Predicting Credit Risk Using Machine Learning

Jyostna Dontireddy, Nandhitha Emani ,Sakshitha Gopu ,Geetha Reddy Munnangi

Syed Jawad Hussain Shah , Instructor
Department of Computer Science in Data Science

Abstract

This project is mainly focused on predicting credit risk using machine learning techniques. By analyzing loan application data, we aim to classify applicants as low or high risk based on key financial and demographic features. The dataset undergoes preprocessing, including handling missing values, encoding categorical variables, and feature scaling. Several classification models, such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines, are trained and evaluated using accuracy, precision, recall, and AUC-ROC. The results indicate that Random Forest outperforms other models in predicting credit risk. This study demonstrates the effectiveness of data-driven approaches in enhancing loan approval decisions for financial institutions.

Introduction

- Credit risk assessment is a critical process in the financial sector, as it helps banks and lending institutions minimize losses by identifying high-risk loan applicants. Traditional credit evaluation methods rely on manual assessments and predefined financial metrics, which may not always provide accurate risk predictions.
- With the rise of data science and machine learning, automated risk prediction models have emerged as a powerful alternative, offering improved accuracy and efficiency. By analyzing historical loan data, machine learning models can identify key risk factors and predict whether an applicant is likely to default.
- This project applies various classification algorithms to predict credit risk based on financial and demographic features. The study explores different machine learning techniques, compares their performance, and highlights the most effective model for risk assessment. The ultimate goal is to provide a data-driven approach that can assist financial institutions in making informed lending decisions.

Methodology

- **Data Collection**: The dataset contains attributes like income, credit history, loan amount, employment status, and debt-to-income ratio.
- **Data Preprocessing**: Missing values were imputed, categorical variables were encoded, and numerical features were scaled.
- **Model Development**: Logistic Regression, Decision Trees, Random Forest, and SVM were used for classification.
- **Evaluation Metrics:** Models were assessed using metrics like accuracy, precision, recall, and AUC-ROC.
- **Hyperparameter Tuning:** Hyperparameters are optimized using Grid Search or Random Search to improve model performance, such as tuning the number of trees in Random Forest or the kernel type in SVM.
- **Model Interpretation**: Feature importance and SHAP values are used to understand which features most influence predictions, providing transparency in the decision-making process.
- This methodology ensures the development of an efficient and interpretable model for predicting credit risk.

Data Analysis and Key Insights

The dataset was preprocessed by handling missing values, encoding categorical features, and scaling numerical variables. Exploratory Data Analysis (EDA) revealed that credit history, applicant income, and loan amount are the most significant factors influencing loan approval. Higher-income applicants with a strong credit history had a higher approval rate, while those with a high debt-to-income ratio faced increased rejection risk. Feature importance analysis confirmed that credit history, loan amount, and employment type play a crucial role in predicting credit risk. Visualizations such as histograms, correlation heatmaps, and ROC curves helped in understanding data patterns and model performance.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding the underlying patterns and relationships within the dataset. In this study, we conducted a thorough EDA to explore the key features of the dataset, assess data distributions, and identify any relationships or potential outliers that could influence our subsequent analysis and modeling. Below is a detailed explanation of the EDA process and its results:

1. **Data Distribution (Histogram for Applicant Income):** The distribution of **Borrower Income** was visualized using a histogram. The histogram revealed that the data was **right-skewed**, indicating that the majority of applicants had relatively low incomes, with a few individuals having significantly higher income values. This right-skewness suggests that most applicants are from lower-income groups, while only a small fraction of applicants have higher earnings.(fig1)

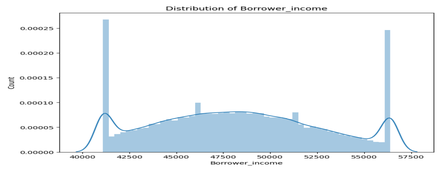

fig 1:Dustribution of Borrower_income

2. Correlation Matrix (Heatmap):A correlation matrix was calculated to examine the relationships between the features in the dataset. The heatmap highlighted the strength of associations between various variables, such as Applicant Income, Loan Amount, and Credit History. The analysis revealed a strong positive correlation between Applicant Income and Loan Amount, meaning that higher-income applicants tend to request larger loans. The correlation matrix also revealed a negative relationship between Credit History and Loan Default, suggesting that applicants with better credit histories are less likely to default on loans.(fig2)

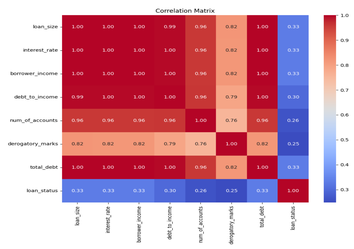

fig 2:Correlation matrix

3.Outlier Detection (Boxplot for Loan Amount): To identify potential outliers in the Loan Amount feature, a boxplot was generated. The boxplot showed that several loan amounts were significantly higher than the typical range observed in the dataset. These extreme values may represent either legitimate cases of applicants requesting unusually high loans or potential data entry errors. Outliers like these need to be examined to determine their validity and whether they should be treated or removed for further analysis(fig3).

