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 preprocessing 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

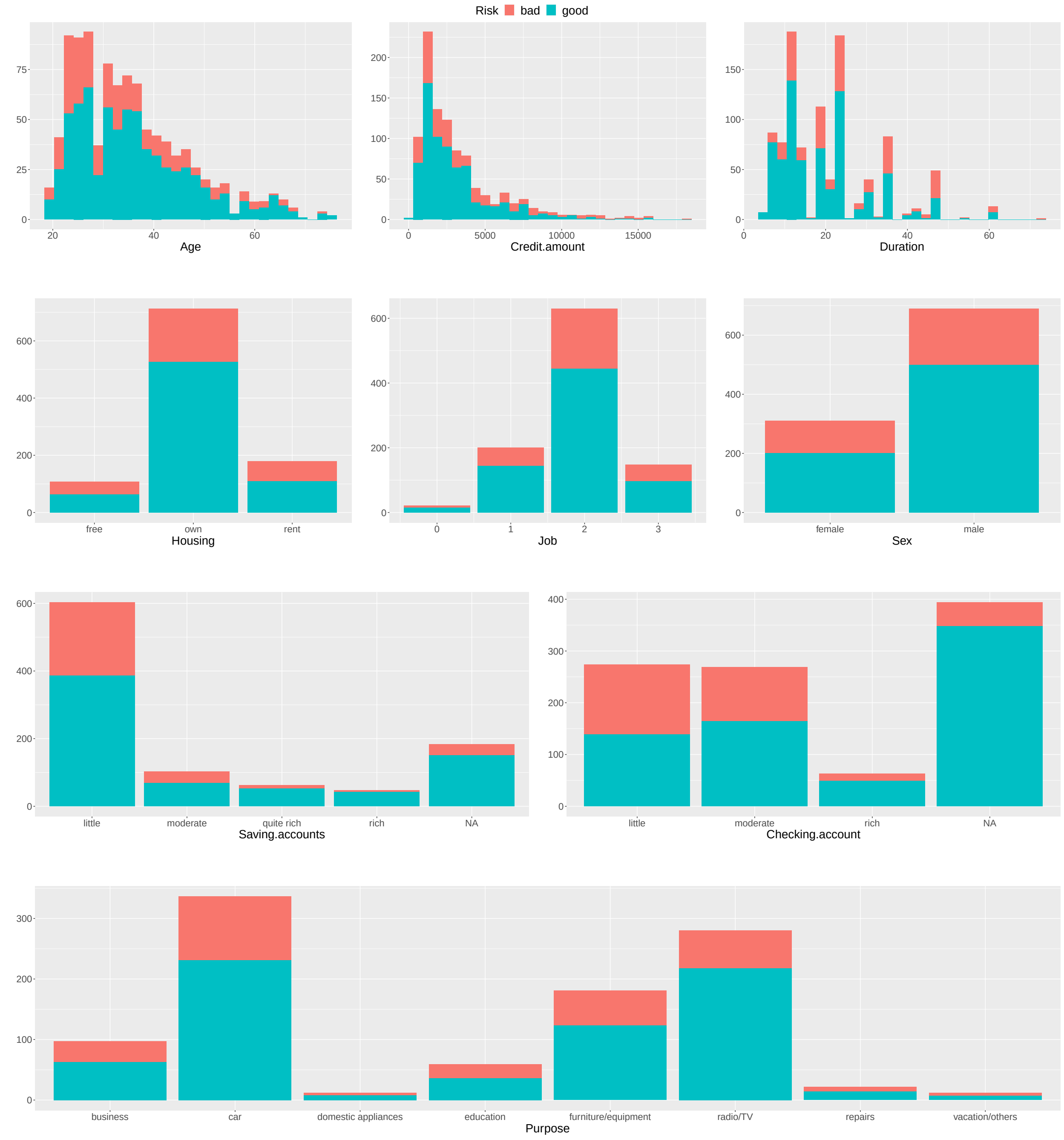
2.1 Analysis problem and existing analysis

The goal of our project is to build a non-hierarchical and a hierarchical model to identify the riskiness of a loan and classify them into good loan and bad loan. We then compare these two models to find out which model is better in predicting the riskiness of a loan.

There are around 80 other people who also tried to solve this problem on Kaggle. After some research on how the other competitors have tried to solve this, we see that logistic regression is a common approach for this problem. However, there is no solution with a complete Bayesian approach, and that is where our model differ from the work of other Kagglers.

2.2 Data preprocessing and cleaning

We can explore the data by plotting histograms for the features corresponding with their target values. The histograms can provide insights regarding the correlation between the explanatory variables (features) with the target variables Risk.



