EXECUTIVE SUMMARY - PROJECT ROSETTA
ERDŐS INSTITUTE - SPRING 2022

HANNAH ALPERT, AMIN IDELHAJ, AND SOUMYA SANKAR

Problem Description. In this project, we study prediction methods for credit default risk, based on past credit history, and minimal demographic data. The primary stakeholders who might have use for such an analysis are banks and credit card companies, who look to maximize revenue and for default-protection. Based on such data obtained from existing clients, a bank could preemptively offer forbearance or some other intervention, to try to recover the payment, or to make decisions about extending credit limits.

Data source and description. We analysed the "Default of credit card clients" data set from the UC Irvine Machine Learning Repository. This data set is associated with the 2009 paper "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," by I.C. Yeh and C.H. Lien in the journal *Expert Systems with Applications*. Each row of the data spreadsheet corresponds to a credit card client of a particular Taiwanese bank, who either did or did not default on the payment due in October 2005. Approximately 22% of the recorded clients did default in this month. The spreadsheet includes payment history (bill and payment amounts and lateness record) for the six months preceding October 2005, plus some demographic information.

Questions and Key Performance Indicators. Our goal is to understand which of the studied models is the better predictor of default, given the data. In order to do so, we use the area ratio of the cumulative gains curve, which is related to the area under the ROC curve by (Area ratio) = 2×(AUC-ROC) -1. We choose area ratio as our primary KPI since it is independent of the probability threshold chosen to predict default. We also support our conclusions by analysing some secondary KPIs, which can be found in notebooks 3 and 5.

Models. We compare the following algorithms: K-Nearest Neighbors, Linear and Quadratic discriminant analysis, Naive Bayes, Decision Tree, Random Forest, Extra Trees, AdaBoost and Gradient Boost. For each of these algorithms, we perform a cross-validation to find the optimal parameters for highest average area ratio. We then compare these models to each other and to the baseline model. A standard baseline is completely random ranking, which gives an area ratio of 0. We also use a second baseline, ranking clients solely based on the lateness of their most recent payment, which gives a ratio of about 0.42.

Feature Engineering. We performed feature tuning based exploratory data analysis as well as the random forest tools on scikit-learn. By testing various subsets of columns, we were able to make better predictions by taking ratios and differences of certain columns in the data. While payment lateness and bill amounts were the most important predictors, two derived variables in particular seemed to play a critical role: Charge amounts (new charges with each bill) and Bill fractions (ratio of the bill amount to the credit limit).

Conclusions. The best performance came from random forest, gradient boosting and adaptive boosting—none of which were tested by the 2009 paper—and had better area ratio scores than the best performing model in the paper. On the test set, a random forest classifier with 100 estimators, a depth of 7 and maximum sample size of 7000, gave an area ratio of 0.556, higher than the higher validation area ratio in the paper (0.54). A gradient boost classifier with 96 estimators, a maximum depth of 2 and, and a learning rate of 0.4 gave an area ratio of 0.553. The worst performing models were discriminant analysis and naive Bayes methods. Of intermediate performance were the kNN, decision tree, and extra trees methods.

Recommendations. We recommend the tree-based ensemble learning models such as random forest, adaboost and gradient boost for better area ratios. These methods are also simpler to run and interpret than neural networks. We also recommend feature engineering for better performance, as well as more comprehensive data collection. Given that the area ratios of even the best performing models in our analysis were bounded above by 0.6—which is in the acceptable range, albeit not in the excellent range—it is quite likely that the collection of data about more variables, such as income, dependents, etc. would improve the predictive power of these models.

We would like to thank the Erdős institute for organizing the Data Science Bootcamp in Spring 2022. We would also like to thank Kung-Ching Lin for his help.