- |
| **Age** | | **Observations** |
|  | A graph with lines and numbers  Description automatically generated | The majority of the data is concentrated on the left side, particularly around the ages of 20-30, with a gradual decrease in counts as the age increases. This indicates that there are fewer occurrences of higher age values compared to lower ones. |
| **Income** | | **Observations** |
|  | A graph with a red rectangle and blue lines  Description automatically generated | The majority of the data is concentrated on the left side, with higher counts for lower income values. As the income increases, the counts gradually decrease, indicating fewer individuals with higher incomes. |
| **Loan Amount** | | **Observations** |
|  | A graph with a red rectangle and blue lines  Description automatically generated | The count of loans generally increases with loan amounts. Beyond 200,000, the frequency begins to decrease slightly.  The visual implies that higher loan amounts are more frequent, with a peak around the middle-high range before tapering off towards the upper limit. |
| **Credit Score** | | **Observations** |
|  | A graph with red rectangle and blue lines  Description automatically generated | The digrams shows a nearly uniform distribution of credit scores between 300 and 800, with slight variations.  There are some small fluctuations in frequency, but most credit score ranges (from 400 to 800) show relatively similar counts, around 12,000 to 13,000. |
| **Months Employed** | | **Observations** |
| A graph with red lines and numbers  Description automatically generated |  | The highest frequency occurs for those with very short employment durations (around 0-10 months), with the count peaking close to 18,000.  As the number of months employed increases, the frequency generally decreases, with some fluctuations. |
| **Interest Rate** | | **Observations** |
| A graph with red lines  Description automatically generated | A graph with lines and numbers  Description automatically generated | The Negative skew in interest rates means that most loans have higher rates, which could increase the risk of default if borrowers find it difficult to make their payments. |
| **DTI Ratio** | | **Observations** |
| A graph with red lines and numbers  Description automatically generated | A graph with lines and a rectangle  Description automatically generated | A symmetric DTI ratio distribution indicates a balanced financial health among the individuals in the dataset, with no extreme outliers significantly affecting the overall average |

To uncover relationships between features and the target variable, I checked the correlation of features with the target (Figure 9). The correlation analysis reveals several insights about the likelihood of loan default. Age has a small negative correlation (-0.17), indicating that as age increases, the likelihood of loan default decreases slightly. Similarly, income shows a small negative correlation (-0.099), suggesting that higher income individuals are slightly less likely to default on their loans. Conversely, the loan amount has a small positive correlation (0.087), implying that larger loans are slightly more likely to be defaulted on.

Credit score exhibits a very small negative correlation (-0.034), indicating that individuals with higher credit scores are slightly less likely to default, though the effect is minimal. Months employed also shows a small negative correlation (-0.097), suggesting that individuals with longer employment histories are slightly less likely to default. The number of credit lines has a very small positive correlation (0.028), indicating that individuals with more credit lines are slightly more likely to default, but the effect is minimal. Interest rate has a small positive correlation (0.13), suggesting that loans with higher interest rates are slightly more likely to be defaulted on. Loan term shows an extremely small positive correlation (0.00054), indicating that longer loan terms have a very minimal effect on the likelihood of default. Lastly, the debt-to-income (DTI) ratio has a very small positive correlation (0.019), suggesting that a higher DTI ratio has a very minimal effect on the likelihood of default.