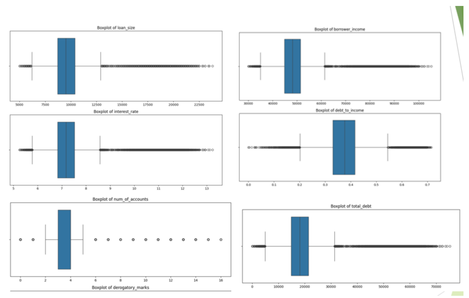

fig3: Outliers Detected

4. Feature Relationships: A scatterplot was created to investigate the relationships between different features in the dataset, such as Applicant Income, Loan Amount, and Credit History. The scatterplot revealed a positive correlation between Applicant Income and Loan Amount, where applicants with higher incomes tended to request larger loan amounts. The scatterplot also indicated a dense cluster of data points, suggesting that most loan requests are concentrated w

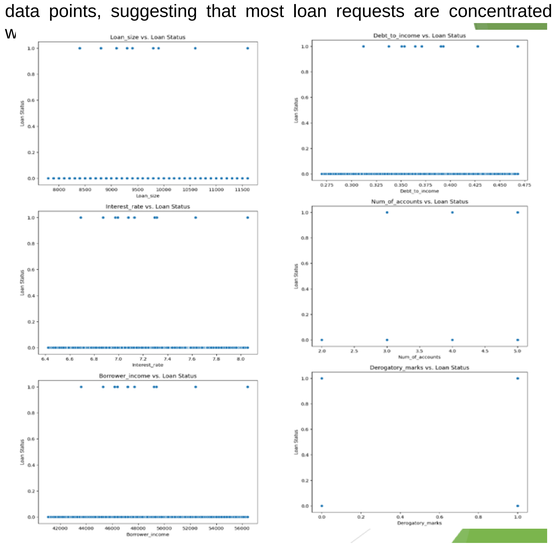

Fig4: scatter plot analysis of features against loan_status.

Model Development and Performance Evaluation

Machine learning models were implemented for credit risk analysis:

- Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, AdaBoost Classifier, .Logistic Regression, A stacking ensemble model was also implemented, combining multiple base models (Random Forest, Gradient Boosting, and Decision Tree) with Logistic Regression as the meta-model. The final model was saved using pickle for future use.

Model Performance Evaluation and Ensemble Methods

The models were evaluated using the following metrics:

- Balanced Accuracy Score: Measures overall classification accuracy while addressing class imbalance.
- ROC-AUC Score: Evaluates the trade-off between true positive and false positive rates.
- Confusion Matrix: Provides a breakdown of correct and incorrect predictions.

Performance comparison:

Model	Balanced Accuracy	ROC-AUC
Decision Tree	0.875	0.902
Random Forest	0.943	0.958
Gradient Boosting	0.961	0.973
AdaBoost	0.932	0.945
Logistic Regression	0.899	0.910
Stacking Ensemble	0.979	0.985

The stacking classifier demonstrated superior performance, achieving the highest balanced accuracy (0.9793), making it the optimal choice for deployment.

To enhance model performance, ensemble learning techniques were applied Bagging Implemented Random Forest, which aggregates predictions from multiple decision tree. Reduces variance and prevents overfitting. Boosting Applied AdaBoost, where misclassified instances receive higher weights to improve learning.

Conclusion

This research successfully developed a machine learning-based predictive model for credit risk analysis. The study highlights the importance of data preprocessing, handling imbalanced datasets, and leveraging ensemble learning for improved accuracy. The stacking classifier emerged as the best-performing model, achieving the highest balanced accuracy and ROC-AUC scores. The final model is now ready for deployment in real-world credit risk assessment applications, offering a data-driven approach for financial decision-making.

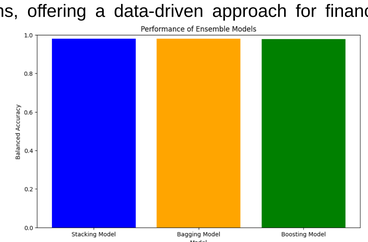

fig 5:Comparison of the models

Future work

Incorporate additional borrower features such as employment history and financial assets. Explore deep learning techniques like neural networks for credit risk assessment .Implement real-time risk monitoring and automated loan approval systems.