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 credits. However, when lenders offer home loans, auto loans, or business loans, there is an inherent risk that borrowers will default on their payments, and 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 to determine the borrowers' ability to meet debt obligations as well as the risks involved, and this process is known as credit risk analysis.

The reason this topic is chosen is because of its meaningfulness and significance to the credit market, and the fact that there has been limited Bayesian data analysis project performed on the topic. Our goal is to use Logistic Regression to identify the riskiness of a loan by classifying them into good and bad loan. A pooled and hierarchical models will be applied to this problem, and through results analysis, we can compare and choose a stronger-performed model for this problem.

The report consists of the following parts: Introduction, Data description, Models, Results, Discussion, Conclusion, and Appendices. First, we formulate the problem and show how we handle the data through pre-processing and feature selection process. 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. Finally, 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 models being used in this report.

2 Data description

The data set is obtained through Kaggle. It contains 1000 data points with 20 categorical attributes created by Professor Hofmann. In this data set, each entry represents a person taking a credit by a bank, and each person is classified as good or bad credit risks according to the set of attributes. The full data set can be found here.

The data consists of 10 columns, 9 explanatory variables and 1 target variable, defined as follows:

Explanatory variables:

- Age (numerical): The age of the subject.
- Sex (textual/categorical): The gender of the subject. This includes “`male`” and “`female`”.
- Job (numerical/categorical): The level of employment of the subject. This includes `0` - unskilled and non-resident, `1` - unskilled and resident, `2` - skilled, `3` - highly skilled.
- Housing (textual/categorical): The type of housing of the subject. This includes “`own`”, “`rent`” and “`free`”.
- Saving.accounts (textual/categorical): The level of wealth of subject's saving account. This includes “`little`”, “`moderate`”, “`quite rich`”, “‘rich“”.
- Checking.account (textual/categorical): The level of wealth of subject's checking account. This includes “`little`”, “`moderate`”, “`quite rich`”, “‘rich“”.
- Credit.amount (numerical): The amount of credit of the subject's loan.
- Duration (numerical): The contracted loan due time (in month).
- Purpose (textual/categorical): The purpose of the subject's loan.

Target variable:

- Risk (textual/categorical): The fact that the subject has defaulted (not repaying the loan). This includes `good` - good loan (the subject has repaid on time), `bad` - bad loan (the subject has defaulted).

