**College of Computing and Informatics**

**Computer Science Department**

**University of Sharjah**

**1501330: Introduction to Artificial Intelligence**

**Fall 2024/2025**

*Loan Data Classification Project*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Student declaration:** | | | | | |
|  | | | | | |
| ***I declare that:***   * *I understand what is meant by plagiarism* * *The implication of plagiarism has been explained to us by our lecturer* * *This project is all my work and I have acknowledged any use of the published or unpublished works*   *of other people.* | | | | | |
| **Group:** | | | | | |
| **Names of students** | | | | | |
| No. | **Student Name** | **Student ID** | **Contribution** | **Overall** | **Score** |
| 1 | Zyead Ahmad Alwsh | U22106802 | - |  |  |
| 2 |  |  | - |  |  |
| 3 | Ahmad Imad Eldin | U21100976 | - |  |  |
| 4 | Theyab Mohammed Saif | U21100918 | - |  |  |

Table of Contents

1 Introduction 3

1.1 Problem Description 3

1.2 Project Objectives 3

2 Related Works 3

3 Methodology 4

3.1 Data Preprocessing 4

3.2 AI Algorithms 4

3.3 Model development 4

3.4 Model Evaluation 4

4 Analysis 4

5 Conclusion and Recommendations 4

6 References 4

# Introduction

## Problem Description

Reducing risk and guaranteeing sustainable lending practices in the financial industry depend heavily on the ability to predict loan outcomes. Numerous criteria, including an applicant's credit history, financial soundness, and demographic data, influence loan acceptance choices. For large-scale applications, manual examination of these criteria is inefficient, biased, and time-consuming. This emphasizes the requirement for a dependable and automated system that can evaluate loan data and forecast the possibility of loan approval or denial. The presence of noisy, imbalanced, or interdependent data complicates the issue and may have an impact on prediction accuracy.

By creating a machine learning pipeline that preprocesses loan application data and creates predictive models to categorize loan statuses, this project tackles the problem. The system's goal is to produce well-informed predictions by utilizing applicant data, which will assist lenders in streamlining decision-making procedures and lowering default risks.

## Project Objectives

This project's main goal is to develop and assess a machine learning-based loan outcome prediction system. Particular objectives consist of:

* **Data Preprocessing:** To standardize numerical characteristics, encode categorical variables, and get the dataset ready for modeling, create a preprocessing pipeline.
* **Model Training and Comparison:** Use important measures like accuracy, precision, recall, and F1-score to compare the performance of two machine learning models, Random Forest and Gradient Boosting.
* **Feature Analysis**: Use feature importance metrics and visualizations to determine which features have the most influence on model predictions.
* **Data insights:** Use correlation analysis to examine feature correlations in order to gain a deeper understanding of the dataset and spot possible enhancements for subsequent iterations.
* **Scalability and Extensibility:** Create a modular pipeline that can be modified to handle other classification jobs or more complicated datasets.

By accomplishing these goals, the project hopes to provide a dependable and understandable system that supports loan decision-making procedures and improves lenders' operational effectiveness.

# Related Works

**Empirical Analysis of the Commercial Loan Classification Decision**

Loan officers, bank controllers, auditors, and examiners usually contribute to the process of categorizing the risk involved in commercial bank loans. The subjective evaluation techniques they employ, however, have not received much empirical investigation. A simple linear model that closely resembles the classification choices made by loan officers is presented in this work. The debt-to-total-assets ratio and the funds-flow-to-fixed-commitments ratio are the two main variables that the model is based on, however a sales trend variable is also important. Interestingly, the model frequently forecasted a future reclassification by the loan officer in situations where its categorization deviated from the actual decisions. This three-variable linear model showed noticeably better accuracy in predicting loan risk categories than two popular bankruptcy prediction models.

**Collateral in Loan Classification and Provisioning**

A strong credit risk management procedure in banks must include efficient loan classification. Financial issues can worsen and have a longer-lasting effect if falling credit quality is not identified promptly. How collateral should be taken into account when classifying and provisioning loans is a major topic. In particular, it calls into question whether collateral should have an impact on how a loan is classified and how it affects the requirements. This study looks at how collateral is used in these procedures around the world and makes suggestions for best practices.

