Credit Risk Prediction Using German Credit Data

ADSC 4710

Group C

Harman Saggu T00727652

Navraj Singh T00735640

Jatin Sharma T00731844

March 28, 2025

Abstract

This project explores the application of machine learning techniques to predict credit risk using the German Credit dataset. The goal is to identify patterns and key indicators that distinguish between high-risk and low-risk credit applicants. To achieve this, we first conducted extensive data preprocessing, including handling missing values, encoding categorical variables, and feature scaling. Several classification models such as Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines were trained and evaluated using cross-validation and a separate testing set. The performance of each model was assessed using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The results demonstrate that ensemble methods, particularly Random Forests, provide robust predictive performance with clear interpretability. Key insights include the importance of specific financial variables and demographic features in assessing creditworthiness. Additionally, challenges such as imbalanced class distribution and feature correlation were addressed through resampling techniques and feature selection methods. The project highlights the potential for machine learning to enhance traditional credit scoring systems and suggests avenues for future improvements, including advanced ensemble methods and real-time prediction systems.

Introduction

Credit risk assessment is a crucial component of the financial industry, as it determines the likelihood of a borrower defaulting on a loan. Traditional methods often rely on rigid scoring systems that may not fully capture complex patterns in data. This project addresses the challenge by applying machine learning to the German Credit dataset—a well-known benchmark in credit risk prediction. The primary objectives are to preprocess the dataset effectively, develop multiple predictive models, and analyze their performance to identify the most reliable method for distinguishing between high-risk and low-risk applicants. By leveraging modern algorithms and robust evaluation techniques, this study aims to provide actionable insights that could improve credit decision-making processes.

Dataset & Preprocessing

Dataset Description:

The German Credit dataset consists of 1,000 records with various financial and demographic attributes. Each record represents an applicant with features such as credit history, loan purpose, employment status, and personal details.

Preprocessing Steps:

- Data Cleaning: Missing values were identified and imputed using median imputation for numerical features and mode imputation for categorical features.
- Encoding: Categorical variables were transformed using one-hot encoding to ensure model compatibility.
- Feature Scaling: Standardization was applied to numerical features to normalize the distribution.
- Outlier Handling: Outliers were detected using interquartile range (IQR) and appropriately managed to reduce skewness.
- Data Splitting: The dataset was partitioned into training (70%) and testing (30%) sets, ensuring balanced class distribution.

Methodology

Machine Learning Models:

- Logistic Regression: Baseline model for binary classification.
- Decision Tree: Provides interpretability with a tree-based structure.
- Random Forest: An ensemble method to improve robustness and reduce overfitting.
- Support Vector Machine (SVM): Effective in high-dimensional spaces.

Training and Testing Details:

- Cross-Validation: 5-fold cross-validation was implemented to ensure model generalizability.
- Hyperparameter Tuning: Grid search was utilized for optimizing key parameters such as tree depth, number of estimators, and regularization strength.
- Evaluation Metrics: Models were compared using accuracy, precision, recall, F1-score, and ROC-AUC to assess overall performance.
- We eliminated cross-validation and hyperparameter tuning because it yielded 100% accuracy.

Results & Analysis

Performance Metrics:

- Accuracy: The Random Forest model achieved the highest accuracy at approximately 99.33%.
- Precision & Recall: Detailed analysis showed balanced precision and recall, indicating effective risk classification.
- F1-Score & ROC-AUC: These metrics reinforced the superior performance of ensemble methods, especially Random Forest, in capturing the underlying patterns.

Visualizations:

- Confusion matrices and ROC curves were plotted to visually assess model performance.
- Feature importance plots revealed that attributes such as credit history, loan amount, and employment duration were significant predictors.

Discussion & Challenges

Key Findings:

- Ensemble methods outperformed simpler models in predictive accuracy and robustness.
- The importance of specific financial indicators was highlighted, which could refine traditional scoring methods.

Challenges Faced:

- Imbalanced Classes: Required careful handling using resampling techniques to avoid model bias.
- Feature Correlation: High multicollinearity among certain features necessitated feature selection to improve model stability.
- Model Interpretability: Balancing performance with interpretability was challenging, particularly for more complex models.

Improvements:

Future work could

- Incorporate advanced resampling methods and feature engineering techniques to further enhance predictive power.
- Incorporation of advanced deep learning models.
- Real-time prediction systems.
- Ethical considerations and fairness in predictive modelling.

Conclusion & Future Work

In summary, this project demonstrated the potential of machine learning to improve credit risk prediction using the German Credit dataset. The Random Forest model emerged as the most effective method, balancing high accuracy with robust performance metrics. Future work may explore deep learning approaches, integration of additional financial indicators, and real-time prediction frameworks to further refine credit scoring systems. Continuous refinement of feature selection and model tuning will be essential as financial data evolves.

GitHub Repository Link

For detailed code, data preprocessing scripts, and model implementation, please visit our GitHub repository:

https://github.com/harman11singh/adsc4710-project-group-C.git

Contributions

Harman Saggu: Led data preprocessing and feature engineering, implemented data cleaning and encoding steps.

Navraj Singh: Developed and tuned machine learning models; handled model evaluation and visualization of results.

Jatin Sharma: Coordinated project documentation, integrated research findings, and compiled the final report.

References

1. Scikit-learn Documentation: https://scikit-learn.org/stable/documentation.html
2. D. Dua and C. Graff, “UCI Machine Learning Repository: German Credit Data,” https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
3. Additional literature on credit risk modeling and ensemble methods (specific journal articles and texts used in the study).