Machine Learning and Statistical Methods in Credit Risk Management: System for Developing a Credit Risk Scorecard

Jorge Reyes

University of Tennessee, Knoxville

COSC 522 Machine Learning

December 3, 2024

Contents

1 Introduction

Abstract

The goal of this project is to showcase machine learning and statistical methods for developing a loan approval scorecard for proper credit risk management. While we do explain the motivation and background of the topics covered in this paper, this paper will not be covering the mathematical theory behind each topic. The first part of any data science project consist of collecting data and then conducting pre-processing operations. This project is no different. In fact, more resources were spent understanding and cleaning the dataset than any other step in the project. This dataset, although the exact source is unknown, comes from a kaggle repository and is rated as 'gold' by kaggle admins due to its usability. The dataset is split into two tables, one contains fields outlining the life of the loan and the other houses all customer characteristics. We will be referring to the aforementioned tables as 'credit' and 'app' respectively throughout this paper. One last note on the dataset, there is not a clear definition if/when a default occurred. Knowing this, as well as acknowledging the extremely small target class if increasing the number of days outstanding, we chose to identify 'bad' borrowers as those having an overdue balance greater than 60 days. This gives leniency to the borrower and allows us to include more records for our target class than having, for example, a cutoff overdue balance greater than 90 days. Next, we will be creating two classification models, a logistic regression and random forest classifier, that will predict whether customers are 'good' or 'bad' borrowers. Before creating our time-to-default (in this case, time-to-bad?) model, we will be running a special type of EDA called vintage analysis. Our goal for the vintage analysis is to show the cumulative distribution of 'bad' borrowers in our loan portfolio (ie. dataset) by their origination date. For our time-to-default models, we will be aiming at predicting the probability that a borrower, given their characteristics, will turn 'bad' as the loan approaches 5 years on the books. As with any ML or statistical model, we will be performing various tests and performance measures and scores. To conclude, an a example of implementing a loan approval scorecard, as a result of our analysis, will be shown.

2 Setting the Stage

2.1 Introduction to the Data

Summarize relevant previous work. Highlight the gaps in the literature that your research addresses. The introduction provides the background and context for the research. Discuss the motivation, objectives, and scope of the study. This is a reference to a study [1].

3 Dealing with Class Imbalance

Describe the methods used in your research. Include details about data collection, experimental design, tools, or techniques employed.

4 Classification Models

Present the findings of your research. Use tables, figures, and charts to illustrate your results if necessary.

5 Time-To-Default: Survival Analysis

Interpret your results and discuss their implications. Compare your findings with those of previous studies.

6 Applying Results

Summarize the key points of your research. State the main conclusions and suggest potential areas for future research.

7 Conclusion

Summarize the key points of your research. State the main conclusions and suggest potential areas for future research.

References

[1] Alex Example. *Guide to Effective Research*. `https://example.com/research-guide`. Accessed: 2025-01-15. 2023.