Responsible Lenders Data Science Bootcamp - Autumn 2023
Team: Craig Franze, Patrick Millican, Chao Sun, Alex Melendez, Andre Martins
Github: https://github.com/Responsible-Lenders/credit-default

Overview: In our commitment to advancing responsible financial lending practices, our team has embarked on a project with a clear objective: to identify key factors in credit evaluation and to enhance the accuracy of creditworthiness predictions. This initiative is not just about data analysis; it's about making a significant impact in the realm of financial decision-making.

Our analysis reveals that variables like Sex, Marriage, Education, and Age have minimal impact on default probability, challenging traditional beliefs and pointing to more significant factors. We found potential biases due to underrepresentation in some categories and a class imbalance in the default rate, necessitating data stratification. The most crucial predictor of default is the recent payment history, especially the latest month's payment status (Pay_1). Additionally, the amount of credit given, along with Education and Age, shows notable correlations with default. These insights are essential for creating accurate predictive models and strategies to reduce credit risk.

XGBoost, a tree-based machine learning algorithm, was selected for its efficiency and ability to handle categorical variables without scaling data. Its hyperparameters effectively address class imbalance. However, it has drawbacks, including its complexity which hampers explainability and a tendency to overfit, along with sensitivity to outliers. Performance metrics vary based on focus: for recall-driven models, it achieved 66% recall and 45% precision with 74.65% accuracy. When precision-driven, it showed 38% recall, 68% precision, and 82.23% accuracy. Lastly, focusing on F1-score, it attained 55% recall, 52% precision, and 78.97% accuracy.

Logistic regression, valued for its simplicity and interpretability, has its own set of pros and cons. While its straightforward nature facilitates understanding the output, this simplicity may lead to underperformance compared to more complex algorithms. In our comparison of two logistic regression algorithms, "L-BFGS" and "liblinear," based on decision cutoffs, "liblinear" emerged as the superior performer. Notably, its F1 Score and Accuracy were optimized when cutoff values ranged between 0.25 and 0.4. This resulted in a 3% increase in Accuracy and a significant 50% boost in F1 Score compared to a baseline model that predicts non-default in every instance.

This analysis confirms the intuitive assumption that recent payment history is the most predictive of default likelihood. In choosing a model, XGBoost strikes a better balance between recall and precision compared to logistic regression, though both models show potential. Variables like education and marital status emerged as less critical. Rebalancing data, borrower age, and loan size were also of lesser importance. A key trade-off identified is that precision and recall can be maximized independently, but not simultaneously. The need for additional borrower data is apparent to strengthen the model, suggesting that broader datasets could yield more robust predictions.