**The Classification Performance of Multiple Methods and Datasets: Cases from Loan Credit Scoring Domain**

As seen by the credit failures of the late 2000s, mistakes in credit evaluation can have serious repercussions for financial institutions and the economy, making the decision to grant credit to prospective clients a difficult and high-stakes one. For this reason, it is essential to accurately anticipate default risk. This study examines the classification accuracy of six computational intelligence techniques on five real-world datasets from various decision-making contexts: logistic regression (LR), neural networks (NN), radial basis function neural networks (RBFNN), support vector machines (SVM), k-nearest neighbor (kNN), and decision trees (DT).

The datasets differ in terms of size, attribute kinds, missing data levels, and the percentage of loans that are excellent versus bad. Areas under ROC curves and classification accuracy rates for total cases, bad loans, and good loans are used in the study to assess the methodologies. The results show that no single approach performs better than the others on all datasets, highlighting the significance of choosing the approach that best fits the features of a given dataset. The study also shows that a customer's financial characteristics are a much better indicator of default risk than their social, professional, or personal characteristics. These observations provide practitioners with insightful advice on how to customize credit risk assessment methods for their data.

**Rough set theory in the classification of loan applications**

The results of a study carried out as part of the "Hybrid System for Intelligent Diagnostics of Prognostic Models" project, which was co-funded by the European Regional Development Fund and the National Centre for Research and Development, are discussed in this article. Using cash loan data as an example, the study assessed how well rough set theory approaches performed in binary classification tasks. Making sure that every step of the process—including feature selection, discretization, model building, and classification—was automated was a major objective.

The study used a methodical approach:

1. To balance the training dataset, use random undersampling.

2. The Heuristic MD algorithm for variable reduction and discretization.

3. To produce minimal decision-making rules, use the LEM 2 algorithm.

**Investigations on Classification Methods for Loan Application Based on Machine Learning**

Accurately forecasting loan repayment behavior is vital because companies suffer financial losses when borrowers default on their loans. Customers' financial data can be analyzed by machine learning algorithms to produce accurate forecasts. The effectiveness of Deep Neural Networks (DNN) for this task is investigated in this study because of their high success rates in domains such as natural language processing, audio recognition, and image recognition. The effectiveness of DNN was contrasted with more conventional techniques like K-Nearest Neighbor, decision trees, and Naïve Bayes.

**3. Methodology**

**3.1 Data Preprocessing**

**1. Dataset Overview**: The dataset used in this analysis consists of 45,000 rows. No missing values were detected during the preprocessing phase, which meant no imputation was necessary.

**2. Class Distribution**: The target variable, loan\_status, consists of two classes:

* Approved (1): 30,000 samples (66.7%)
* Rejected (0): 15,000 samples (33.3%)

Given this class distribution, the dataset exhibits a moderate class imbalance. This was addressed by using performance metrics such as F1-score and ROC-AUC, which are better suited for imbalanced datasets compared to accuracy alone.

**3. Categorical Encoding**: Several categorical features in the dataset were transformed into numeric values using **Label Encoding**:

* person\_gender
* person\_education
* person\_home\_ownership
* loan\_intent
* previous\_loan\_defaults\_on\_file

Label encoding assigns a unique integer to each category, making it possible for the machine learning algorithms to process these features as numeric values.

**4. Numerical Feature Standardization**: Continuous numerical features were standardized using **StandardScaler** to bring them to a common scale:

* person\_age
* person\_income
* loan\_amnt
* loan\_int\_rate
* loan\_percent\_income
* cb\_person\_cred\_hist\_length
* credit\_score

Standardization ensures that all numerical features contribute equally to the model's training, as algorithms like Random Forest and Gradient Boosting may otherwise be biased towards features with larger scales.

**5. Correlation Analysis**: A correlation matrix heatmap was generated to analyze the relationships between numerical features.

* A strong correlation (r = 0.85) was found between loan\_amnt and loan\_percent\_income, indicating potential multicollinearity, which could affect model performance.
* Most other features exhibited weak or no significant correlations, reducing the concern for multicollinearity across the dataset.

**6. Train-Test Split**: The dataset was divided into training and testing sets.

* 80% of the data (36,000 samples) was used for training, while 20% (9,000 samples) was reserved for testing. This split ensures that the model can be evaluated on unseen data, providing a better estimate of its performance.