There are some issues that can be observed from the histograms:

- The variables Saving.accounts and Checking.accounts have `NA` values.
- The continuous variables Credit.amount and Duration have strong outliers, as their histograms clearly show a few larger values seperating themselves from the rest of the population.
- The discrete variables Saving.accounts and Purpose have a high sparse level. Their histograms show that the data was categorized into too many groups while some groups are relatively low in population.

We have provide some fix to these issues:

- In order to resolve the high `NA` values presented in two of the features, we decided to impute the missing data with the mode (most frequent value) of data, which, from the plots, is the "`little`" value. Mode imputation is an popular method for filling missing data of categorical data, especially when the data is unbalanced and highly skewed towards the most frequent value like Saving.accounts and Checking.accounts.

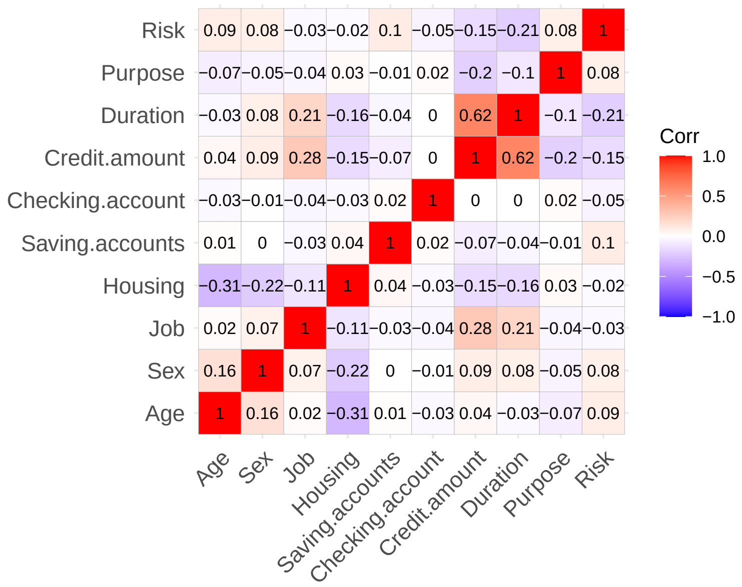
- To resolve the outliers presented in the Credit.amount and Duration, because they are very little in population, we can just remove them from the data. For Credit.amount, we remove all data points containing values larger than 15000; for Duration, we remove all data points with values larger than 65.

- To resolve the high sparse level presented in Saving.accounts and Purpose, a possible solution could be merging the less populated groups with the relevant counterparts. With Saving.accounts, we merge the `rich` and `quite rich` values with each other; with Purpose, we merge `domestic appliances` with `funiture/equipment`, `repairs` with `vacation/others`. This fix will help to data to be less scattered and make the models constructed later to predict stronger results.

One additional, but important, step that could be made to our data is standardization. The input consists of different variables (categorical, numerical, textual) with distinct scales so standardizing them is vital for the models to perform better. In addition, standardizing can facilitate the process of prior choosing for the features, since all variables are on the same scale. To standardize, we transform each explanatory variables to numerical, then scale them so that their mean = 0 and standard deviation = 1.

2.3 Explanatory variables analysis

We can select the final explanatory variables for the models by using a correlation matrix to shows the correlated level of each variables with each other.



From the correlation plot, we observe no correlation greater than 0.9 so no significant multicollinearity is presented. However, it is noticable that the Housing variable shows the lowest correlation with the target variable (only -0.02), indicating that it is quite insignificant to the sampling process. This should be noted, as later on, some actions will be taken to to diminish its effect on the results.

In conclusion, the explanatory variables are chosen for the pooled are Age, Sex, Job, Housing, Saving.accounts, Checking.account, Duration, Purpose, with Housing being the less correlated with the target values. For the hierarchical model, however, Purpose will instead be used as a levels grouping variable and will be excluded from the explanatory variables. # Models description

2.4 Pooled Logistic Regression model

The Logistic Regression model is used to classify the risk y_i in lending a loan based on the observation x_i. Its predictor p is parameterized with an intercept β_0 and explanatory coefficients $\beta_1, \beta_2, ..., \beta_k$ as follows:

$$p = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1 - e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}, \text{ or } p = logit^{-1}(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)$$

The target variable y for the observations x follows the distribution:

$$y \sim Bernoulli(p) = Bernoulli(logit^{-1}(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k))$$

2.4.1 Prior justification

2.4.2 Running the model

2.5 Hierarchical Logistic Regression model

2.5.1 Prior justification

2.5.2 Running the model

3 Results analysis

3.1 Covergence diagnostics

3.2 Model comparison

3.3 Posterior predictive checks

3.4 Predictive performance assessment

3.5 Prior sensitivity analysis

4 Conclusion and potential improvements

5 Self-reflection

6 Appendices and references

6.1 Stan code appendices

6.2 References