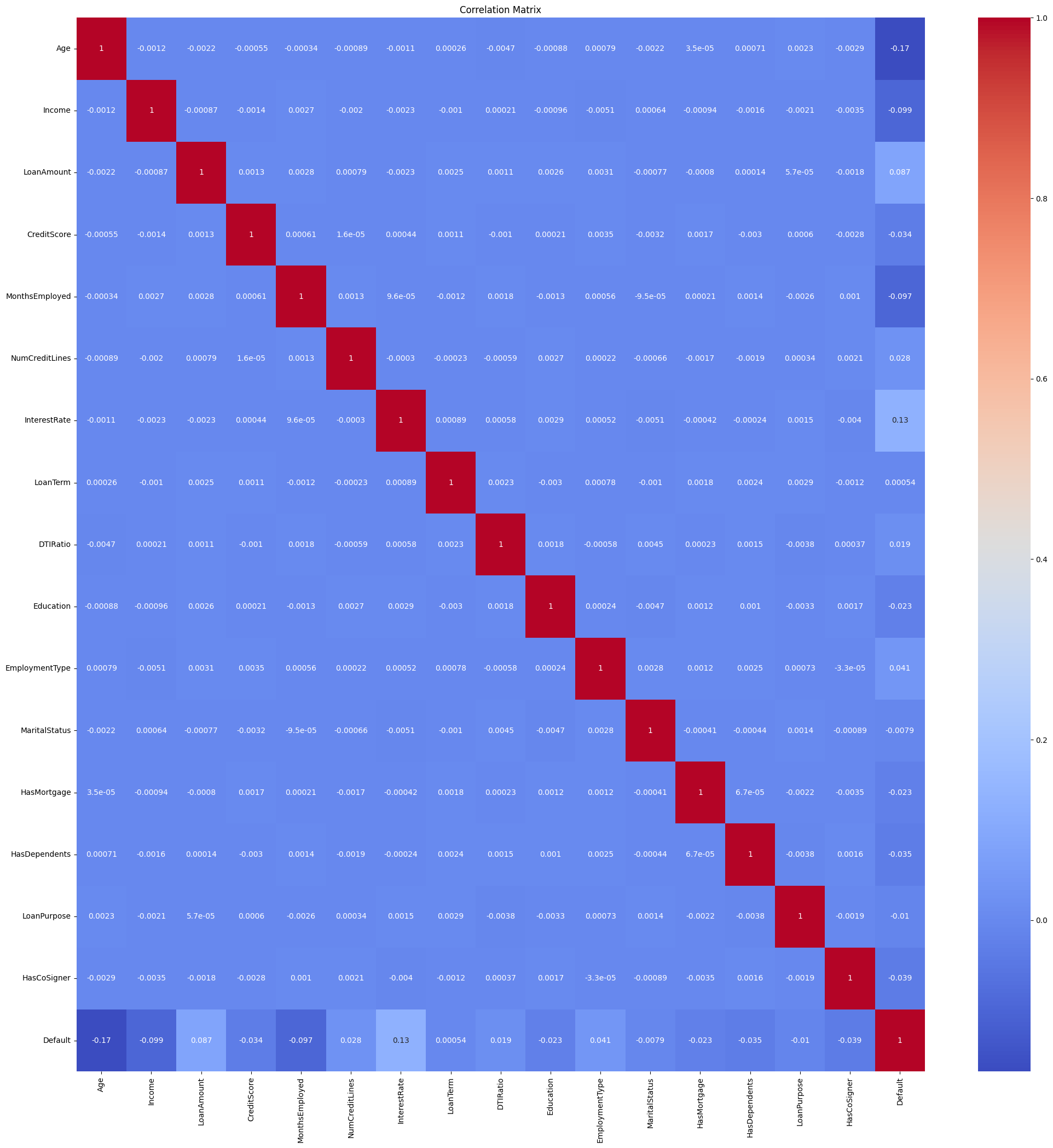


Figure 9 Shows the correlation analysis

In Figure 10, Analysis of loan default data shows that divorced individuals have the highest default rate, followed by single individuals and married individuals have the lowest default rate. Unemployed borrowers face the greatest risk of default, followed by part-time and self-employed, while full-time employees are the least vulnerable. Age appears to be inversely related to the risk of default, as older borrowers are less likely to default. Interestingly, higher interest rates are associated with lower default rates. For loan purposes, commercial loans represent the highest failures, followed by auto, education and other categories, while home loans represent the lowest. Borrowers without Dependent are more likely to default than those with Dependent. Longer employment history is associated with a reduction in the likelihood of default, while the increase in loan terms reduces default rates to 36 months, after which the risk begins to rise again.

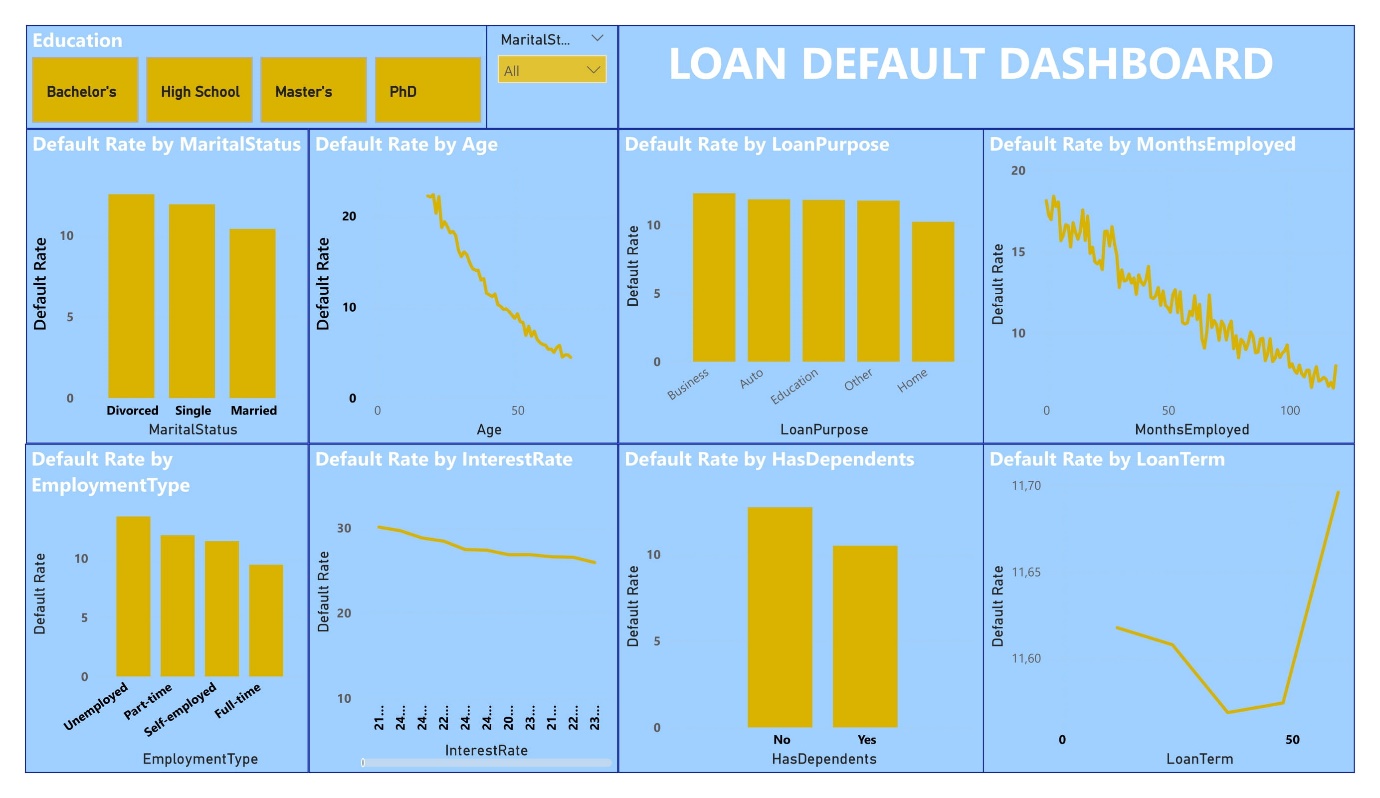


Figure 10 Snippet Dashboard of Loan Default Analysis.

### Split Preprocessed Data

To split the dataset into training and testing sets with 70% of the data for training and 30% for testing (Figure 10).

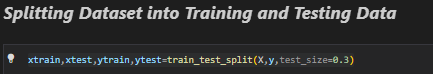


Figure 11 Splitting Dataset into Training and Testing

### Applied Machine learning Models

#### Random Forest

In this step, the Random Forest model, an ensemble learning method, was employed to predict loan defaults. The Random Forest algorithm is particularly effective because it combines the outputs of multiple decision trees to improve predictive performance and reduce overfitting. This model was selected due to its ability to handle large datasets with high dimensionality, which aligns with the complexity of our loan default dataset. The implementation process began by initializing the Random Forest classifier and fitting it to the training data (Figure 12). The model was then evaluated using key performance metrics, including accuracy and F1 scores, to assess its predictive power on both the training and test sets.



Figure 12 Code snippet for Random Forest and Decision Tree

The model was then evaluated using key performance metrics, including accuracy, precision, recall and F1 scores, to assess its predictive power on both the training and test sets (Figure 13).

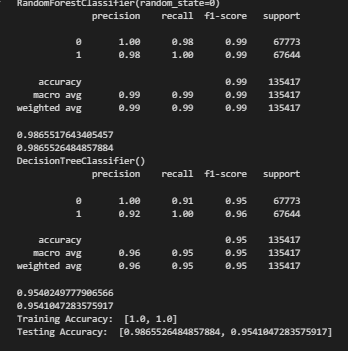


Figure 13 Performance metrics for Random Forest and Decision Tree.

