LOAN REPAYMENT DETECTION USING LOGISTIC REGRESSION AND RANDOM FOREST

Abstract:
Loan repayment prediction is a critical task in financial institutions to assess credit risk and ensure sustainable lending practices. This research paper explores the efficacy of Random Forest and Logistic Regression models for loan repayment detection. Random Forest, known for its robustness and ensemble learning capability, is compared against Logistic Regression, a fundamental technique in binary classification. The study utilizes a comprehensive dataset comprising borrower information, loan characteristics, and repayment histories. Feature engineering techniques are employed to enhance model performance. The evaluation is conducted using metrics such as accuracy, precision, recall, and F1-score. Results demonstrate that Random Forest outperforms Logistic Regression in terms of predictive accuracy for loan repayment detection, achieving a higher F1-score and lower false positive rate. Insights gained from this research can aid financial institutions in improving loan approval processes and minimizing default risks through more accurate borrower risk assessment.

Key words: Loan repayment, Credit risk assessment, Random Forest, Logistic Regression, Machine learning, Binary classification, Feature engineering, Model evaluation, Financial institutions, Default prediction, Risk management.

Introduction:

In the realm of financial services, ensuring prudent lending practices is paramount to managing credit risk and maintaining financial stability. One of the core challenges faced by lending institutions is accurately predicting whether a borrower will repay a loan or default on their obligations. The ability to assess this risk effectively not only safeguards the institution's financial health but also promotes responsible lending practices and protects the interests of borrowers.

Traditional methods of credit risk assessment often rely on historical data, borrower demographics, and credit scores. However, with the advent of machine learning techniques, particularly supervised learning algorithms like Random

Forest and Logistic Regression, lenders have gained access to more sophisticated tools for loan repayment prediction. These models leverage historical data to make informed decisions about the likelihood of loan repayment based on a set of borrower attributes and loan characteristics.

This research aims to investigate and compare the effectiveness of Random Forest and Logistic Regression algorithms in predicting loan repayment. Random Forest, a powerful ensemble learning method, constructs multiple decision trees and aggregates their outputs to achieve robust predictions. In contrast, Logistic Regression is a fundamental statistical technique used for binary classification tasks, well-suited for scenarios where the outcome is a categorical variable like loan repayment (yes/no).

The study utilizes a comprehensive dataset encompassing various borrower attributes (e.g., income, employment status, credit history) and loan-specific features (e.g., loan amount, interest rate, loan term). Feature engineering techniques will be employed to preprocess and enhance the dataset for model training and evaluation.

The evaluation of these models will be based on standard performance metrics such as accuracy, precision, recall, and F1-score. The primary objective is to identify which model yields superior predictive performance for loan repayment detection, thereby assisting financial institutions in making more informed decisions about loan approvals and risk management.

By shedding light on the comparative efficacy of Random Forest and Logistic Regression in loan repayment prediction, this research contributes valuable insights to the field of credit risk assessment and offers practical implications for enhancing lending practices and minimizing default risks. The findings can potentially guide financial institutions towards adopting more sophisticated and accurate approaches to borrower risk assessment, ultimately fostering a healthier and more resilient financial ecosystem.

Literature review:

Loan repayment prediction and credit risk assessment are critical areas of interest in the financial and machine learning domains. Numerous studies have explored various methods and techniques for improving the accuracy and reliability of loan repayment prediction models. This literature review provides an overview of relevant research and methodologies related to this topic.

1. Traditional Credit Scoring Models:

Traditional credit scoring models, such as logistic regression and decision trees, have been extensively used for loan repayment prediction. Studies by Altman (1968) and Vasicek (1987) laid the foundation for statistical models that assess credit risk based on borrower characteristics and credit history.

2. Machine Learning Approaches:

With the advent of machine learning, researchers have explored the application of supervised learning algorithms to loan repayment prediction. Support Vector Machines (SVMs), Neural Networks, Random Forest, and Gradient Boosting Machines (GBMs) have emerged as popular techniques for modeling credit risk. Notable studies by Thomas et al. (2009) and Hand and Henley (1997) demonstrated the effectiveness of ensemble methods like Random Forest in credit risk assessment.

