Loan Fraud Prediction

Yuxuan Zhang

Westcliff University

AIT 506

Instructor: Professor Noha Bhairy

Date: Nov 8, 2025

Load Fraud Prediction

This project builds a classification model for predicting, given customer attributes such as income, education level and marital status as well as the load amount, whether the customer will default on the load. The dataset used is the Loan Default Prediction Dataset on kaggle. In the Jupyter notebook, I perform data analysis steps such as data cleaning, preprocessing and exploratory analysis. I also explore different machine learning methods, perform hyperparameter tuning and evaluate the results on a held-out test set. In this report, I will provide more details on each of the steps.

The dataset has 255347 entries and 18 columns. We drop the column for load id since that is irrelevant to the task. Of the remaining columns, 8 are categorical variables and the remaining are numeric variables. No missing values are present in the dataset, so we keep all records. In preprocessing, I use one-hot encoding for all categorical variables, dropping the first class to avoid correlation between the generated columns. For the numerical features, I perform a Z score normalization so that all features are on the same scale (de Amorim, 2023). This is important to models that are sensitive to feature scales such as regularized logistic regression and linear regression. In exploratory analysis, some noteworthy findings are

1. Customers who default have lower income on average.
2. Loans that are defaulted have a higher amount.
3. Somewhat unexpectedly, education and dependents don’t seem to be correlated with defaulting.

The dataset is highly imbalanced. The ratio of default is 0.11. This will affect modelling and evaluation decisions (Sun et al., 2009), which will be discussed below. A train-test split of 0.8/0.2 is performed.

The models that are explored are linear regression (the output is viewed as a probability) logistic regression, SVM and decision trees. For each model, I train a baseline using the original training dataset. To address the issue of imbalance in the data, I used SMOTE (Chawla, 2002) to create a new resampled dataset that oversamples the minority class. I then train a version of the model with the resampled dataset. Cross validation – with SMOTE applied within each fold – is applied to find the best parameters with respect to the F1 score. For evaluation, I use accuracy, precision, recall and the F1 score. The best F1 score is achieved with logistic regression trained on the SMOTE resampled dataset with a probability threshold of 0.65. But the results from other models are fairly similar. The decision tree is plotted for interpretation, the first levels of which show decisions based on the variables age, interest rate and number of credit lines, which are seen to be correlated with the target in exploratory analysis. The detailed results are contained in the Jupyter notebook.

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

de Amorim, L. B. V., Cavalcanti, G. D. C., & Cruz, R. M. O. (2023). The choice of scaling technique matters for classification performance. Applied Soft Computing, 133, 109924. https://doi.org/10.1016/j.asoc.2022.109924

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4), 687-719.