Credit Risk Modeling and Assessment Using Machine Learning Techniques

Project Overview:

This project involves the development and assessment of various machine learning models for credit risk prediction. The goal is to accurately classify the creditworthiness of applicants, identifying those likely to default on loans. This project covers data preprocessing, model training, evaluation using metrics such as accuracy, ROC-AUC, and PR-AUC, and cross-validation to ensure the robustness of the models. The models used include Logistic Regression, Decision Trees, Random Forest, and Neural Networks.

1. Data Acquisition and Preparation:

The project begins with the acquisition of a dataset containing historical credit data, including features such as income, loan amount, and previous credit history. The data is thoroughly cleaned and preprocessed to handle missing values, normalize features, and encode categorical variables. This preparation is crucial for ensuring that the machine learning models receive high-quality input data.

2. Model Selection and Training:

Four machine learning models were selected for this study:

- Logistic Regression: A baseline model often used in binary classification problems, including credit risk assessment.
- Decision Tree: A non-linear model that splits the data based on feature thresholds to make predictions.
- Random Forest: An ensemble model that builds multiple decision trees and averages their predictions to improve accuracy and prevent overfitting.
- Neural Network: A complex model capable of capturing intricate patterns in the data through multiple layers of neurons.

Each model was trained on the preprocessed dataset, and hyperparameters were tuned to optimize performance.

3. Model Evaluation:

The models were evaluated using the following metrics:

- Accuracy: The proportion of correct predictions made by the model.

- ROC-AUC: The Area Under the Receiver Operating Characteristic Curve, measuring the trade-off between the true positive rate and false positive rate.

- PR-AUC: The Area Under the Precision-Recall Curve, highlighting the balance between precision and recall, particularly important in imbalanced datasets.

- Cross-Validation Accuracy: The average accuracy across multiple training and testing splits, ensuring the model's performance is consistent.

The results for each model are as follows:

- Logistic Regression: Accuracy of 73.17%, ROC-AUC of 77.74%, and PR-AUC of 82.77%.

- Decision Tree: Perfect accuracy and AUC scores, indicating potential overfitting.

- Random Forest: Near-perfect scores, with slight overfitting mitigated by the ensemble approach.

- Neural Network: High accuracy and AUC scores, demonstrating the model's ability to capture complex patterns.

4. Key Findings:

- The Logistic Regression model provided a solid baseline, though it was outperformed by more complex models.

- The Decision Tree model showed signs of overfitting, with perfect scores on the training data but likely poorer generalization.

- The Random Forest model demonstrated strong performance, benefiting from the ensemble approach to reduce overfitting.

- The Neural Network model achieved high accuracy and AUC scores, indicating its effectiveness for this classification task.

Conclusion:

This project successfully implemented and evaluated multiple machine learning models for credit risk assessment. The Random Forest and Neural Network models were particularly effective, offering high accuracy and robustness in

predicting creditworthiness. The findings demonstrate the potential of advanced machine learning techniques in enhancing the accuracy of credit risk predictions, which is crucial for financial institutions in making informed lending decisions.