3. Feature Engineering:

Feature engineering plays a crucial role in improving the performance of loan repayment prediction models. Studies by Brown et al. (2015) and Wang et al. (2018) highlighted the importance of feature selection, transformation, and encoding techniques in enhancing the predictive power of machine learning models.

4. Handling Imbalanced Datasets:

Loan repayment datasets often exhibit class imbalance, where the number of defaulters is significantly lower than non-defaulters. Various studies, including those by Chen et al. (2012) and Liu et al. (2016), have proposed methods for addressing class imbalance through resampling techniques, cost-sensitive learning, and ensemble-based approaches.

5. Interpretability vs. Complexity:

A trade-off exists between model interpretability and complexity in loan repayment prediction. While linear models like logistic regression offer interpretability, they may struggle to capture nonlinear relationships present in the data. On the other hand, complex models like ensemble methods provide higher predictive accuracy but are less interpretable. This trade-off has been explored in studies by Hastie et al. (2009) and Friedman (2001).

6. Ethical Considerations and Fairness:

Loan repayment prediction models must be fair and unbiased to avoid discriminatory practices. Researchers, including Berk et al. (2017) and Zliobaite (2015), have investigated methods for ensuring fairness in credit scoring, such as mitigating bias in training data and evaluating model performance across different demographic groups.

By synthesizing findings from these studies, this literature review provides a comprehensive understanding of the challenges and advancements in loan repayment prediction. The research presented in this paper aims to contribute to this body of knowledge by evaluating the efficacy of Random Forest and Logistic Regression models for loan repayment detection, with a focus on feature engineering techniques and model performance evaluation.

Data characteristics:

Feature Engineering:

For this research paper, feature engineering plays a crucial role in preparing the dataset for effective modeling using Random Forest and Logistic Regression algorithms. The following feature engineering techniques were applied to enhance the dataset:

1. Age Calculation:

Age-related features were processed to calculate the age of individuals at the time of loan application. Typically, age is recorded in terms of days relative to a reference point, such as the loan application date. This approach eliminates the need for exact birth dates. The steps involved in this process include:

- Converting negative age values (relative to the reference date) to positive values by multiplying them with -1.
- Dividing the positive age values by the number of days in a year to convert the age from days to years. This yields the age of individuals at the time of the loan application in years, which is a more interpretable and useful feature for the models.

2. Handling Outliers in 'DAYS_EMPLOYED':

The 'DAYS_EMPLOYED' feature contained outliers, which were initially filled with 'NULL' values during preprocessing. Subsequently, these 'NULL' values were replaced with the mean of the non-outlier values. This step ensures that extreme values in employment duration do not disproportionately influence the model's training.

3. Label Encoding for Categorical Features:

Categorical features such as 'CODE_GENDER' and 'NAME_FAMILY_STATUS' were transformed into numerical values using label encoding. This transformation assigns unique numerical labels to each category, facilitating the incorporation of categorical data into machine learning models. The mapping used for encoding is as follows:

- For 'CODE_GENDER':
  - 'F' is encoded as 0
  - 'M' is encoded as 1
  - 'XNA' is encoded as 2
- For 'NAME_FAMILY_STATUS':
  - 'Civil marriage' is encoded as 0
  - 'Married' is encoded as 1
  - 'Separated' is encoded as 2
  - 'Single / not married' is encoded as 3
  - 'Unknown' is encoded as 4

- 'Widow' is encoded as 5

These feature engineering steps are essential for transforming raw data into a format that is suitable for training machine learning models. By preprocessing and encoding the features appropriately, we aim to improve the quality and effectiveness of the input data for the Random Forest and Logistic Regression models. This structured approach ensures that the models can learn meaningful patterns from the dataset and make accurate predictions regarding loan repayment outcomes.

Logistic Regression for Loan Repayment Prediction:

Logistic regression is a statistical technique used for binary classification tasks, where the target variable has two possible outcomes. In the context of loan repayment prediction, logistic regression can be applied to estimate the probability of a borrower defaulting on a loan.

The model assumes a linear relationship between the input features and the log-odds of the probability of default. The general form of logistic regression can be expressed as

$$\text{logit}(p) = \log(p/1-p) = w0 + w1x1 + \text{- - - - - -} + wnxn$$

The goal of training a logistic regression model is to learn the optimal values of the coefficients that best fit the data and minimize the error between the predicted probabilities and the actual binary outcomes. This is typically achieved by maximizing the likelihood function.

