STAT 445 Term Project

German Credit Analysis

Youneng Zou

April 12, 2018

Table of Contents

Abstract

In this report, we will explore in this data analysis the behavior of German borrowers. In the dataset, there are 1,000 observations and 11 variables. The first thing we will do is the data cleaning. After the data cleaning, we will do the variable selection and model selection. Finally, we will develop a model to predict whether a loan will have a good or bad risk and we will also analyze the key patterns for a bad risk loan.

Keywords:

Loan risk, age, sex, job, housing, saving accounts, checking account, credit amount, duration, purpose

Introduction

A credit risk is the risk of default on a debt that may arise from a borrower failing to make required payments. To reduce the lender's credit risk, the lender wants to perform a credit check on the borrower, may require the borrower to take mortgage insurance. In this project, we want to consider the credit risk of the borrower, which contain two categories (good or bad), as the response variable, and other attributes of the borrower as the explanatory variables.

Dataset:

The dataset had 1,000 observations with 10 variables and it is shared at Kaggle.com. It contains 1,000 borrowers and their behaviors, such as their age, job, wealth, the amount they want to borrow and the purpose and duration. There are some missing values and unrelated variables in the dataset so I have to do data cleaning first. After data cleaning, the dataset has following variables.

Response variables:

1. Risk: Categorical (good, bad)

Explanatory variables:

1. Age: Numerical
2. Sex: Categorical (male, female)
3. Job: Categorical (unskilled and non-resident, unskilled and resident, skilled, highly skilled)
4. Housing: Categorical (own, rent, free)
5. Saving.accounts: Categorical (little, moderate, quite rich, rich)
6. Checking.account: Categorical (little, moderate, rich)
7. Credit.amount: Numerical
8. Duration: Numerical
9. Purpose: Categorical (business, car, domestic appliances, education, furniture/equipment, radio/TV, repairs, vacation/others)

Goals:

1. To construct and determine an appropriate model as the final model.
2. To interpret the effects of the explanatory variables in the final model, and analyze some key explanatory variables and their effects.

Methods and Models

In this project, we wanted to develop an appropriate model to classify the credit risk of the borrower. Since the response variable is the credit risk, which is categorical, the basic regression model we choose is logistic regression model. We first constructed the full model, which includes all the explanatory variables:

$$Logit\big(\pi(Age, \dots, Purpose)\big) = \alpha + \beta_1 * Age + \beta_2 * Sex + \beta_3 * Job + \beta_4 * Housing + \beta_5 * Saving.accounts + \beta_6 * Checking.account + \beta_7 * Credit.amount + \beta_8 * Duration + \beta_9 * Purpose$$

The full model was initially fit on a training dataset, that includes 80% of the observations. Successfully, the fitted model was used to predict the responses for the validation dataset, that includes 20% of the observations. Furthermore, we calculated the accuracy of this model.

Analysis of the Full Model

The full model we constructed has 9 explanatory variables and 6 of them are categorical, with the brief summary of the model below:

Summary of full model				
Residual deviance	635.68	Significant Predictors	Estimated Value	p-value
Degrees of freedom	500	Saving.accountsrich	-1.345	0.02161 *
		Checking.accountrich	-1.029	0.00585 **
		Duration	0.05104	0.00000216 ***
		Purposeeducation	1.097	0.04499 *

In the full model, there are 5 significant predictors with p-value < 0.05.

Consequently, in order test whether there are unnecessary variables, Least Absolute Shrinkage and Selection Operator (LASSO) was used. We also calculated the predicted accuracy of the LASSO regularization.

Variable Selection

In order to test whether there are unnecessary variables, we used LASSO regularization.

Summary for LASSO	
23 x 1 sparse Matrix of class "dgCMatrix"	
(Intercept)	0.63716620

Duration	-0.02075617

For the LASSO regularization, the output showed that only the duration of the loan that we had determined to be significant on the basis of p-values have non-zero coefficients. The coefficients of all other variables have been set to zero by the algorithm. It doesn't make sense to remove too many predictors and we also mentioned another reason before, so we decided not to use this one.

