Prediction of Credit Risk

Executive Summary

This project focused on developing a robust credit risk assessment model to predict loan default probabilities using a comprehensive dataset of borrower attributes. By leveraging machine learning techniques, including Logistic Regression, Random Forest, and Support Vector Machine (SVM), we aimed to identify the most effective model for classifying loan outcomes. The analysis included data preprocessing, feature engineering, model training, evaluation, and visualization to ensure actionable insights for lending decisions. The Logistic Regression model emerged as the top performer, achieving an accuracy of 99.46% and a ROC-AUC score of 0.9971, making it suitable for deployment in credit risk evaluation.

Introduction

Credit risk assessment is a critical process for financial institutions to minimize losses from loan defaults while maximizing lending opportunities. This project utilized a dataset containing borrower financial metrics to build predictive models capable of distinguishing between loans likely to be repaid and those at risk of default. The primary objectives were to preprocess the data, engineer relevant features, evaluate multiple machine learning models, and provide visual insights into model performance and feature importance.

Data Description

The dataset, lending_data.csv, included the following features:

- **loan_size**: The total amount of the loan.
- **interest_rate**: The interest rate applied to the loan.
- **borrower_income**: The annual income of the borrower.
- **debt_to_income**: The ratio of the borrower's debt to their income.
- **num_of_accounts**: The number of credit accounts held by the borrower.
- **derogatory_marks**: The number of negative marks on the borrower's credit report.
- **total_debt**: The total debt owed by the borrower.
- **loan_status**: The target variable (0 for non-default, 1 for default).

A new feature, loan_to_income_ratio, was engineered by dividing loan_size by borrower_income to capture the relative burden of the loan on the borrower's financial capacity.

Methodology

Data Preprocessing

1. **Data Loading**: The dataset was imported using pandas, ensuring proper handling of the CSV file.
2. **Missing Values**: An initial check confirmed no missing values across all features, allowing immediate progression to feature engineering.
3. **Feature Engineering**: The loan_to_income_ratio was added to enhance the model's ability to assess loan affordability.

4. **Data Splitting**: The dataset was split into training (70%) and testing (30%) sets, with stratification to maintain the distribution of loan_status.
5. **Feature Scaling**: Features were standardized using StandardScaler to ensure compatibility with distance-based algorithms like SVM.

Model Development

Three machine learning models were trained and evaluated:

1. **Logistic Regression**: A linear model suitable for binary classification, chosen for its interpretability and efficiency.
2. **Random Forest**: An ensemble method leveraging decision trees, selected for its robustness to overfitting and ability to handle non-linear relationships.
3. **Support Vector Machine (SVM)**: A linear kernel SVM was used for its effectiveness in high-dimensional spaces and binary classification tasks.

Each model was trained on the scaled training data, and predictions were made on the test set. Performance metrics included accuracy, precision, recall, ROC-AUC, and cross-validated ROC-AUC scores.

Evaluation Metrics

The models were assessed using the following metrics:

- **Accuracy**: The proportion of correct predictions.
- **Precision**: The ratio of true positive predictions to total positive predictions, critical for minimizing false positives in loan approvals.
- **Recall**: The ratio of true positives to actual positives, important for identifying default cases.
- **ROC-AUC**: The area under the Receiver Operating Characteristic curve, measuring the model's ability to distinguish between classes.
- **Cross-Val ROC-AUC**: The mean ROC-AUC score from 5-fold cross-validation, ensuring robustness.

Visualization

Visualizations were generated to aid interpretation:

1. **ROC Curves**: Plotted for each model to compare their ability to separate classes.
2. **Correlation Heatmap**: Displayed pairwise correlations between features to identify potential multicollinearity.
3. **Feature Importance**: For the Random Forest model, feature importance scores were visualized to highlight key predictors.

