Credit Risk Analysis

Hung Nguyen (1022029) - Hiep Nguyen (1022799)

November, 2022

Contents

1 Introduction

This project is part of the Aalto University Bayesian Data Analysis 2022 Course.

One of the most important services of banks that attract their customers is providing credit. When lenders offer home loans, auto loans, or business loans, there is an inherent risk that borrowers will default on their payments, this is termed as Credit Risk. Credit risk is universally known as the possibility of a loss for a lender due to a borrower's failure to repay a loan. When the credit risk is mishandled by the lenders, the consequences can be catastrophic. The collapse of the housing market in 2008 and the ensuing recession were one of the best illustration of how severe the outcome can be: in just a few months, the banking sector nearly collapsed due to a significant overexposure to credit defaults. Therefore, it is essential for the banks to determine the lenders' ability to meet debt obligations, and this process is known as Credit Risk Analysis

1.1 Motivation and problem definition

For our project, we choose our topic to be Credit Risk Analysis due to its meaningfulness, significance, and the fact that there has not been any previous Bayesian data analysis performed on the data set. Our goal is to use Bayesian data analysis method to identify the riskiness of a loan and classify them into good loan and bad loan.

The report consists of the following parts: introduction, data description, models, results, discussion, conclusion, and appendix. First, we formulate the problem and show how we handle the data through pre-processing and feature selection process. Then in the Models section, we describe the two Stan models used and justify their likelihood and justification of their choice. In the Result section, we perform convergence analysis, posterior predictive checks, predictive performance assessment, model selection, and prior sensitivity analysis. In the Discussion and Conclusion section, we will discuss issues and potential improvements for our models, as well as some interesting insights that we have learned while doing the project. The Appendix part will include all the Stan model code. The complete model and R code can also be found at https://github.com/Hungreeee/Credit-Risk-Analysis

1.2 Research goals

Our goal is to use Bayesian data analysis method to identify the riskiness of a loan and classify them into good loan and bad loan using Stan and R programming language. Non-hierarchical and hierarchical models will be applied to this problem, and through convergence analysis, posterior predictive checks, predictive performance assessment, and prior sensitivity analysis, we can choose the better methods for this problem and also know how we can further improve our Bayesian-approach analysis for this problem.

2 Data description

This data set is used in a Kaggle competition - Predicting Credit Risk. The original data set contains 1000 data points with 20 categorical/symbolic attributes prepared by Prof. Hofmann. In this data set, each entry represents a person who takes a credit by a bank, and each person is classified as good or bad credit risks according to the set of attributes. The data set can be found at: https://www.kaggle.com/datasets/kabure/german-credit-data-with-risk