Full model			**LASSO**		
	bad	good		bad	good
bad	23	16	bad	13	4
good	16	42	good	26	54
Accuracy: 67.01%			**Accuracy: 69.07%**		

From the table above, we saw although the LASSO has higher accuracy, the full model predicts the bad credit risk much better. Remember in this project, we want to detect the bad credit risk borrowers so we don't approve their loan. The full model definitely did better here, so we still chose the full model which contains all the variables.

After variable selection, we did the model selection. Cross-validation was used to compare different models and pick up the suitable final model. We chose 10-fold cross-validation, the original observations were randomly partitioned into 10 equal size subsamples. Of these 10 subsamples, a single subsample was retained as the validation set for testing the model, and the remaining 9 subsamples were used as training set. The cross-validation process was then

repeated 10 times, with each of the 10 subsamples used exactly once as the validation set. The two model we compared were Logistic Regression and Linear Discriminant Analysis (LDA):

Logistic			**LDA**		
	bad	good		bad	good
bad	115	71	bad	114	67
good	115	219	good	116	223
Accuracy: 64.23%			**Accuracy: 64.80%**		

Although LDA has slightly better accuracy, we still chose logistic model as our final model since logistic model predict bad credit risk slightly better as we mentioned before.

Results

	Coefficient
Intercept	-0.6492
Age	-0.009741
Sex (male)	-0.3691
Job (skilled)	0.1152
Job (unskilled and non-resident)	0.3189
Job (unskilled and resident)	0.07172
Housing (own)	-0.2672

Housing (rent)	0.06059
Saving.accounts (moderate)	-0.1212
Saving.accounts (quite rich)	-0.7095
Saving.accounts (rich)	-1.345
Checking.account (moderate)	-0.1882
Checking.account (rich)	-1.029
Credit.amount	0.000006004
Duration	0.05104
Purpose (car)	0.3597
Purpose (domestic appliances)	0.7246
Purpose (education)	1.097
Purpose (furniture/equipment)	0.07653
Purpose (radio/TV)	0.09924
Purpose (repairs)	0.3798
Purpose (vacation/others)	-0.3488

According to the coefficients we have above, since it is logistic regression model, we can only tell how much the log-odds of the probability of bad credit risk increase. But we can simply say, the larger the coefficient is, the higher the probability to have a bad credit risk. We made following statement:

1. Per unit increase in age, the estimated decrease in the log-odds of the probability of bad credit risk is 0.009741.
2. If the sex changes from female to male, the estimated decrease in the log-odds of the probability of bad credit risk is 0.3691.
3. If the job changes from highly skilled to unskilled and resident, skilled, unskilled and non-resident, the estimated increase in the log-odds of the probability of bad credit risk are 0.07172, 0.1152, 0.3189.
4. If housing condition changes from free to own, the estimated decrease of the log-odds of the probability of bad credit risk is 0.2672. If housing condition changes from free to rent, the estimated increase of the log-odds of the probability of bad credit risk is 0.06059.
5. If the borrower's saving account is increasing from little to moderate, quite rich, rich, the estimated decrease of the log-odds of the probability of bad credit risk is 0.11212, 0.7095, 1.345.
6. If the borrower's checking account is increasing from little to moderate, rich, the estimated decrease of the log-odds of the probability of bad credit risk is 0.1882, 1.029.
7. Per unit increase in the credit amount of the loan, the estimated decrease in the log-odds of the probability of bad credit risk is 0.000006004.
8. Per unit increase in the duration of the loan, the estimated decrease in the log-odds of the probability of bad credit risk is 0.05104.
9. If the purpose changes from business to vacation/others, the estimated decrease in the

log-odds of the probability of bad credit risk is 0.3488. All the others purposes increase the probability to have bad credit risk.