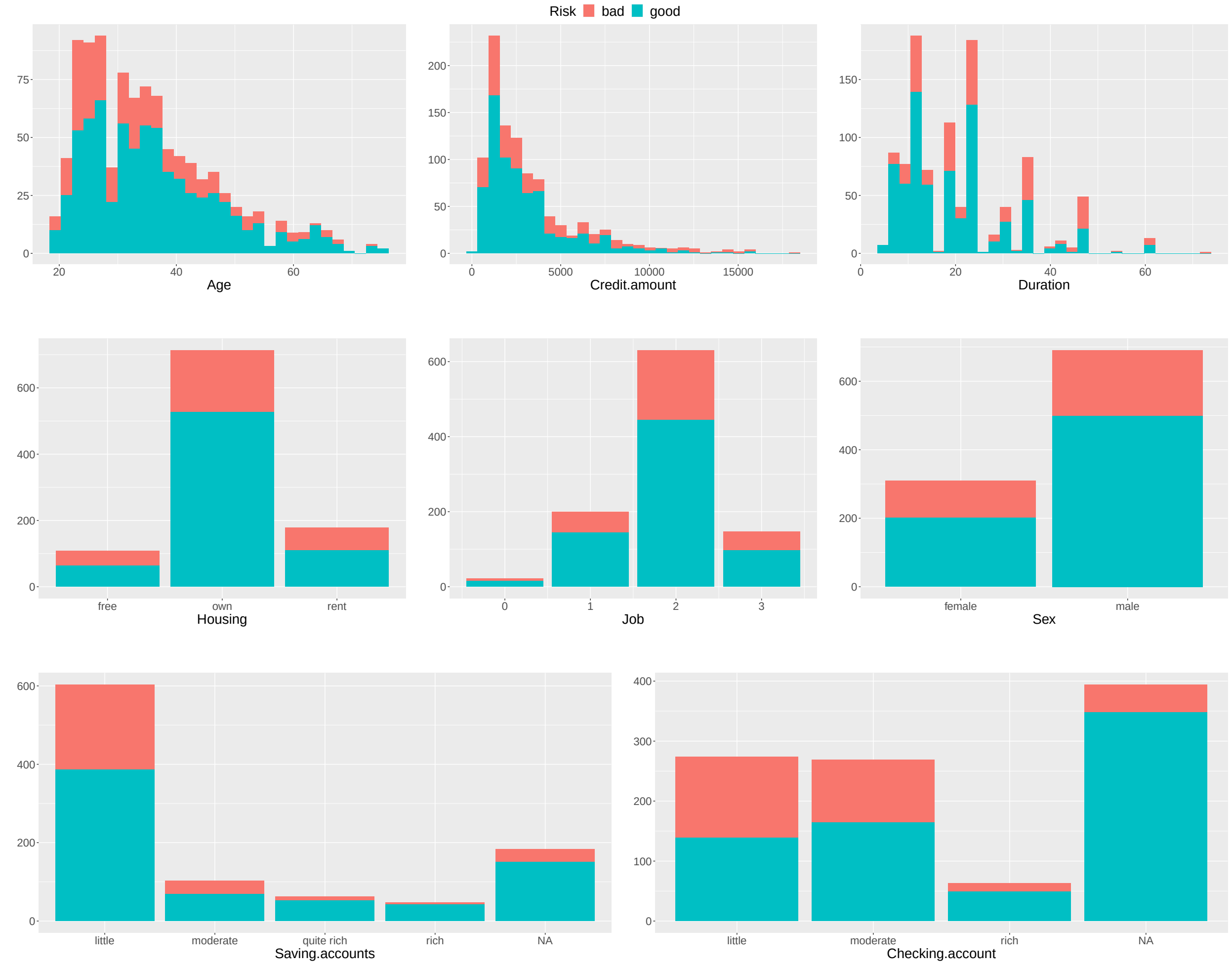
The data set have missing data: the columns Saving.accounts and Checking account both contain `NA` values (missing variables), which takes 18.3% and 39.4% of the total observations, respectively.

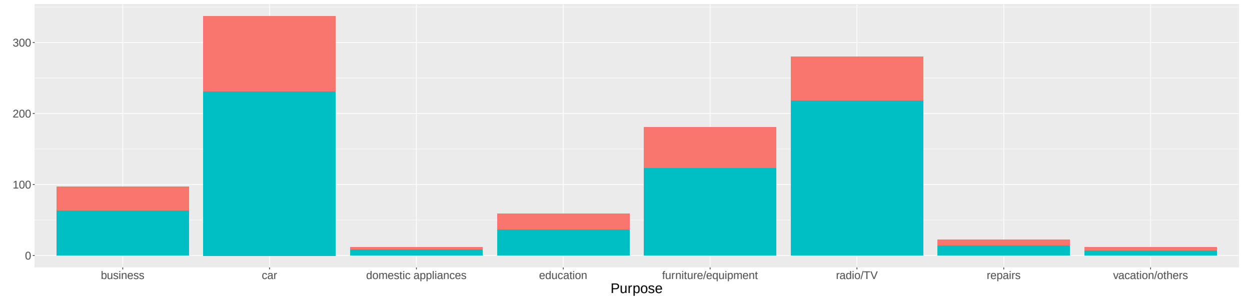
```
isNA <- function(x){sum(is.na(x))/length(x)*100}
apply(credit_risk, 2, isNA)
```

```
##              Age              Sex              Job          Housing
##              0.0              0.0              0.0              0.0
##  Saving.accounts Checking.account    Credit.amount         Duration
##             18.3             39.4              0.0              0.0
##          Purpose             Risk
##              0.0              0.0
```

2.1 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.account have `NA` values (which is also suggested at the Data set description section above).
- The continuous variables Credit.amount and Duration have great outliers, as their histograms clearly show a few significantly larger values clearly separating themselves from the rest of the population.
- The discrete variables Saving.accounts and Purpose have a high scatter level. That is, 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.

