Capstone 2 Project Proposal: Risk Assessment for Small Business Loans

Ray Karpman

The goal of this project is to build a model which classifies SBA-guaranteed small business loans as high or low risk. High risk loans have a significant chance of going into default, while low risk loans do not. The potential client is a bank which lends money to small businesses. Equipped with a data-driven risk assessment model, the client could maximize profit by focusing on low-risk loans and avoiding financial losses from defaults.

The data for this project is from the United States Small Business Administration (SBA), and contains information on over 899,000 small business loans dating from 1987 through 2014. The data was retrieved and partially cleaned by professors Min Li, Amy Mickel and Stanley Taylor, for use as a case study in undergraduate and graduate level statistics courses (Li et al.)

Founded in 1953, the SBA's mission is to promote small business, by guaranteeing a portion of certain small business loans. If the borrower defaults on an SBA-guaranteed loan, the SBA will pay a portion of the balance. Small businesses have the potential to create jobs and help foster strong local economies, and these benefits may outweigh the economic impact of loan defaults. However, financial institutions face serious risks when approving a small-business loan, even if guaranteed by the SBA. (Li et al.) Hence there is a need for better tools to quantify and manage the risk associated with SBA-guaranteed small-business loans.

To build our model, we will use the data set compiled by Li, Mickel and Taylor, which has been uploaded to Kaggle by Mirbek Toktogareav (Toktogareav). This rich dataset contains twenty-seven variables, which describe the company applying for the SBA loan, and the specifics on the loan application itself. Some features that may be of interest include what industry the business belongs to, whether the business is new or pre-existing, and how many jobs were created or retained as a result of the loan. Also of interest are geographic predictors, namely U.S. state and urban vs. rural. I plan to test several different models on the data, to find the one which performs best. Logistic regression and random forest algorithms may be worth exploring here.

The deliverables for this project are a Github repository containing reproducible code for obtaining, cleaning, exploring, and preprocessing the data, as well as training the final model. This repository will also include a slide deck and project report, which will explain our main results concisely, and offer recommendations to a potential client. Following Springboard's suggested timeline, the project will be completed by November 15, 2021.

There are several possible challenges to building our predictive model. We are using a large historical dataset, which provides a wealth of detailed information about hundreds of thousands of loans. However, historical data may not always be directly relevant to predicting current risk levels. For example, loans active during a recession may have a higher risk of default. Hence a machine-learning model might penalize other features common to recession-era loans, whether or not those features contribute to risk. We could try addressing this problem by weighting data from different years differently, or by adding an additional variable that was an indicator of prevailing economic conditions. Similarly, a large number of defaults might occur in a given state following a natural disaster. It will be important to explore the data carefully, and note any anomalous spikes in default rates.

References

Li, Min, Amy Mickel and Stanley Taylor. "'Should This Loan be Approved or Denied?': A Large Dataset with Class Assignment Guidelines," *Journal of Statistics Education*, vol. 26, no. 1, 2018, pp. 55-66. doi:10.1080/10691898.2018.1434342

Toktogaraev, Mirbek. *Kaggle.* "Should this loan be approved or denied?", March 17 2020. www.kaggle.com/mirbektoktogaraev/should-this-loan-be-approved-or-denied. Accessed September 6, 2021.