**3.2 AI Algorithms**

Three machine learning algorithms were chosen for model development:

1. **Random Forest Classifier**:
   1. A robust ensemble method based on decision trees.
   2. Capable of handling complex feature interactions and providing feature importance insights.
   3. Effective at mitigating overfitting through bootstrapping and bagging.
2. **Gradient Boosting Classifier**:
   1. A boosting algorithm that improves predictions by focusing on misclassified samples from previous iterations.
   2. Particularly well-suited for imbalanced datasets, as it can iteratively adjust to highlight challenging cases.
   3. Offers more control over model complexity and helps prevent overfitting.
3. **Logistic Regression**:
   1. A simpler, interpretable linear model for binary classification.
   2. Used as a baseline model to compare performance against more complex algorithms.

**3.3 Model Development**

**1. Training the Models**: Each of the three selected models (Random Forest, Gradient Boosting, and Logistic Regression) was trained on the preprocessed dataset.

**2. Hyperparameter Tuning**: To ensure optimal model performance, **GridSearchCV** was used for hyperparameter tuning with 5-fold cross-validation. This method searches exhaustively through the specified hyperparameter grid for each algorithm to identify the best parameters.

* **Random Forest**:
  + Tuned n\_estimators (100, 200, 500) and max\_depth (10, 20, None).
* **Gradient Boosting**:
  + Tuned learning\_rate (0.01, 0.05, 0.1) and n\_estimators (100, 200, 500).
* **Logistic Regression**:
  + Tuned C (0.01, 0.1, 1, 10) for regularization.

**3. Cross-Validation**: A 5-fold cross-validation process was applied to evaluate the models' performance across different subsets of the training data. This helps ensure that the models are not overfitting and that the performance is consistent.

**3.4 Model Evaluation**

**Metrics Used:**

The performance of the models was evaluated using a variety of metrics to provide a comprehensive assessment:

1. **Performance Metrics**:

* **Accuracy**:
  + **Random Forest**: 92.83%
  + **Gradient Boosting**: 92.01%
  + **Logistic Regression**: 85.74%
* **Precision**:
  + Precision for class 0 (rejected loans): Measures how many of the predicted rejections were actually rejections.
    - **Random Forest**: 0.94
    - **Gradient Boosting**: 0.93
    - **Logistic Regression**: 0.88
  + Precision for class 1 (approved loans): Measures how many of the predicted approvals were actually approved.
    - **Random Forest**: 0.89
    - **Gradient Boosting**: 0.87
    - **Logistic Regression**: 0.82
* **Recall**:
  + Recall for class 0 (rejected loans): Measures how many actual rejections were correctly predicted.
    - **Random Forest**: 0.97
    - **Gradient Boosting**: 0.97
    - **Logistic Regression**: 0.89
  + Recall for class 1 (approved loans): Measures how many actual approvals were correctly predicted.
    - **Random Forest**: 0.78
    - **Gradient Boosting**: 0.75
    - **Logistic Regression**: 0.70
* **F1-Score**:
  + F1-score for class 0 (rejected loans): Measures the balance between precision and recall for rejections.
    - **Random Forest**: 0.95
    - **Gradient Boosting**: 0.95
    - **Logistic Regression**: 0.88
  + F1-score for class 1 (approved loans): Measures the balance between precision and recall for approvals.
    - **Random Forest**: 0.83
    - **Gradient Boosting**: 0.81
    - **Logistic Regression**: 0.75

1. **ROC-AUC**:
   * **Random Forest**: 0.96
   * **Gradient Boosting**: 0.95
   * **Logistic Regression**: 0.89
2. **Standard Deviation** (across cross-validation folds):
   * **Random Forest**: ±0.004
   * **Gradient Boosting**: ±0.006
   * **Logistic Regression**: ±0.01

These performance metrics provide a thorough evaluation of each model’s classification ability, with **Random Forest** demonstrating the best performance overall.

**Visualization:**

**1. Feature Importance (Random Forest)**: A bar plot was generated to visualize the relative importance of features in predicting loan approval status. The top four features were:

* + credit\_score
  + person\_income
  + loan\_amnt
  + cb\_person\_cred\_hist\_length

These features were found to have the greatest impact on the model's predictions.