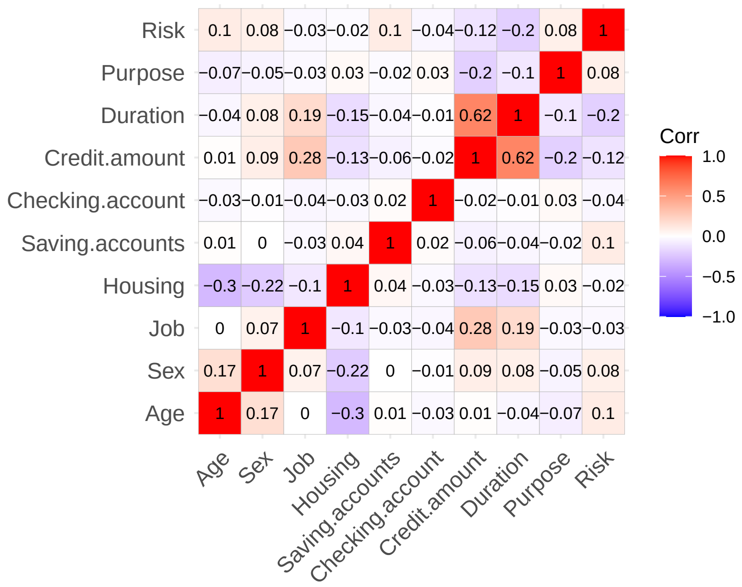
- To resolve the outliers presented in the Credit.amount and Duration, because they are very little in population, we can just remove them from the data. For Credit.amount, we remove all data points containing values larger than 13000; for Duration, the outlier is just a single data point with value equals to 72 so we just delete it from the data.

- 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 into "`rich`"; with Purpose, we merge `domestic appliances` with `funiture/equipment` into "`funiture/equipment`", `repairs` with `vacation/others` into "`others`". This fix will help to data to be less scattered and make the models constructed later to predict more robust 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 all variables (including target) to numerical. Then, we scale the explanatory variables them so that their mean = 0 and standard deviation = 1.

2.2 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.



	Age	Sex	Job	Housing	Saving.accounts	Checking.account	Credit.amount	Duration	Purpose	Risk
Risk	0.1	0.08	−0.03	−0.02	0.1	−0.04	−0.12	−0.2	0.08	1
Purpose	−0.07	−0.05	−0.03	0.03	−0.02	0.03	−0.2	−0.1	1	0.08
Duration	−0.04	0.08	0.19	−0.15	−0.04	−0.01	0.62	1	−0.1	−0.2
Credit.amount	0.01	0.09	0.28	−0.13	−0.06	−0.02	1	0.62	−0.2	−0.12
Checking.account	−0.03	−0.01	−0.04	−0.03	0.02	1	−0.02	−0.01	0.03	−0.04
Saving.accounts	0.01	0	−0.03	0.04	1	0.02	−0.06	−0.04	−0.02	0.1
Housing	−0.3	−0.22	−0.1	1	0.04	−0.03	−0.13	−0.15	0.03	−0.02
Job	0	0.07	1	−0.1	−0.03	−0.04	0.28	0.19	−0.03	−0.03
Sex	0.17	1	0.07	−0.22	0	−0.01	0.09	0.08	−0.05	0.08
Age	1	0.17	0	−0.3	0.01	−0.03	0.01	−0.04	−0.07	0.1

From the correlation plot, we observe no correlation greater than 0.9 between the features 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.

3 Models description

3.1 Pooled Logistic Regression model

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

$$\begin{aligned} p &= \frac{e^{\beta_0+\beta_1 x_{n,1}+...+\beta_k x_{n,k}}}{1-e^{\beta_0+\beta_1 x_{n,1}+...+\beta_k x_{n,k}}} \\ &= logit^{-1}(\beta_0+\beta_1 x_{n,1}+...+\beta_k x_{n,k}) \end{aligned}$$

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

$$y_n \sim Bernoulli(p) = Bernoulli(logit^{-1}(\beta_0+\beta_1 x_{n,1}+...+\beta_k x_{n,k}))$$

where β_0 is the intercept and each $\beta_1, \beta_2, ..., \beta_k$ models the regression coefficient for each feature.

3.1.1 Prior justification

Because of the standardized data, we may assume a commonly-used generic weakly informative priors for the explanatory coefficients:

$$\beta_0, \beta_1, ..., \beta_k \sim normal(0,1)$$

A Normal distribution with location parameter of 0 expresses our little knowledge of whether a change in input variables can affect the predicted target, while a standard deviation of 1 widens the shape of the distribution so that it, softly concentrates below the scale, which is weakly informative enough for our standardized data.