#### Decision Tree

The Decision Tree model was applied in this step to predict loan defaults. Decision Trees are simple yet powerful models that work by recursively splitting the data based on the feature that provides the most significant information gain at each step. This approach is particularly useful for interpretability, as it provides a clear representation of the decision-making process. To implement the Decision Tree model, the classifier was initialized and trained using the dataset (Figure 12). The model's performance was then evaluated using accuracy, F1 scores, recall and precision on both the training and test sets (Figure 13). Decision Trees offer a straightforward and interpretable method for understanding the underlying patterns in the data, making them a valuable tool in predictive modeling.

#### Logistic Regression

In this step, the Logistic Regression model was employed to predict loan defaults. Logistic Regression is a widely used statistical model that estimates the probability of a binary outcome based on one or more predictor variables. It is particularly well-suited for classification tasks where the goal is to predict a dichotomous variable, such as whether a borrower will default on a loan. The Logistic Regression model was trained using the scaled training dataset. Once trained, the model was used to predict the outcomes on the test set, and the probabilities of class 1 (loan default) were computed (Figure 14).

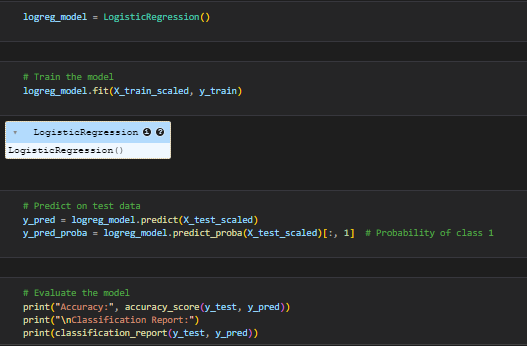


Figure 14 Implementation of Logistic Regression Model

The model's performance was then assessed using accuracy and a detailed classification report, which includes accuracy, precision, recall, and F1 scores (Figure 15).

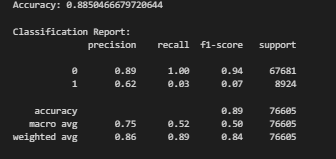


Figure 15 Performance matrix for Logistic Regression

#### K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm was utilized to predict loan defaults. KNN is a simple, non-parametric method used for classification tasks. It works by identifying the 'k' closest data points (neighbors) to a given test point and classifying the test point based on the majority class among these neighbors. The simplicity and effectiveness of KNN make it a popular choice, particularly when the dataset does not assume any specific form for the underlying data distribution. For this project, a KNN model was trained using the scaled training data, with the number of neighbors set to 5. The trained model was then applied to the test dataset to predict outcomes, and the probabilities for class 1 (loan default) were calculated (Figure 16).

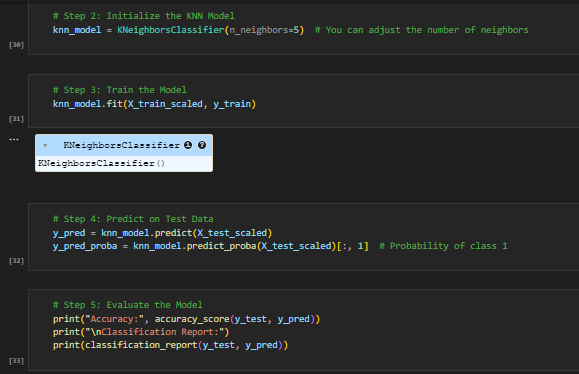


Figure 16 Implementation of KNN model

The model's accuracy and detailed classification metrics, such as accuracy, precision, recall, and F1 scores, were evaluated to assess its performance (Figure 17).

A screenshot of a computer

Description automatically generated

Figure 17 Performance Matrix of KNN model.

#### Naive Bayes

The Naive Bayes algorithm was applied to predict loan defaults. Naive Bayes is a probabilistic classifier based on Bayes' theorem, which assumes that the features are independent given the class label. Despite this strong independence assumption, Naive Bayes often performs surprisingly well, especially in high-dimensional datasets or when the features are mostly categorical. In this project, the Gaussian Naive Bayes model, which is appropriate for continuous data, was used. The model was trained on the scaled training dataset and then tested on the test dataset. The predicted outcomes and the probabilities for class 1 (loan default) were generated (Figure 18).

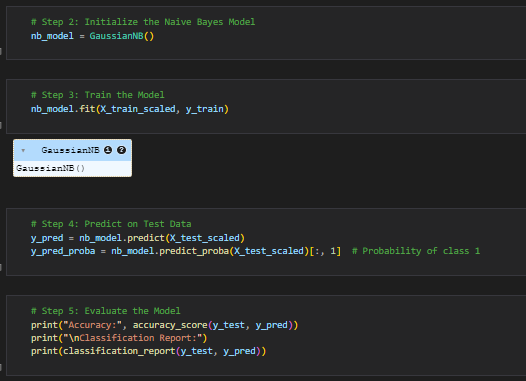


Figure 18 Implementation of Naive Bayes model.

The model's performance was evaluated using accuracy and a detailed classification report, which included accuracy, precision, recall, and F1 scores (Figure 19).

A screenshot of a computer

Description automatically generated

Figure 19 Performance Matrix for Naive Bayes model.

#### Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) was implemented as part of the machine learning models for predicting loan defaults. LDA is a method used in statistics and machine learning for dimensionality reduction while preserving as much of the class discriminatory information as possible. It works by modelling the difference between classes and is particularly effective when the data is linearly separable. In this project, the LDA model was initialized and trained using the scaled training dataset. Once trained, the model was tested on the scaled test dataset to predict loan defaults. Both the predicted class labels and the probabilities for class 1 (loan default) were generated (Figure 20).

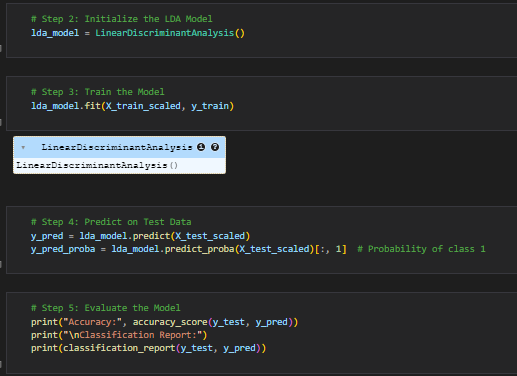


Figure 20 Implementation of the Linear Discriminant Analysis (LDA).

The model's performance was assessed through accuracy and a comprehensive classification report, which provided insights into accuracy, precision, recall, and F1 scores (Figure 21).

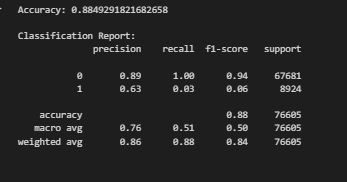


Figure 21 Performance Matrix of Linear Discriminant Analysis (LDA).

