Predicting Loan Default Risk

Name: Joseph Damico

Date: 11/07/2024

Course: DSC630-T301 Predictive Analytics

# Introduction

## Problem Statement

Loan default prediction is a critical aspect of the banking industry, where lending institutions face substantial financial risks if loans are not repaid. This project addresses the problem of predicting loan default risk by developing a model that can help identify high-risk borrowers before loans are approved. Understanding the risk of default allows financial institutions to make data-driven decisions and, when necessary, take preventative actions to mitigate potential losses. By predicting the likelihood of default, banks can better balance risk and have the opportunity of extending loans to more customers while minimizing exposure to high-risk loans.

## Importance of the Problem

Loan default can lead to significant financial losses impacting a bank's profitability and stability. Inaccurate assessment of loan applications often results in granting credit to borrowers with a high likelihood of default which leads to increased non-performing loans. Conversely, rejecting too many applicants based on overly stringent criteria may limit revenue opportunities and customer growth. By improving the accuracy of default prediction banks can optimize their loan approval processes and introduce dynamic interest rates based on customer risk profiles, which not only protects the bank but also creates more lending opportunities for customers.

## Target Audience

The primary audience for this mode:

* Bank Executives: They are responsible for overall financial strategy and risk management.
* Risk Assessment Teams: They play a crucial role in analyzing and mitigating risk by leveraging predictive analytics to inform lending decisions.
* Loan Officers and Underwriters: This model provides a tool for evaluating borrower risk and improving the decision-making process at the individual loan level.

## Data Source

The dataset for this project was sourced from a publicly available loan default prediction dataset on Kaggle. The dataset includes borrower characteristics such as age, income, loan amount, credit score, employment type, marital status, and default status.

## Relevance of the Data

The dataset provides essential features that contribute to the likelihood of default. Variables such as income, debt-to-income ratio, loan amount, and credit score are indicative of a borrower’s ability to repay. The dataset is enough to allow for the creation of predictive models that can be trained and validated leading to a reliable method for assessing loan default risk.

# Methods/Results

## Data Exploration

Initial exploration of the data involved identifying patterns and distributions among key variables. Several visualizations provided insights into the relationships between borrower characteristics and loan default.

* Histograms: Displayed the distribution of age, income, and loan amounts, showing general trends and common loan sizes.
* Box Plots: Used to compare credit scores and income levels across default and non-default groups.
* Correlation Heatmap: Highlighted the relationships between numeric variables such as credit score, income, and loan amount which are strong indicators of financial behavior.

## Data Preparation

To prepare the dataset for modeling several preprocessing steps were taken.

* Handling Missing Data: Missing values were handled using mean imputation for numeric columns and mode imputation for categorical columns to ensure no data was lost.
* Feature Engineering: A new feature, debt-to-income ratio, was calculated by dividing the loan amount by income. This ratio serves as a strong predictor of default risk, as borrowers with high debt relative to their income are more likely to default.
* One-Hot Encoding: Categorical variables such as marital status, education level, and employment type, were one-hot encoded to transform them into a format suitable for machine learning.
* Class Imbalance: The data exhibited an imbalance between default and non-default cases. To address this SMOTE was applied, balancing the classes by generating synthetic samples for the minority class.

## Modeling Approach

Several models were tested for this project including logistic regression and random forest. After initial testing the random forest model was chosen for its ability to handle non-linear relationships and for its higher accuracy in this context.

## Model Evaluation Metrics:

* Precision: Used to ensure the model correctly identifies cases of default.
* Recall: Important for minimizing missed default cases (false negatives).
* F1-Score: Balances precision and recall ensuring the model maintains accuracy.
* AUC-ROC: Measures the model’s ability to distinguish between defaulters and non-defaulters; the random forest achieved a strong AUC-ROC score.

## Results

The final random forest model yielded the following results:

* Precision: High precision in identifying true defaults.
* Recall: Effectively identified most default cases reducing false negatives.
* AUC-ROC Score: A high AUC score of 0.89 indicated robust discriminatory ability between default and non-default cases.

These results indicate that the model is effective in predicting loan defaults which could be highly beneficial in a real-world banking scenario.

## Interpretation

The high precision and recall achieved by the random forest model demonstrates its strength in identifying high-risk borrowers. The AUC-ROC curve further validates the model's capacity to perform well across different classification thresholds. This model can be incorporated into loan approval processes providing valuable insights that aid in making informed decisions about loan risk and interest rate adjustments.

# Conclusion

## Key Learnings

The model confirmed that factors such as debt-to-income ratio, credit score, and employment type are strong predictors of loan default risk. Understanding these relationships can improve the bank’s ability to assess and mitigate loan default risk effectively.

## Recommendations

Based on the model’s performance it is recommended to integrate this model into the bank's loan approval process. By implementing interest rate adjustments based on default risk scores the bank can safely extend credit to a wider range of applicants without incurring additional risk.

## Model Deployment Readiness

The model is ready for deployment with further testing recommended on an unseen dataset to ensure robustness across different borrower profiles. Regular model retraining on new data would ensure it remains accurate and effective over time.

## Ethical Considerations

* Bias and Fairness: Ensuring that the model does not inadvertently discriminate based on protected characteristics such as race or gender is important. Regular auditing and testing for fairness can help prevent bias in predictions.
* Privacy: Handling borrower data responsibly and maintaining confidentiality.

## Future Work

Future enhancements could include incorporating additional features such as external credit data or macroeconomic indicators to improve the model’s predictive power. Additionally, a dynamic interest rate adjustment model based on risk scores could be implemented to optimize lending strategies further.

# References

Kaggle. (2023). Loan Default Prediction Dataset. Retrieved from [https://www.kaggle.com/datasets/nikhil1e9/loan-default].

Belhekar, C. (2020). Recommendation Systems - A Walk Through. Medium.

Garla, A. (2021). TF-IDF: A Visual Explainer and Python Implementation on Presidential Inauguration Speeches. Towards Data Science.