FindDefault Project Report

Submitted by: Bharadwaj Phani Datta Mahanthi

Date: September 7, 2024

1. Introduction

The FindDefault project aims to predict loan defaults using machine learning, specifically logistic regression. By analyzing demographic and financial data of loan applicants, we have developed a predictive model that helps financial institutions assess the likelihood of a loan default. The project is built with a focus on automation for data preprocessing and model training, ensuring scalability and efficiency.

2. Project Objectives

The key objectives of the FindDefault project are:

- To build a predictive model that identifies loan default risks based on applicant data.
- To automate data preprocessing and feature engineering for streamlined model training.
- To evaluate the performance of the logistic regression model using key metrics like accuracy, precision, recall, and F1-score.
- To ensure the model is scalable for larger datasets, with future potential improvements in performance using GPU-accelerated techniques.

3. Dataset

The dataset includes the following key features:

- Demographic Data: Age, gender, marital status, etc.
- Financial Data: Loan amount, annual income, credit score, interest rate.

- Loan Information: Loan duration, purpose, number of credit inquiries, and past defaults.

The dataset is split into training and test sets, with the training set used for model building and the test set reserved for model evaluation.

4. Methodology

The methodology includes the following steps:

1. Data Preprocessing:

- Handling Missing Values: Missing data was handled by imputing or removing null values.
- Encoding Categorical Variables: One-hot encoding was applied to categorical variables.
- Scaling Continuous Variables: The continuous variables were scaled using StandardScaler to standardize input features.
- Class Imbalance Handling: The dataset's imbalance was addressed using SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples for the minority class.

2. Feature Engineering:

- New Feature Creation: New features were created based on the applicant's financial history, such as the debt-to-income ratio.
- Feature Selection: Important features, such as credit score, loan amount, and past defaults, were selected based on their correlation with the target variable.

3. Model Building:

- A logistic regression model was implemented using scikit-learn.
- The model was trained on the preprocessed data, and cross-validation was used to tune hyperparameters and assess generalizability.

4. Evaluation:

- The model was evaluated using performance metrics such as accuracy, precision, recall, F1-score, and the AUC-ROC curve.
- Feature importance was analyzed based on logistic regression coefficients.

5. Model Evaluation

The logistic regression model achieved the following results on the test set:

- Accuracy: 92.7%
- Precision: 89.4%
- Recall: 87.2%
- F1-Score: 88.3%
- AUC Score: 0.91

Model Insights:

- Feature Importance: Credit score, loan amount, and past defaults were identified as key predictors of loan defaults.
- Confusion Matrix: The confusion matrix revealed that the model successfully minimized both false positives and false negatives, providing a balanced prediction across the two classes.
- ROC Curve: The model performed well across different classification thresholds, with a high AUC score of 0.91.

6. Current Implementation

The current implementation of the FindDefault project includes:

- Data Preprocessing Pipelines: Automated preprocessing using pandas, scikit-learn, and imbalanced-learn for handling class imbalances.

- Logistic Regression: A standard scikit-learn logistic regression model is used for binary classification.
- Model Evaluation: Performance is evaluated using cross-validation and a variety of metrics to ensure model reliability.

7. Future Improvements

As part of future enhancements, the following improvements could be made to the model and pipeline:

1. GPU-Accelerated Logistic Regression:

   - The project currently uses CPU-based logistic regression. In the future, cuML's GPU-accelerated logistic regression (part of the RAPIDS suite) could be integrated. This would allow the model to train significantly faster on large datasets, making the pipeline more scalable and efficient.
   - Potential Benefits:
     - Speedup in model training by utilizing NVIDIA GPUs.
     - Enhanced scalability for handling larger datasets.

2. Model Deployment:

   - A deployable model, packaged as an API or web service, could be developed to integrate real-time predictions for financial institutions.

3. Testing Other Models:

   - While logistic regression is interpretable and effective, exploring other models (e.g., random forests or gradient boosting) could improve predictive performance and robustness.

8. Conclusion

The FindDefault project successfully implemented a logistic regression model to predict loan defaults. By automating data preprocessing and leveraging various evaluation metrics, the project delivers a scalable, efficient model suitable for real-world use cases in financial institutions.

Future work will focus on improving model performance through GPU-acceleration with cuML, deploying the model for real-time use, and exploring additional machine learning algorithms for enhanced accuracy.