#### Gradient Boosting

Gradient Boosting was employed in this project to enhance the predictive accuracy for loan defaults. Gradient Boosting is a powerful ensemble technique that builds multiple decision trees sequentially, where each tree tries to correct the errors of the previous ones. The model is known for its robustness and effectiveness in handling various types of data, making it a preferred choice for complex datasets like the one used in this project. In this phase, the Gradient Boosting model was initialized and trained using the scaled training dataset. The model was then used to make predictions on the scaled test dataset, with both the predicted class labels and probabilities for class 1 (loan default) being recorded (Figure 22).

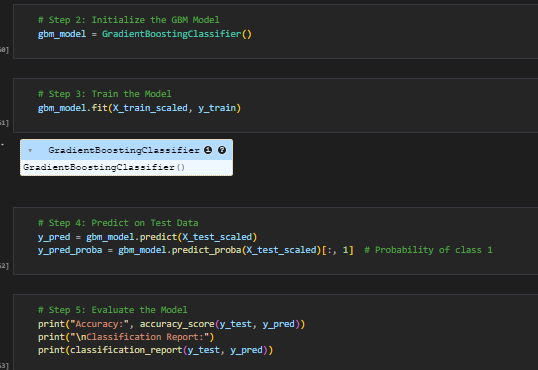


Figure 22 Implementation of the Gradient Boost model.

The performance evaluation was conducted by calculating the model's accuracy and generating a detailed classification report, which included metrics such as accuracy, precision, recall, and F1 scores (Figure 23).

A screenshot of a computer screen

Description automatically generated

Figure 23 Performance matrix of Gradient Boost model.

### Results

Table 3 below show confusion matrix and the ROC curve for each machine learning algorithm that was implemented. For each machine learning algorithm observations were made.

Table 3 Shows the performance results for all the Implemented models

|  |  |  |
| --- | --- | --- |
| **LOGISTIC REGRESSION (LR)** | | **Observations** |
|  | A line graph with orange and blue lines  Description automatically generated | * 67492 correctly classifed non defalter for loan and 189 where missclassifed. * 307 correctly classified as defaulter for loan with 8612 misclassied. * AUC=0.75 indicates the model has a fair ability to distinguish between defaulter and non defaulter |
| **K-NEAREST NEIGHBOUR (KNN)** | | **Observations** |
| A chart of a number of different colors  Description automatically generated with medium confidence | A graph with blue and orange lines  Description automatically generated | * 66556 correctly classifed non defalter for loan and 1125 where missclassifed. * 451 correctly classified as defaulter for loan with 8473 misclassied. * AUC=0.60 indicates the model has a modest ability to distinguish between defaulter and non defaulter |
| **NAIVES BAYES (NB)** | | **Observations** |
|  |  | * 67541 correctly classifed non defalter for loan and 140 where missclassifed. * 222 correctly classified as defaulter for loan with 8702 misclassied. * AUC=0.75 indicates the model has a fair ability to distinguish between defaulter and non defaulter |
| **LINEAR DISCRIMINANT ANALYSIS (LDA)** | | **Observations** |
|  | A line graph with orange and blue lines  Description automatically generated | * 67529 correctly classifed non defalter for loan and 152 where missclassifed. * 261 correctly classified as defaulter for loan with 8663 misclassied. * AUC=0.75 indicates the model has a fair ability to distinguish between defaulter and non defaulter |
| **GRADIENT BOOSTING (GB)** | | **Observations** |
|  | A graph of a curve  Description automatically generated | * 67443 correctly classifed non defalter for loan and 239 where missclassifed. * 460 correctly classified as defaulter for loan with 8464 misclassied. * AUC=0.75 indicates the model has a fair ability to distinguish between defaulter and non defaulter |
| **RANDOM FOREST (RF)** | | **Observations** |
| A screenshot of a graph  Description automatically generated | A graph with a line and a point  Description automatically generated with medium confidence | * 66249 correctly classifed non defalter for loan and 1524 where missclassifed. * 67347 correctly classified as defaulter for loan with 297 misclassied. * AUC=0.99 indicates the model has a very high ability to distinguish between defaulter and non defaulter |
| **DECISION TREE (DT)** | | **Observations** |
|  | A graph of a function  Description automatically generated | * 61816 correctly classifed non defalter for loan and 5957 where missclassifed. * 67419 correctly classified as defaulter for loan with 225 misclassied. * AUC=0.95 indicates the model has a very high ability to distinguish between defaulter and non defaulter |

Base on Table 3 Random Forest model is the best performing model with 66249 correctly classified non defaulter for loan and 1524 where misclassified also with 67419 correctly classified as defaulter for loan with 225 misclassed. The AUC of 99% indicates the model has a very high ability to distinguish between defaulter and non-defaulter. The second-best model will be the Decision Tree model with AUC of 95% indicates the model has a very high ability to distinguish between defaulter and non-defaulter. To further support the result above, Table 4 provides a comprehensive comparison of various machine learning algorithms based on several performance metrics: Precision, Recall, F1-Score, Accuracy, and AUC score.

Logistic Regression (LR) shows good accuracy and AUC score, indicating a decent ability to distinguish between classes. However, its macro recall is relatively low, suggesting it may struggle with imbalanced classes. K-Nearest Neighbour (KNN) has lower precision and recall compared to other models, and its AUC score is the lowest, indicating it may not be the best choice for this dataset. Naive Bayes (NB) performs similarly to Logistic Regression, with good accuracy and AUC score, but its macro F1-Score is slightly lower. Linear Discriminant Analysis (LDA) shows similar performance to Logistic Regression and Naive Bayes, with good accuracy and AUC score, but slightly lower macro recall. Gradient Boosting (GBM) performs well across all metrics, with slightly better precision and F1-Score compared to Logistic Regression, Naive Bayes, and LDA.