However, as noted, for the potential irrelevant variable Housing, which corresponds to β_3, we assume an informative prior:

$$\beta_3 \sim normal(0, 0.01)$$

This informative prior serves as an powerful *regularization prior* scaling the insignificant coefficient β_3 to nearly zero, but not completely disregard them from the model, therefore shrinking its effect on the results. This effect shrinkage idea is to justify model selection, which is obtain from Paul et al. work.

3.1.2 Running the model

Before running, the input data was separated into training data and testing data with equal proportions (half) of the data.

```
train_size <- round(nrow(data_pooled)/2, 0)
test_size <- nrow(data_pooled) - train_size
cat("Train size:", train_size, "/ Test size:", test_size)
```

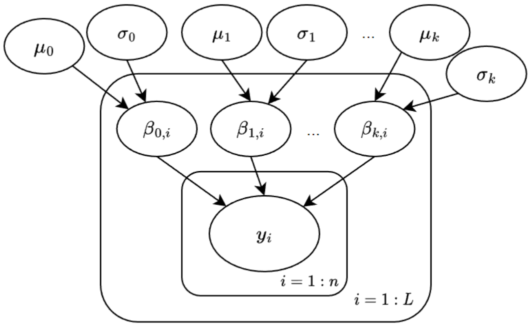
```
## Train size: 493 / Test size: 493
```

The model is then run with 4 chains, 2000 iterations and 1000 warm-ups.

```
data_train <- head(data_pooled, train_size)
data_test <- tail(data_pooled, test_size)

y_train_pooled = data_train$Risk
y_test_pooled = data_test$Risk

X_train_pooled = subset(data_train, select=-c(`Risk`))
X_test_pooled = subset(data_test, select=-c(`Risk`))

credit_data_pooled <- list(N_train=nrow(X_train_pooled),
                           D=ncol(X_train_pooled),
                           X_train=X_train_pooled,
                           y_train=y_train_pooled,
                           N_test=nrow(X_test_pooled),
                           X_test=X_test_pooled)

pooled_fit <- stan(file="pooled.stan", data=credit_data_pooled,
                   chains=4, iter=2000, warmup=1000)
```

3.2 Hierarchical Logistic Regression model

The hierarchical model, though resembles the pooled model in using Logistic Regression, is a relatively more complex model. As mentioned in the Explanatory variables analysis section, the data is categorized into L groups using the categorical data in the Purpose variable. The coefficients $\beta_0, \beta_1, ..., \beta_k$ is then sampled separately for each of L groups using their own *hyper-parameters* μ and σ. The structuring of the model can be given as below:



With each of L groups having their own $\beta_0, \beta_1, ..., \beta_k$ sampled separately, we can build a Logistic Regression equation for the target variable:

$$y_n \sim Bernoulli(logit^{-1}(\beta_{0,L(n)} + \beta_{1,L(n)}x_{n,1} + ... + \beta_{k,L(n)}x_{n,k})),$$
$$\text{with } L(n) \text{ being the group of the } n^{th} \text{ data}$$

3.2.1 Prior justification

We can estimate the coefficients $\beta_{0,L(n)}, \beta_{1,L(n)}, ..., \beta_{k,L(n)}$ using a Normal prior with the hyper-parameters μ and σ:

$$\beta_{i,L(n)} \sim normal(\mu_i, \sigma_i), \text{ with } i = 1 : k$$

For the hyper-parameter μ_i, we assume a the same generic weakly-informative prior:

$$\mu_i \sim normal(0, 1)$$

Note that because of the standardization step, this prior is considered to be weakly-informative enough to not creating biases. We have explained the choice of this prior above in the pooled model.

For the hyper-parameter σ, we assume a weakly-informative inverse Gamma prior:

$$\sigma_i \sim \Gamma^{-1}(0.5, 1)$$

3.2.2 Running the model

4 Results analysis

4.1 Covergence diagnostics

4.2 Model comparison

4.3 Posterior predictive checks

4.4 Predictive performance assessment

4.5 Prior sensitivity analysis

5 Conclusion and potential improvements

6 Self-reflection

7 Appendices and references

7.1 Stan code appendices

7.2 References

Hahn, Paul & Carvalho, Carlos. (2014). Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective. Journal of the American Statistical Association. 110. 10.1080/01621459.2014.993077.