Results

Model Performance

The performance metrics for each model on the test set are summarized below:

Model	Accuracy	Precision	Recall	ROC-AUC	Cross-Val ROC-AUC
Logistic Regression	0.9946	0.8681	0.9827	0.9971	0.9941
Random Forest	0.9924	0.8737	0.8947	0.9970	0.9936
SVM	0.9947	0.8684	0.9853	0.9971	0.9941

- **Logistic Regression** achieved the highest ROC-AUC (0.9971) and accuracy (99.46%), with a strong recall of 0.9827, indicating excellent identification of default cases.
- **SVM** performed similarly, with a slightly higher recall (0.9853) but identical ROC-AUC and marginally better accuracy (99.47%).
- **Random Forest** had the lowest accuracy (99.24%) and recall (0.8947), though it maintained a competitive ROC-AUC (0.9970).

The Logistic Regression model was selected as the best performer due to its balance of high accuracy, ROC-AUC, and interpretability, making it ideal for deployment in a financial context.

Feature Importance

The Random Forest model provided feature importance rankings, revealing the following order of influence on loan status prediction:

1. **interest_rate** (0.2595): Higher interest rates were strongly associated with default risk.
2. **borrower_income** (0.2185): Lower income levels increased the likelihood of default.
3. **debt_to_income** (0.1401): Higher debt-to-income ratios correlated with default risk.
4. **total_debt** (0.1222): Greater total debt burdens were predictive of default.
5. **loan_to_income_ratio** (0.1157): The engineered feature proved valuable, indicating loan affordability's role.
6. **loan_size** (0.1085): Larger loans posed higher risks.
7. **num_of_accounts** (0.0354): The number of accounts had a moderate impact.
8. **derogatory_marks** (0.0001): Negative credit marks had minimal influence, possibly due to low variance.

Visual Insights

- **ROC Curves**: All models exhibited near-perfect ROC curves, with AUC values close to 1, confirming their strong discriminative power. The curves were saved as roc_curves.png.
- **Correlation Heatmap**: The heatmap revealed high correlations between borrower_income, total_debt, and debt_to_income, suggesting potential redundancy.

However, all features were retained due to their distinct contributions to the model. The heatmap was saved as correlation_heatmap.png.

- **Feature Importance Plot**: The Random Forest feature importance was visualized, emphasizing the dominance of interest_rate and borrower_income.

Model Deployment

The best-performing Logistic Regression model and the StandardScaler were saved as best_credit_risk_model.pkl and scaler.pkl, respectively, using joblib. These files enable seamless integration into a production environment for real-time credit risk assessment. The model can be loaded, and new borrower data can be scaled and evaluated to predict loan default probabilities.

Discussion

The Logistic Regression model's superior performance highlights its suitability for credit risk assessment, balancing predictive power with interpretability. Its high recall ensures minimal missed defaults, critical for risk management, while its precision reduces false positives, optimizing loan approvals. The feature importance analysis underscores the importance of financial metrics like interest rates and income, aligning with domain knowledge in lending.

However, the dataset's lack of missing values and relatively clean structure may not reflect real-world scenarios, where data quality issues are common. Future work could involve testing the model on noisy or imbalanced datasets to assess robustness. Additionally, exploring advanced techniques like XGBoost or neural networks could further enhance performance, though at the cost of interpretability.

Conclusion

This project successfully developed a credit risk assessment model using machine learning, with Logistic Regression emerging as the optimal choice. The model achieved exceptional performance metrics, supported by insightful visualizations and feature engineering. The saved model and scaler provide a practical solution for financial institutions to evaluate loan applications efficiently. Future enhancements could focus on handling diverse datasets and incorporating additional features to further refine predictions.

Recommendations

1. **Deploy the Logistic Regression Model**: Integrate the saved model into the lending decision pipeline for real-time risk assessment.
2. **Monitor Model Performance**: Regularly evaluate the model on new data to ensure sustained accuracy and recalibrate if necessary.
3. **Enhance Feature Set**: Incorporate additional features, such as credit scores or employment history, to improve predictive power.
4. **Address Data Limitations**: Test the model on datasets with missing values or class imbalances to validate its robustness.