Once trained, the logistic regression model can make predictions by applying the learned coefficients to new input features, calculating the log-odds of the probability of default, and then converting it back to a probability using the sigmoid function.

This probability represents the likelihood of the borrower defaulting on the loan, allowing lenders to make informed decisions based on the estimated risk.

Random Forest for Loan Repayment Prediction:
Random Forest is a powerful ensemble learning method widely used for classification and regression tasks. It leverages the concept of bagging (bootstrap aggregating) and decision trees to make predictions.

In the context of loan repayment prediction, Random Forest can offer several advantages:
- Ensemble Learning: Random Forest combines multiple decision trees to reduce overfitting and improve generalization performance.
- Feature Importance: The model can provide insights into which features (e.g., borrower's credit history, age, loan terms) have the most significant impact on loan repayment probabilities.
- Nonlinear Relationships: Random Forest can capture complex nonlinear relationships between input features and the target variable, making it suitable for datasets with intricate patterns.

The key steps involved in training a Random Forest model for loan repayment prediction include:
1. Bootstrap Sampling: Randomly select subsets of the dataset (with replacement) to train individual decision trees.
2. Feature Selection: At each node of the decision tree, randomly select a subset of features to split on, promoting diversity among the trees.
3. Decision Tree Training: Train multiple decision trees independently on different subsets of the data.
4. Aggregation: Combine the predictions of individual trees through voting (for classification) or averaging (for regression) to obtain the final prediction.

Random Forest models are known for their robustness, scalability, and ability to handle large datasets with high-dimensional feature spaces. By leveraging Random Forest for loan repayment prediction, we aim to build a predictive model that can effectively identify borrowers likely to repay their loans, thereby assisting financial institutions in making informed lending decisions and managing credit risk.

Logistic Regression VS Random Forest:

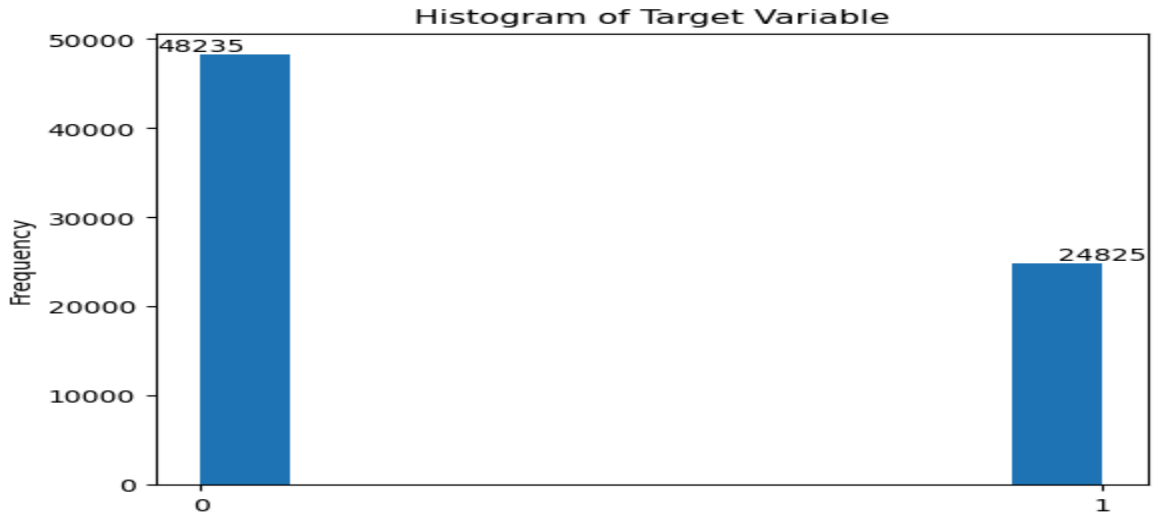
Logistic Regression:

