Analysis of Credit Default Likelihood

Group 2: Carolyn Nina Fiore and Syed Hani Haider

MOTIVATION

Financial institutions no longer operate on hard goods, but on credit; money is primarily abstract, rather than physical cash. Banks invest deposited money and offer clients credit in the form of loans or credit cards, and need some measure of assessing whether clients will be able to repay their debts. We aim to understand the measurable factors of a user's financial profile well enough to predict whether a user will default on their credit card. Approximately 83% of adults in America have at least one credit card, with the average adult owning 3.8 of the 1.1 billion credit cards in the U.S.[i] Almost half of households can't pay off their balances each month, the average debt $5,270.[ii] Over 14 million households in the U.S. owe $10,000 or more in credit card debt.[iii] The importance of this is connected to the 2008 recession. A surfeit of cheap credit lines and loans caused a housing bubble; when that burst, financial institutions were left without sufficient capital to cover the debts. High housing prices and interest rates due to current inflation have economists worried that another recession may be on the horizon, which makes it even more critical for financial institutions to ensure that no repetition of the bad credit and loans occurs that could initiate another recession.

Analysis of existing data on credit card defaults can provide insight into the financial reliability of a potential client. Someone who defaults on a credit card or bank loan is a higher-risk client. We have accessed an open-source dataset from American Express[iv] which they posted as a Kaggle competition for data scientists to create models predicting which users will default. It has 189 potential predictor variables, 177 numerical and 11 categorical. These variables are split into 5 categories: delinquency, spending, payment, balance, and risk. The dataset includes information from over 900,000 clients. Using data science, we can determine user qualities that indicate likelihood of default, which can be used by financial institutions assessing potential clients for credit cards, loans, and mortgages.

Due to the high prize amount for the Kaggle competition, the top contending teams' work was private. The public teams' approaches are primarily random forest classifier, one team used this technique with ROC; generally these teams reached accuracy between 65-75%. A mitigating factor is the size of the dataset, which can limit the teams' approach based on available time; this will be a consideration for our team as well. We do not seek to replicate this competing work but to branch from it based on our knowledge from our machine learning classes. From our research, there is currently no completed work – published articles, tutorials, or reference book chapters - using this dataset.

METHODOLOGY

We have broken down the project goals into milestones for low, medium and high risk steps. Our immediate goal, which is low-risk, is to create a supervised model for this dataset using stepwise selection to determine which of the 189 variables are the best predictors for users who will default. This should enable us to reduce the dimensions without loss of predictive accuracy. It also enables early feedback to any stakeholders on immediate focus areas. Estimated completion for this phase is 31 January 2023.

The medium risk goal is to program a classification model using the variables selected from the first phase. We want to compare the results of XGBoost and Random Forest in terms of accuracy, specificity, and sensitivity based on the formulas in Figure 1, where TP and TN stand for true positive and true negative, FP and FN stand for false positive and false negative, and the lowercase *n* is the total number of measurements. Estimated completion for this phase is 10 February, 2023.

Our high risk goal is to apply the results of Almustfa, Khatir, and Bee's 2022 paper[v] to our data. They concluded that random oversampling improved the accuracy of tree-based models when they were classifying credit score data. We believe that this will hold true for our data as well. Assuming satisfactory completion of phase 2 with positive results, we will initiate this attempt to improve our models. Estimated completion for this phase is 15 February, 2023.

IMPACT

Our approach provides two main benefits. First, the use of free, existing data minimizes company expenses and enables us to determine the usefulness of various measures. If follow-on research is authorized, data collection can be targeted categorically based on our team's results, which will save both financial and work-hour resources. Additionally, the results of phase 1's low-risk supervised model provide actionable information for any teams working in parallel, and can be continuously refined as the project progresses. Using the results of the best-performing model enables American Express, and, more generally, financial institutions, to prevent the compromise of their financial holdings due to poor credit and loans, which is beneficial not only for their stockholders but for the economy at large.

FIGURES

$$\text{Acc.} = \frac{TP + TN}{n}, \quad \text{Sens.} = \frac{TP}{TP + FN}, \quad \text{Spec.} = \frac{TN}{TN + FP}.$$

Figure 1.[vi]

[i] Flynn, Jack. "30+ Credit Card Statistics [2023]: Credit Card Debt, Fraud, Usage, And Ownership Facts" [20 December 2022]. *Zippia*. [Online] available: https://www.zippia.com/advice/credit-card-statistics/.

[ii] Hopkins, Christopher. "Credit Card Balances Surge – and are likely to get worse" [1 December 2022]. *Chattanooga Times Free Press.* [Online] available: https://www.timesfreepress.com/news/2022/dec/01/credit-card-balances-surge-tfp/.

[iii] Ibid.

[iv] American Express – Default Prediction. *Kaggle.* [Online] available: https://www.kaggle.com/competitions/amex-default-prediction/.

[v] Ibid.

[vi] Almustfa, A., Khatir, H. and M. Bee [24 August 2022]. "Machine Learning Models and Data-Balancing Techniques for Credit Scoring: What is the Best Combination?" Risks, vol. 10: pp. 169. [Online] available: https://doi.org/10.3390/risks10090169/.