Table 4 Performance Analysis and comparison of different Machine learning Algorithms

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Precision | | Recall | | F1- Score | | Accuracy | | AUC score |
| Macro average | Weighted average | Macro average | Weighted average | Macro average | Weighted average | Training | Testing |
| 1. LR | 0.75 | 0.86 | 0.52 | 0.89 | 0.50 | 0.84 | 0.89 | 0.89 | 0.75 |
| 1. KNN | 0.59 | 0.82 | 0.52 | 0.87 | 0.51 | 0.83 | 0.89 | 0.87 | 0.60 |
| 1. NB | 0.75 | 0.85 | 0.51 | 0.88 | 0.49 | 0.83 | 0.89 | 0.88 | 0.75 |
| 1. LDA | 0.76 | 0.86 | 0.51 | 0.88 | 0.50 | 0.84 | 0.89 | 0.88 | 0.75 |
| 1. GBM | 0.77 | 0.86 | 0.52 | 0.89 | 0.52 | 0.84 | 0.89 | 0.89 | 0.75 |
| 1. RF | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 1.00 | 0.99 | 0.99 |
| 1. DT | 0.96 | 0.96 | 0.95 | 0.95 | 0.95 | 0.95 | 1.00 | 0.95 | 0.95 |

Random Forest (RF) shows outstanding performance across all metrics, with near-perfect precision, recall, F1-Score, accuracy, and AUC score. This indicates it is highly effective for this dataset. Decision Tree (DT) also performs very well, with high precision, recall, F1-Score, accuracy, and AUC score. However, it is slightly less effective than Random Forest. Based on the provided metrics, Random Forest (RF) is the best model for this dataset. It achieves near-perfect scores across all performance metrics, indicating it is highly reliable and effective for classification tasks in this context. Random Forest being the be best model in terms of performance is supported by studies like (Egwa, 2022; Li & Wu, 2024; Wang et al., 2020) , and these studies used different datasets.

### Hyper-Tuning parameters for Random Forest

In the Random Forest model implementation, after determining it as the best-performing model, a detailed hyperparameter tuning process was conducted using RandomizedSearchCV. The tuning involved experimenting with various combinations of key hyperparameters to optimize the model's performance. The hyperparameters included the number of estimators, which defines the number of trees in the forest, the maximum depth of the trees, and the maximum features considered for splitting a node. The search also included different criteria for measuring the quality of a split (criterion) and whether to bootstrap samples when building trees (bootstrap) (Figure 24).

The RandomizedSearchCV method was used with 20 iterations, which allowed for an efficient exploration of the hyperparameter space without the computational expense of an exhaustive search. The cross-validation (cv=3) ensured that the model's performance was robust across different subsets of the training data. The optimal set of hyperparameters identified were n\_estimators = 200, max\_features = log2, max\_depth = 44, criterion = gini, and bootstrap = False. These settings yielded the best validation score of 0.9755, indicating that the model was well-tuned to generalize effectively to unseen data while maintaining high accuracy (Figure 24).

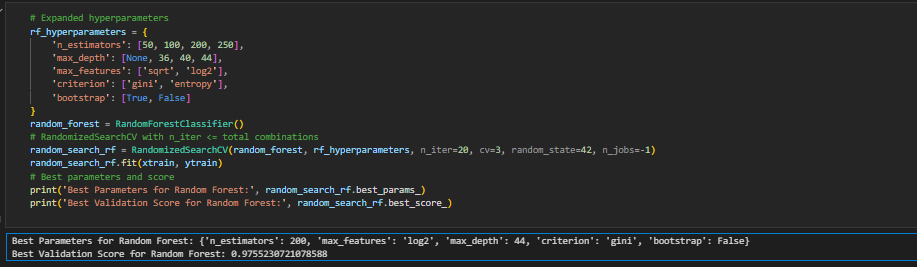


Figure 24 Implementation of Hyper Tuning on the Random Forest model.

The hyperparameter-tuned Random Forest model produced an excellent classification performance, as demonstrated by the classification report. The model achieved a precision of 1.00 for class 0 and 0.99 for class 1, indicating that the model is highly accurate in predicting both classes. The recall values were also remarkably high, with 0.99 for class 0 and 1.00 for class 1, suggesting that the model is equally effective at identifying true positives across both classes. The f1-scores, which balance precision and recall, were 0.99 for both classes, underscoring the model's overall balanced performance. The accuracy of the model across the entire dataset was 0.99, reflecting its robustness in making correct predictions for most instances. Additionally, both the macro average and weighted average for precision, recall, and f1-score were all 0.99, confirming that the model maintained a consistent and high level of performance regardless of the class distribution (Figure 25).

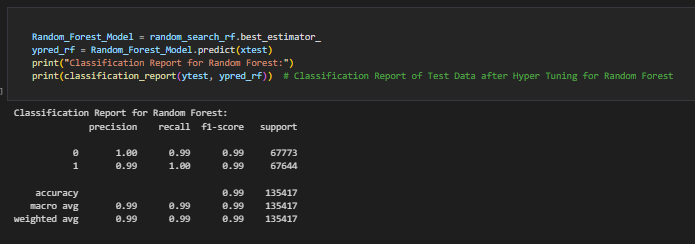


Figure 25 Classification Report of Hyper tuning for Random Forest.

The hyperparameter-tuned Random Forest model exhibited exceptional performance, as highlighted by the classification report and the AUC score. The model achieved a perfect AUC score of 1.00, indicating that it can distinguish between the positive and negative classes with perfect accuracy. This reflects the model's ability to assign higher probabilities to positive instances than to negative ones consistently (Table 5).

Table 5 Confusion Matrix and ROC curve of Random Forest after hyper tuning.

|  |  |
| --- | --- |
| **Random Forest After Hyper Tuning** | |
|  | A graph with a line  Description automatically generated |

### Save Model

After determining the optimal hyperparameters, the best-performing Random Forest model was saved for future use. By using the joblib library, the model was serialized and saved as a file named final. Joblib (Figure 26). This step ensures that the trained model can be efficiently loaded and used for predictions without the need to retrain it each time.



Figure 26 Saved Random Forest Model

### Explainable AI

In this step, Explainable AI (XAI) techniques are applied to interpret and understand the predictions made by the Random Forest model. By using XAI, we can provide transparency into how the model arrives at its decisions, which is crucial for stakeholders, especially in high-stakes applications like loan default prediction. The process begins by loading the saved Random Forest model (`final.joblib`) and the dataset (`Loan\_default.csv`) that was used to train it. The features of interest, such as `Age`, `Income`, `LoanAmount`, and others, are defined and preprocessed. Categorical features are label-encoded to ensure compatibility with the model (Figure 27).

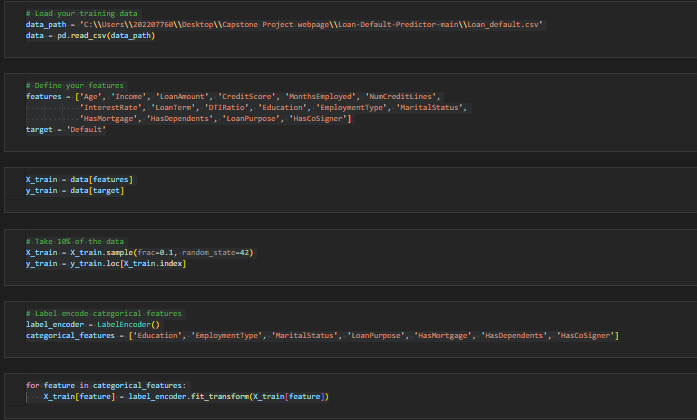


Figure 27 Implementation of Explainable AI.