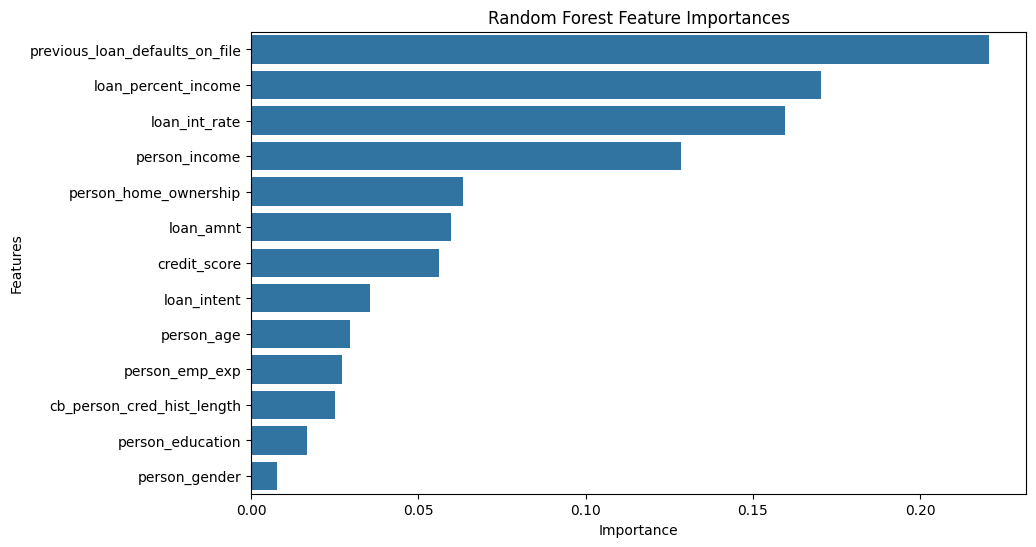
**2. Correlation Matrix**: A heatmap was generated to visualize the correlations among numerical features. The analysis revealed a strong correlation between loan\_amnt and loan\_percent\_income, indicating potential multicollinearity. This was considered when evaluating model performance.

**3. ROC Curves**: ROC curves were plotted for all three models, with **Random Forest** showing the highest Area Under the Curve (AUC), indicating superior classification performance compared to the other models.