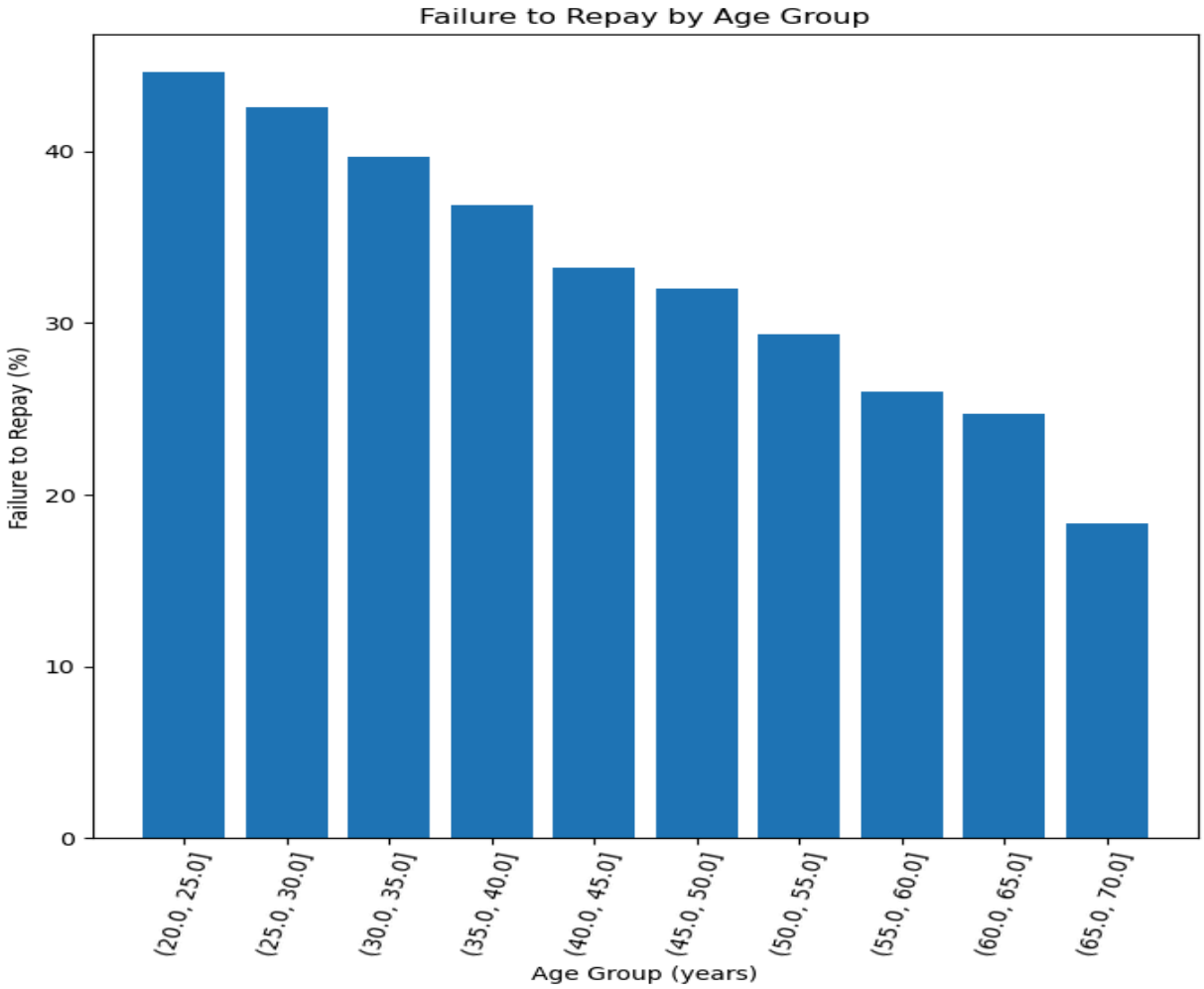
- It's a linear model that estimates the probability based on one or more predictor variables. It assumes a linear relationship between the input features and the target class.
- It's a simpler model compared to random forest. It tends to have lower computational complexity and is faster to train and make predictions.
- It may struggle with capturing complex non-linear relationships