#### Feature Importance

Feature importance analysis is conducted to determine which features had the most influence on the model's predictions. The importance scores are derived from the trained model, and these scores are visualized in a bar chart (Figure 28). The chart highlights the relative importance of each feature, helping to identify the key drivers behind the model's decisions. The interest rate has the highest importance, indicating it plays a crucial role in the model’s predictions. This suggests that the interest rate is a key factor in determining whether a loan will default or not. Age is the second most important feature, implying that age-related factors significantly influence the likelihood of loan default. Income also ranks highly, indicating that financial status is a critical predictor in the model. The loan amount is another essential feature, as it directly impacts the ability to repay or default on a loan.

A graph of blue bars with white text

Description automatically generated

Figure 28 Feature Importance of the model based on Random Forest

The number of months of employment is important, likely reflecting job stability and income consistency, which are crucial for loan repayment. Following months employed, the next important features are credit score, DTI (Debt-to-Income) ratio, loan term, loan purpose, and the number of credit lines. Credit score is crucial as it directly reflects creditworthiness. The DTI ratio is significant as it measures the individual’s ability to manage monthly payments and repay debts. Loan term affects repayment schedules and default risk, while loan purpose can provide insights into the reasons for borrowing and potential risk factors. The number of credit lines indicates the individual’s credit activity and management, which can impact their ability to handle additional debt.

#### Decision Paths for Individual Predictions

The decision paths for individual predictions are examined by extracting and printing the decision rules from a few trees in the Random Forest (Figure 29). This step provides a clear and interpretable view of how specific features influence the outcome in individual trees, offering insights into the model's decision-making process.

A screen shot of a computer

Description automatically generated

Figure 29 Decision Paths for individual Prediction based on Random Forest.

Each decision path represents a sequence of conditions that an individual data point must satisfy to reach a specific prediction. These paths are visualized as decision trees, where each node represents a feature and a threshold value. For example, one path might start with checking if the Employment Type is less than or equal to 0.50, followed by checking if Income is less than or equal to 38270, and so on. Each subsequent node further refines the decision based on additional features such as Age, DTI Ratio, Education, Marital Status,and Loan Purpose.

As the data point traverses down the tree, it encounters various conditions that either lead it to the next node or to a leaf node, which represents the final prediction. For instance, if the Income is less than or equal to 27258.50 and the DTI Ratio is less than or equal to 0.45, the path continues to check other features like Education and Marital Status. If all conditions are met, the path concludes with a prediction, such as classifying the individual as a non-defaulter (class 0.0) or a defaulter (class 1.0).

#### Permutation Feature Importance

Permutation importance is another method used to assess feature importance. By randomly shuffling the values of each feature and measuring the impact on model performance, this approach provides an alternative perspective on feature relevance. The results are visualized in a bar chart (Figure 30).

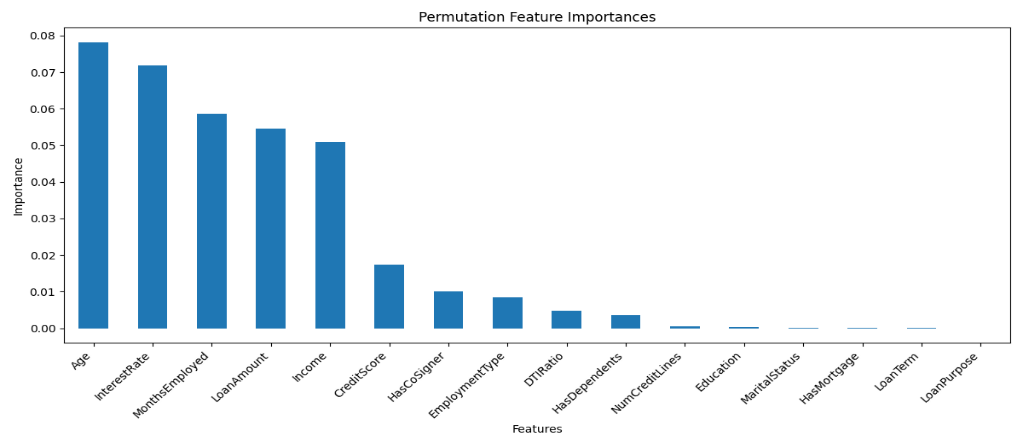


Figure 30 Permutation Feature Importance base on Random Forest.

The chart above shows that age, interest rate, and months employed are the most influential features, with age having the highest importance score. This suggests that age-related factors, interest rates, and job stability significantly impact the model’s predictions. Other important features include loan amount, income, credit score, presence of a co-signer, employment type, Debt-to-Income (DTI) ratio, loan term, loan purpose, and the number of credit lines. These features collectively help the model make accurate predictions by capturing various aspects of financial stability and creditworthiness.

#### Partial Dependence Plots (PDPs)

Partial Dependence Plots are generated for each feature to show the relationship between a feature and the predicted outcome while holding other features constant. These plots help to visualize the marginal effect of each feature on the model's predictions, offering a more nuanced understanding of how changes in a feature influence the model's decisions. Table 6 depict Partial Dependence Plots for different features in the dataset along with comments about observations.

Table 6 Partial Dependence Plots (PDPs) for different features based on Random Forest

|  |  |
| --- | --- |
| **Age** | **Observations** |
|  | The probability of loan default decreases as the borrower's age increases. The plot highlights a steep decline in default probability for younger borrowers, which stabilizes as they get older. |
| **Income** | **Observations** |
| A graph with a line  Description automatically generated | higher income levels are associated with a lower probability of loan default. As income rises, the ability to meet loan obligations improves, leading to a gradual decrease in default probability. |
| **Loan Amount** | **Observations** |
| A graph with a line going up  Description automatically generated | Loan Amount reveals a positive correlation with the probability of loan default. Larger loan amounts may place more financial strain on borrowers, increasing the likelihood of default. The plot clearly shows that as the loan amount increases, so does the risk of default. |
| **Credit Score** | **Observations** |
| A graph with a line  Description automatically generated | Demonstrates a strong negative correlation with loan default probability. Borrowers with higher credit scores, indicative of good creditworthiness, are less likely to default on their loans. The plot underscores the importance of maintaining a good credit score. |
| **Months Employed** | **Observations** |
| A graph with a line  Description automatically generated | Longer employment histories are linked to a lower probability of loan default. Stable employment provides financial security, which is reflected in the decreasing default probabilities as months of employment increase. |
| **Number of Credit Lines** | **Observations** |
| A graph with a line  Description automatically generated | Positive relationship with loan default probability. As the number of credit lines increases, the probability of default increase |
| **Interest Rate** | **Observations** |
| A graph with a line  Description automatically generated | Higher interest rates correlate with a higher probability of loan default. Increased interest rates elevate the financial burden on borrowers, making defaults more likely. |
| **Loan Term** | **Observations** |
|  | As the Loan Term increases, the probability of default increases to a certain extend of Loan term (months). At approximately 35 months then its start to decrease further from that. |
| **Debt to Income Ratio** | **Observations** |
| A graph of a line  Description automatically generated with medium confidence | * Shows a strong positive correlation with loan default probability. A higher DTI ratio suggests that a borrower is allocating a large portion of their income to debt payments, which increases the likelihood of default. |
| **Education** | **Observations** |
| A graph with a line  Description automatically generated | * For Education suggests that higher education levels are associated with a lower probability of loan default. Educated individuals may have better job opportunities and financial management skills, which reduce their default risk |