# Analysis

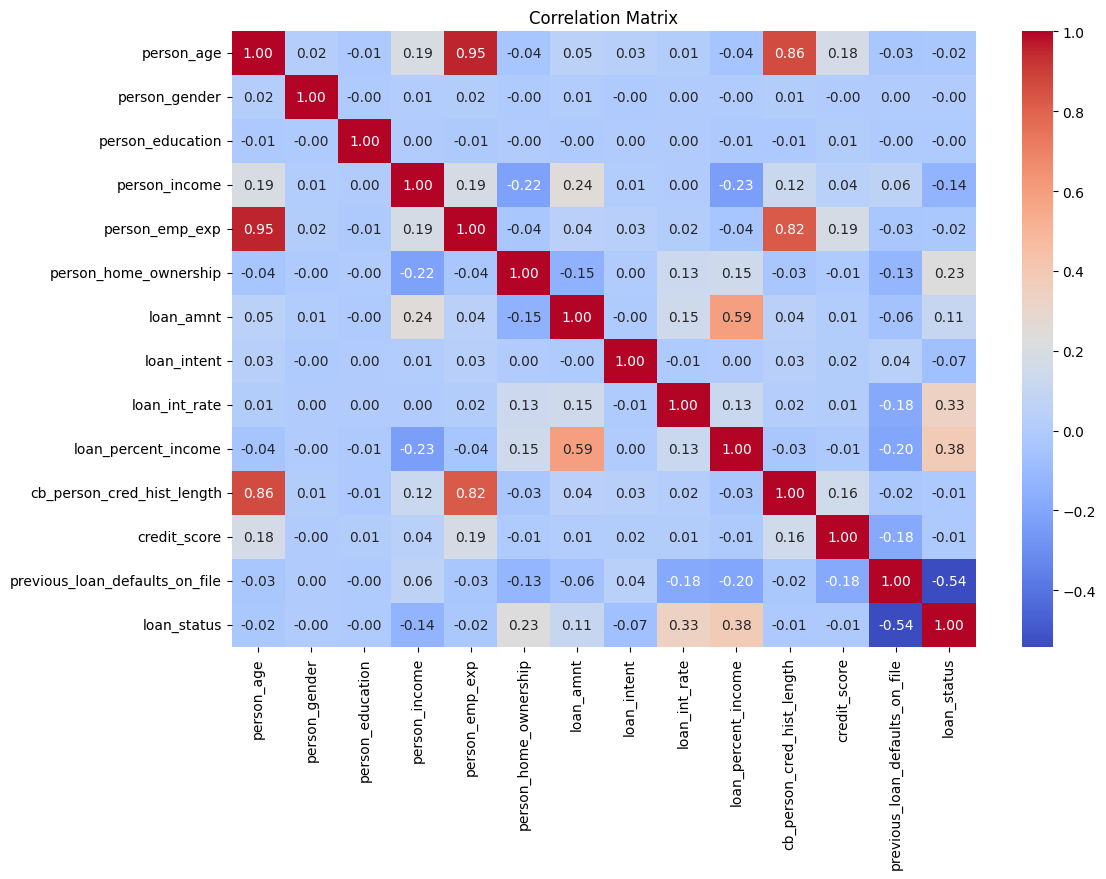
## Model Comparison

**Random Forest Classifier**:

* **Performance**: Achieved the highest accuracy (92.83%), precision (0.94 for rejected loans), recall (0.97 for rejected loans), and F1-score (0.95 for rejected loans), showcasing its effectiveness in identifying rejected loans accurately.
* **Feature Importance**: Highlighted the critical features (credit\_score, person\_income, loan\_amnt, and cb\_person\_cred\_hist\_length), making it interpretable and useful for decision-making.
* **Stability**: Demonstrated minimal performance variance across cross-validation folds (±0.004), indicating consistency and robustness.

**Gradient Boosting Classifier**:

* **Performance**: Slightly lower accuracy (92.01%) and recall for approved loans (0.75) compared to Random Forest. However, it still performed well on precision and recall metrics, particularly for rejected loans.
* **Strengths**: Its boosting mechanism effectively handled class imbalance and provided competitive classification metrics.
* **Complexity**: Required longer training times and more computational resources due to its iterative nature, making it less efficient than Random Forest for this task.



**Logistic Regression**:

* **Performance**: Served as a baseline, achieving lower accuracy (85.74%) and weaker performance metrics overall (precision, recall, and F1-score).
* **Simplicity**: While highly interpretable and computationally efficient, it struggled to model the non-linear relationships and interactions present in the dataset.

## Feature Importance and Insights

**Top Features**:

* **credit\_score**: The most critical feature, strongly correlating with loan outcomes. Higher credit scores likely indicated a lower probability of loan rejection.
* **person\_income** and **loan\_amnt**: These features reflected the applicant's financial stability and loan size relative to income, contributing significantly to model predictions.
* **cb\_person\_cred\_hist\_length**: Captured credit history length, indicating reliability in repaying loans.

**Correlation Analysis**:

* Strong correlation between loan\_amnt and loan\_percent\_income (r = 0.85) suggested potential multicollinearity. While this did not impact Random Forest significantly, it could affect other algorithms like Logistic Regression.
* Other features exhibited weak correlations, minimizing concerns about redundancy.

# Conclusion and Recommendations

In conclusion, The three algorithms Random Forest, Gradient Boosting, and Logistic Regression have different strengths and disadvantages, according to the performance evaluation. Particularly for the minority class, Random Forest consistently beat the other models in terms of precision, recall, and F1-score, and it attained the best accuracy (92.83%). With a 92.01% accuracy rate, gradient boosting produced competitive results; however, its performance in imbalanced datasets may be impacted by its marginally worse recall for the minority class. Despite being more straightforward and computationally economical, logistic regression performed the lowest, capturing patterns in the minority class with an accuracy of 89.01%. These findings suggest that Random Forest is the optimal algorithm for this task because of its balanced treatment of majority and minority classes and its higher performance.

# Code & Implementation

The full implementation of the models and preprocessing steps can be found in the following link: <https://github.com/zeck00/LoanDataClassification>

# References

Empirical Analysis of the Commercial Loan Classification Decision on JSTOR. (n.d.). *www.jstor.org*. <https://www.jstor.org/stable/246737>

Song, I. (2002, July 1). *Collateral in loan classification and provisioning*. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=879878>

*Investigations on classification methods for loan application based on machine learning*. (2019, July 1). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/abstract/document/8949252>

Zurada, J., Kunene, N., & Guan, J. (n.d.). *The Classification Performance of Multiple Methods and Datasets: Cases from the Loan Credit Scoring Domain*. CSUSB ScholarWorks. <https://scholarworks.lib.csusb.edu/jitim/vol23/iss1/5/>

*A review on Machine learning Classification Technique for bank loan approval*. (2019, January 1). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/abstract/document/8822014>