Evaluating the Performance of Classification Algorithms in Predicting Credit Card Default

CIND 820: Big Data Analytics Project

Abstract

Mitchell Cassar

501140751

Dr. Tamer Abdou

January 23, 2023

The ability of financial institutions to evaluate and forecast credit risk is an integral component to the stability of financial markets and national economies (Aaron et al., 2012). So important is this task that the global financial recession of 2008 is largely attributed to firms' failure to properly manage counterparty risk (OCC, 2021). As a result, there is extreme financial and regulatory pressure on institutions to develop and maintain effective solutions to manage risk.

Regulatory changes approved in 1998 revolutionized this area, allowing banks to develop internal mathematical models for the purposes of evaluating credit risk (Gordy et al., 2000). Although institutions are often required to report on potential credit risks of all borrowers and products, this project will focus on individual borrowers and their credit cards accounts. It will evaluate the application of common machine learning models on these borrowers to model their repayment behaviour. The models applied are sourced from the machine learning subset of classification and regression.

These models are to be trained, tested, and evaluated on their performance using a publicly available dataset of 30,000 Taiwanese credit card borrowers (Yeh, I. C. et al., 2009). Please find the referenced dataset here: http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients Each line item represents an individual borrower and contains both quantitative and qualitative attributes. The qualitative information provided for each borrower includes gender, age, education, and marital status. The quantitative attributes are as follows:

- Amount of credit given
- History of past payment (as multiple attributes over several billing cycles)

- Amount of credit card bill statement (as multiple attributes over several billing cycles)
- Amount of previous credit card payment (as multiple attributes over several billing cycles)

In total, each borrower contains 24 attributes. This data will be preprocessed. Statistical description of the data will be provided, including summary statistics and visualizations. These processing operations will include identifying missing data and outliers as well as basic attribute analysis such as correlation, covariance and if necessary, dimensionality reduction using principal component analysis.

The dataset described above will be analyzed to answer the following research question: *To what level can demographic data and recent payment history accurately predict the default of credit consumers?* In the process of answering this high-level question, the following narrower questions will also be considered:

- How different are the results produced by isolated machine learning algorithms and which algorithm is the most accurate?
- Which performance evaluation metrics are most relevant to predicting customer defaults?
- Does a time-series or consolidated view of customers repayment history produce more accurate results?

To answer these questions, models will be developed for standard classification and regression algorithms including K-Nearest Neighbours, Logistic Regression, Decision Tree and Random Forest. For each supervised learning method, the dataset will be split into training and test sets. The models will be trained on the training data and the test data will be used to predict the class variable. The models predictions will then be compared to the class variable of the live test data using evaluation metrics such as a Confusion Matrix, Accuracy, Precision and Recall, and F-Score.

References

Aaron, M., Armstrong, J., & Mark Zelmer. (2012). *An overview of risk management at Canadian Banks - Bank of Canada.* Bank of Canada. Retrieved January 19, 2023, from https://www.bankofcanada.ca/wp-content/uploads/2012/01/fsr-0607-aaron.pdf

Comptroller's Handbook: Allowances for Credit Losses. Office of the Comptroller of the Currency. (2021, April 15). Retrieved January 18, 2023, from https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/allowances-for-credit-losses/index-allowance-credit-losses.html

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Gordy, M., Galai, D., Mark, R (2000). *A comparative analysis of current credit risk models.* Journal of Banking and Finance. 24(1-2), 2. Retrieved January 20, 2023 from http://www.financerisks.com/filedati/WP/Credit%20risk/COMPARATIVE%20CREDIT%20RISK%20MODELS.pdf

Yeh, I. C., & Lien, C. H. (2009). *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients*. Expert Systems with Applications, 36(2), 2473-2480.