#### LIME

LIME (Local Interpretable Model-agnostic Explanations) is an explainability tool that helps to understand the predictions made by complex machine learning models. It works by perturbing the input data slightly and observing how the model’s predictions change. LIME generates a locally interpretable model (such as a linear model) around the specific instance in question, giving insight into which features are most influential for that particular prediction. This makes it useful for understanding black-box models, such as Random Forests or Neural Networks, by simplifying their decisions for specific predictions.

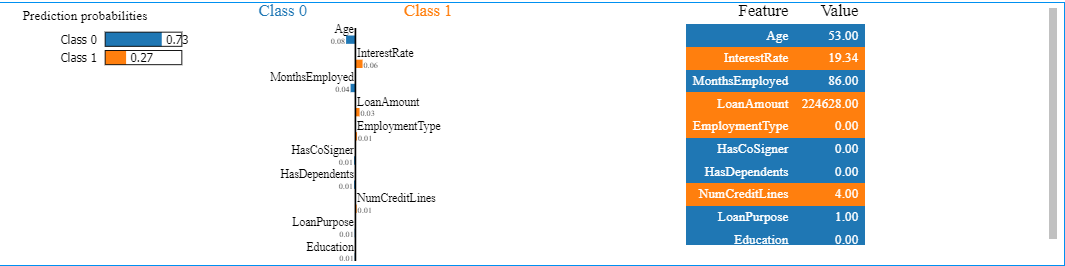


Figure 31 Using LIME for explainability and interpretability of the model.

In (Figure 31) are LIME results for loan default prediction, the model predicted that the applicant has a 73% chance of not defaulting (class 0) and a 27% chance of defaulting (class 1). The listed features, such as age, interest rate, and loan amount, correspond to specific values that describe this applicant. For example, the person is 53 years old, has an interest rate of 19.34%, and has been employed for 86 months. Features like Has Cosigner and Has Dependents been 0, indicating the absence of a co-signer and dependents. LIME would highlight how these specific feature values contributed to the final prediction, identifying which features had the most influence on predicting the applicant’s likelihood of not defaulting.

In the LIME results, the most influential features predicting the applicant will not default (Class 0) are Age (0.08), Interest Rate (0.06), and Months Employed (0.04). Age, at 53, has the strongest positive impact, followed by the interest rate of 19.34% and 86 months of employment. The Loan Amount (0.03**)** also contributes but to a lesser degree. Other features, like Employment Type, Has Cosigner, has Dependents, Num Credit Lines, Loan Purpose, and Education each contribute minimally (0.01), indicating they have a smaller role in the prediction.

### Deployment

The final model (Random Forest) was deployed as a Flask web application. The web app allows users to input relevant loan and customer data to receive a prediction on whether the loan is likely to default. Below is a snippet of the Flask application's app.py (Figure 32).

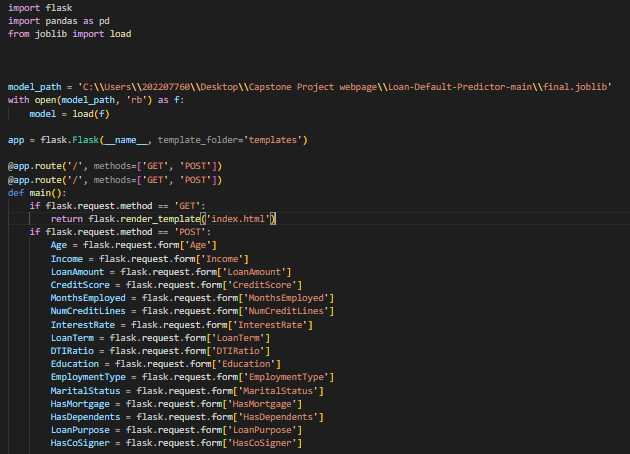


Figure 32 Snippet of Flask application

Figure 33 shows the layout of the website's home page, which includes a button labelled *Predict Approval* that redirects users to a form where they can input customer information and submit it to predict the likelihood of loan default (Figure 34). Another button, *Insight on Loan Default*, redirects users to a dashboard providing meaningful insights about loan defaults. The page also features buttons for *About Us*, *Contact Us*, and *News*.

Figure 33 depict how the home page of website looks like, it has a button called Predict Approval which redirect you to the form where a user will input information about a customer the submit to predict if they are likely to default loan (Figure 34). Another button is called insight on loan default which will redirect a user to a Dashboard where they can get meaningful insight about loan default. Other buttons are About Us, Contact Us, News.