Assuming we know all variables of a borrower, the estimated probability of having a bad credit risk is:

$$Propability\ of\ having\ a\ bad\ credit\ risk = \frac{1}{1+\exp(\alpha+\beta_1*Age+\beta_2*Sex+\beta_3*Job+\beta_4*Housing+\beta_5*Saving.accounts+\beta_6*Checking.account+\beta_7*Credit.amount+\beta_8*Duration+\beta_9*Purpose)}$$

Conclusion

Base on the results, we can make the following conclusion:

1. As age increase, the probability of having a bad credit slight decreases.
2. The gender actually plays an important role in applying for the loan, it decreases the probability of having a bad credit a lot.
3. Based on the job condition, a better job will have a higher probability to avoid bad credit risk.
4. Based on the housing condition, whether the borrower owns a house will influence the probability of ad credit risk.
5. We can consider saving account and checking account as wealth. The higher the wealth, the lower the probability of being a bad credit risk.
6. The higher the credit amount of the loan, the most likely the loan will end up bad.

7. The longer the duration of the loan, the most likely the loan will turn out bad.
8. For the purposes of the loan, vacation/others and business are the less risky purposes.

Potential Problems and Possible Solution

Although we can conclude our final model quite well, there are still potential problems:

1. It is no hard to see the decreased probability from female to male is more than the decreased probability from free housing to own housing and it is unreasonable in the reality. The sex and job table below showed that most females are unemployed compared to male. We may need to further investigate it by considering the interaction of sex and housing.

	Highly skilled	Skilled	Unskilled and non-resident	Unskilled and resident
female	22	98	7	41
male	57	215	7	75

2. As ages increases, the probability of having a bad credit risk shouldn't keep decreasing. The probability-age curve should have a parabola shape curve in reality. One of the possible solutions can be dividing the ages into group such as young, adult, senior, elder.

References

Bilder, Christopher R., and Thomas M. Loughin. *Analysis of Categorical Data with R*. Boca Raton: CRC, 2015.

Ferreira, Leonardo. German_credit_data_with_risk | Kaggle. January 09, 2018. Accessed April 17, 2018. https://www.kaggle.com/kabure/german-credit-data-with-risk/data.

Johnson, Richard Arnold., and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Upper Saddle River: Pearson Education International, 2007.

Appendix

Summary of full model from training dataset:

```
Call:
glm(formula = (Risk == "bad") ~ ., family = binomial(), data = trainset)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.9692  -0.9970  -0.6387   1.0904   2.0766

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -9.225e-01  8.031e-01  -1.149   0.2507
Age                          -7.259e-03  1.006e-02  -0.722   0.4705
Sexmale                      -2.661e-01  2.336e-01  -1.139   0.2546
Jobskilled                    1.765e-01  3.421e-01   0.516   0.6058
Jobunskilled and non-resident  1.139e+00  8.433e-01   1.350   0.1769
Jobunskilled and resident     1.259e-01  3.954e-01   0.318   0.7502
Housingown                   -1.814e-01  3.613e-01  -0.502   0.6156
Housingrent                   3.023e-01  4.178e-01   0.724   0.4693
Saving.accountsmoderate      -5.778e-03  3.480e-01  -0.017   0.9868
Saving.accountsquite rich    -6.004e-01  6.354e-01  -0.945   0.3447
```