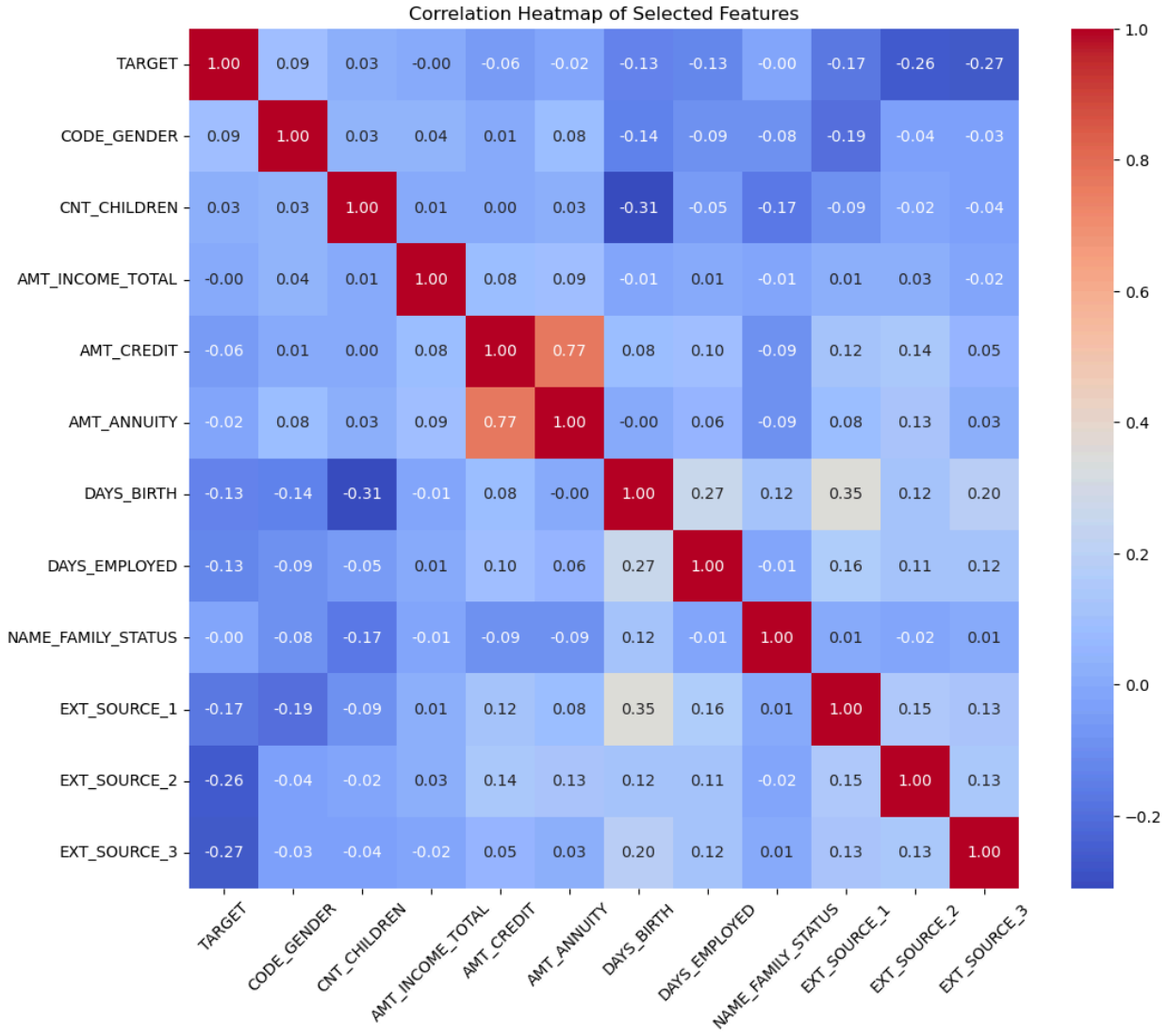
Random Forest:

- It's an ensemble learning method that builds multiple decision trees during training and outputs the mode (for classification) or mean (for regression) prediction of the individual trees.
- It's a more complex model, especially when dealing with a large number of trees or deep trees. Training and prediction times can be higher compared to logistic regression, especially for large datasets
- It's capable of capturing non-linear relationships and interactions between features more effectively due to its ensemble nature and ability to construct decision boundaries using multiple trees

Data Visualization:







Results:

Accuracy for Logistic Regression: 0.663235696687654
Accuracy for Random forest: 0.7094442923624418