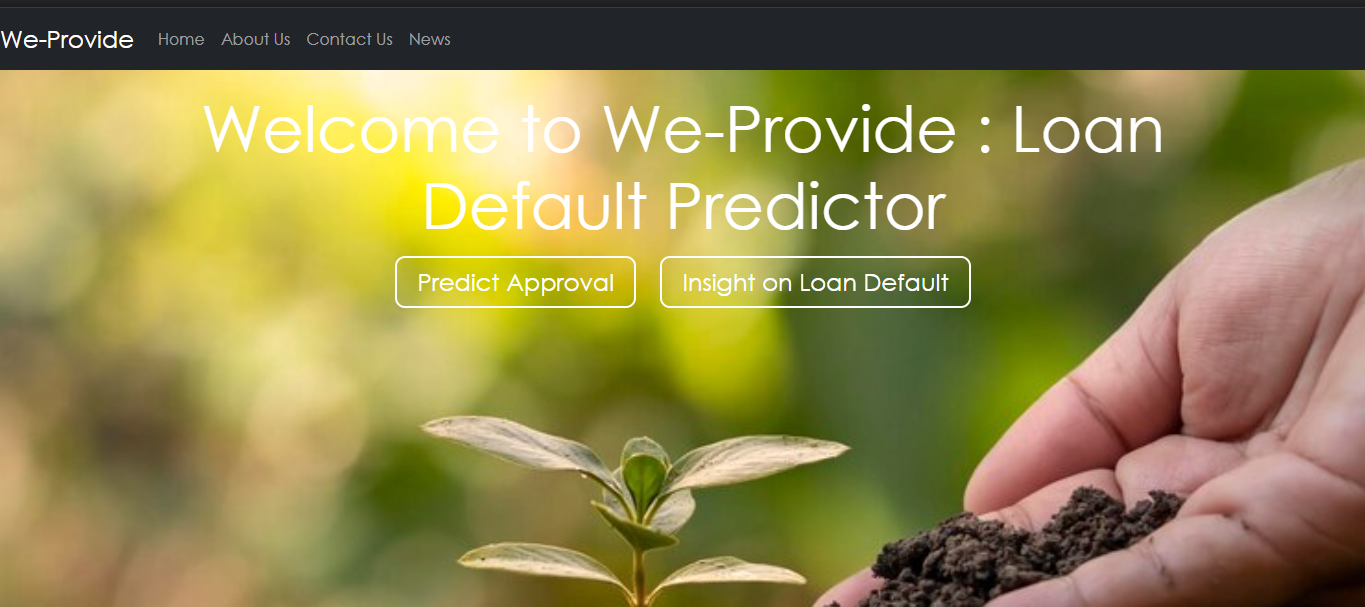


Figure 33 Home page of the Loan prediction website



Figure 34 Snippet of the Form for loan default predictor

# **RECOMMENDATIONS**

## Actionable Insights

The Random Forest model, with its near-perfect AUC score of 0.99 and outstanding precision, recall, and F1 scores, should be implemented in the loan approval process to enhance the accuracy of loan default predictions. By integrating this model, financial institutions can significantly reduce the risk of financial losses associated with defaults. Additionally, the analysis highlights the importance of focusing on high-impact features such as interest rates, age, income, and loan amounts, which play a crucial role in determining the likelihood of default. Institutions are encouraged to prioritize these features in their risk assessment processes, potentially applying stricter lending criteria to high-risk loans, particularly those with higher interest rates and larger amounts.

Furthermore, to continually improve the model's performance, it is essential to ensure accurate and comprehensive data collection, especially regarding key features like employment history, credit scores, and debt-to-income ratios. These factors have been identified as critical predictors in the loan default model. Finally, it is recommended to establish a system for continuous monitoring and updating of the predictive model to maintain its effectiveness in the face of changing economic conditions. This includes retraining the model periodically with new data to ensure it retains its predictive power over time.

## Potential Risks and Mitigation Strategies

Although the Random Forest model offers high precision, there is a risk of over-fitting training data, which could reduce its effectiveness for new invisible data. In order to address this problem, regular assessment of the model with new data is recommended, as well as the use of technologies such as cross validation to ensure that the model can be generalized far beyond the initial data set. Another potential risk is related to data privacy concerns, as the model handles sensitive personal information such as credit score and income. To address this problem, effective data protection measures such as encryption and anonymization should be implemented and strict compliance with relevant data protection regulations such as the General Data Protection Regulation (GDPR) (Deloitte, 2018).

Model biases are another critical risk where models could unintentionally perpetuate biases in training data, resulting in unfair lending decisions. Regular checks of the model are required, and training on different data sets is recommended. If necessary, fairness restrictions should be applied to prevent biased outcomes. In addition, economic fluctuations pose a risk to model accuracy, as economic changes, such as recessions, can affect model prediction power. In order to mitigate this situation, it is recommended to include macroeconomic indicators in the model and conduct scenario analysis to predict different economic conditions and their potential effects (Roijmans, 2020).

## Challenges

Some challenges must be addressed to ensure the success of the loan default prediction model and the continued performance of the loan default prediction model. An important challenge is the quality and availability of the data. Access to high-quality, comprehensive data sets, especially for new or potential customers with no credit history, can limit the accuracy of the model. Another challenge is to integrate the model into existing loan approval systems, which may require significant changes to the old IT infrastructure, leading to higher costs and longer implementation times. In addition, while the random forest model is highly precise, its complexity may make interpretation difficult, preventing trust and adoption among loan officers and decision makers. Ensure that these stakeholders understand how the model makes predictions is crucial for its successful integration into the decision-making process. Finally, compliance with regulations remains a challenge, particularly in ensuring that the model complies with financial regulations on transparency and fairness in automated decision-making. This requires continuous attention and may require modifications to the model as the regulations evolve.

# **REFERENCES**

1. Data Science Process Alliance. (2021). [www.DataScience-PM.com](http://www.DataScience-PM.com)
2. Egwa, A. A. (2022). Default Prediction for Loan Lenders Using Machine Learning Algorithms. SLU Journal of Science and Technology, 5(1 & 2), 1–12. <https://doi.org/10.56471/slujst.v5i.222>
3. Jean Ross. (2023, May 11). Default Would Have a Catastrophic Impact on the Economy. <https://www.americanprogress.org/article/default-would-have-a-catastrophic-impact-on-the-economy/#:~:text=Default%20raises%20interest%20rates%20on,including%20those%20paid%20by%20households>.
4. Li, H., & Wu, W. (2024). Loan default predictability with explainable machine learning. Finance Research Letters, 60. <https://doi.org/10.1016/j.frl.2023.104867>
5. Xianyu, Q., & Hai, M. (2023). Research on Default Prediction Model of Corporate Credit Risk Based on Big Data Analysis Algorithm. Procedia Computer Science, 221, 300–307. <https://doi.org/10.1016/j.procs.2023.07.041>
6. Zhang, L., Wang, J., & Liu, Z. (2023). What should lenders be more concerned about? Developing a profit-driven loan default prediction model. Expert Systems with Applications, 213. <https://doi.org/10.1016/j.eswa.2022.118938>
7. Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. Data Science and Management, 6(3), 123–133. <https://doi.org/10.1016/j.dsm.2023.04.003>
8. Deloitte. (2018). *deloitte-uk-the-impact-of-gdpr-on-the-financial-services*. <https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/risk/deloitte-uk-the-impact-of-gdpr-on-the-financial-services.pdf>
9. Roijmans, S. (2020). *Macroeconomic factors in loan default prediction A machine learning based approach*. <https://www.netspar.nl/assets/uploads/P20201215_MSc016_Roijmans.pdf>
10. Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning ——a case study of bank loan data. *Procedia Computer Science*, *174*, 141–149. <https://doi.org/10.1016/j.procs.2020.06.069>

# **APPENDIX**

GitHub: <https://github.com/Trueprince/Loan-default-Predictor--Capstone-Project/tree/master>