Executive Summary

The Credit Risk Prediction project aims to develop a machine learning model to predict the likelihood of a loan applicant defaulting on a loan. By analyzing various features related to the applicant's personal and financial information, the model helps financial institutions assess credit risk more accurately, thereby minimizing potential losses and improving lending decisions.

Introduction

Credit risk assessment is a crucial process for financial institutions, helping them determine the probability of a borrower defaulting on a loan. This project leverages machine learning techniques to predict loan defaults based on historical data, enabling more informed and efficient credit risk management.

Data Description

The dataset used in this project consists of 12 columns, each representing specific attributes of the loan applicants and their loan requests. Below is a detailed description of each column:

person_age: Age of the applicant (integer).

person_income: Annual income of the applicant in dollars (integer).

person_home_ownership: Home ownership status of the applicant (categorical: RENT, OWN, MORTGAGE).

person_emp_length: Length of employment in months (float).

loan_intent: Intended purpose of the loan (categorical: PERSONAL, EDUCATION, MEDICAL).

loan_grade: Grade assigned to the loan based on credit risk (categorical: A to G).

loan_amnt: Loan amount requested in dollars (integer).

loan_int_rate: Interest rate of the loan (float in percentage).

loan_status: Status of the loan (binary target variable: 0 = not in default, 1 = in default).

loan_percent_income: Percentage of the applicant's income that the loan amount represents (float).

cb_person_default_on_file: Indicates if the applicant has a history of default with the credit bureau (categorical: Y = Yes, N = No).

cb_person_cred_hist_length: Length of the applicant's credit history in years (integer).

Exploratory Data Analysis (EDA)

EDA involves summarizing the main characteristics of the dataset and visualizing patterns to gain insights. Key steps include:

Data Overview: Checking for missing values and understanding the distribution of features.

Statistical Summary: Generating summary statistics to describe the central tendency, dispersion, and shape of the dataset's distribution.

Data Visualization: Creating histograms, box plots, and correlation heatmaps to identify relationships and potential outliers.

Key Findings from EDA:

The dataset contains a mix of numerical and categorical features.

Certain features, such as person_income and loan_amnt, exhibit a wide range of values, indicating variability in the financial status of applicants.

The correlation heatmap reveals relationships between numerical features, which can be leveraged during model training.

Feature Engineering

Feature engineering involves transforming raw data into meaningful features that improve the model's predictive power. Key steps include:

Handling Missing Values: Imputing or removing missing values to ensure data integrity.

Encoding Categorical Variables: Converting categorical variables into numerical values using techniques like one-hot encoding.

Scaling Features: Standardizing features to ensure they are on a similar scale, which is crucial for algorithms that rely on distance calculations.

Model Building

Several machine learning models were considered for this project, including:

- Logistic Regression
- Decision Tree
- Naïve Bayes

- SVM
- XGBoost

The data was split into training and testing sets to evaluate the models' performance. Feature scaling was applied to the training data before fitting the models.

Model Evaluation

Models were evaluated using various metrics:

Accuracy: The proportion of correctly predicted instances.

Precision, Recall, F1-Score: To assess the balance between precision and recall.

Confusion Matrix: To visualize true positives, false positives, true negatives, and false negatives.

ROC Curve and AUC: To evaluate the performance across different threshold values.

Key Results

XGboost emerged as the best-performing model with the highest accuracy and AUC.

The confusion matrix and classification report provided detailed insights into the model's performance across different classes.

Conclusion

The Credit Risk Prediction project successfully developed a machine learning model that accurately predicts loan defaults. The Random Forest model, in particular, demonstrated robust performance, making it a valuable tool for credit risk assessment.

Future Work

Future improvements to this project could include:

Hyperparameter Tuning: Further optimizing model parameters to enhance performance.

Feature Selection: Identifying and utilizing the most predictive features.

Incorporating Additional Data: Including more features or external data sources to improve model accuracy.