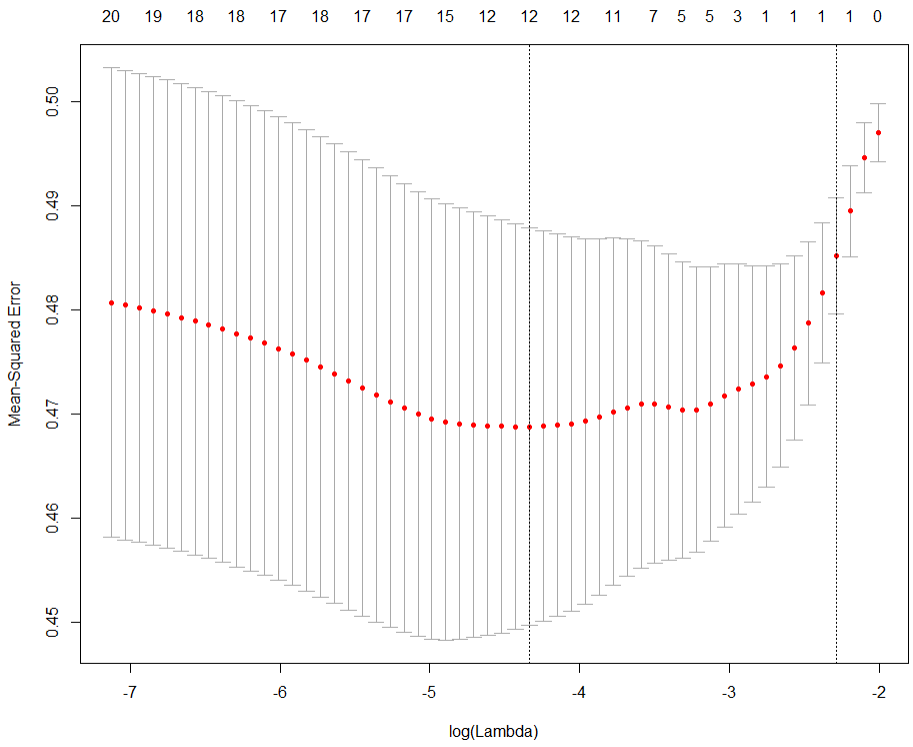
```
Saving.accountsrich          -1.233e+00  6.818e-01  -1.808   0.0706 .
Checking.accountmoderate      -3.654e-01  2.401e-01  -1.522   0.1280
Checking.accountrich         -9.444e-01  3.854e-01  -2.450   0.0143 *
Credit.amount                1.749e-05  5.185e-05   0.337   0.7359
Duration                    4.639e-02  1.177e-02   3.943 8.04e-05 ***
Purposecar                  4.819e-01  4.117e-01   1.170   0.2418
Purposedomestic appliances    5.997e-01  1.021e+00   0.587   0.5571
Purposeeducation             8.655e-01  5.838e-01   1.483   0.1382
Purposefurniture/equipment    1.281e-01  4.342e-01   0.295   0.7680
Purposeradio/TV              2.349e-01  4.195e-01   0.560   0.5756
Purposerepairs               4.788e-01  7.142e-01   0.670   0.5026
Purposevacation/others        1.559e-01  9.836e-01   0.158   0.8741
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 585.21  on 424  degrees of freedom
Residual deviance: 526.10  on 403  degrees of freedom
AIC: 570.1

Number of Fisher Scoring iterations: 4
```

Plot of LASSO:



Coefficients of LASSO:

```
23 x 1 sparse Matrix of class "dgCMatrix"
                                       1
(Intercept)                  0.41714987
(Intercept)                  .
Age                          .
Sexmale                      .
Jobskilled                   .
Jobunskilled and non-resident  .
Jobunskilled and resident    .
Housingown                   .
Housingrent                  .
Saving.accountsmoderate      .
Saving.accountsquite rich    .
Saving.accountsrich          .
Checking.accountmoderate     .
Checking.accountrich         .
Credit.amount                .
Duration                    -0.01047241
Purposecar                   .
```

```
Purposedomestic appliances    .
Purposeeducation             .
Purposefurniture/equipment    .
Purposeradio/TV              .
Purposerepairs               .
Purposevacation/others        .
```

Summary of final model:

```
Call:
glm(formula = (Risk == "bad") ~ ., family = binomial(), data = credit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0993  -0.9812  -0.6104   1.0943   2.1488

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -6.492e-01  7.313e-01  -0.888  0.37470
Age                         -9.741e-03  9.068e-03  -1.074  0.28273
Sexmale                     -3.691e-01  2.123e-01  -1.738  0.08215 .
Jobskilled                   1.152e-01  3.054e-01   0.377  0.70611
Jobunskilled and non-resident  3.189e-01  6.981e-01   0.457  0.64779
Jobunskilled and resident     7.172e-02  3.525e-01   0.203  0.83877
Housingown                  -2.672e-01  3.336e-01  -0.801  0.42305
Housingrent                  6.059e-02  3.804e-01   0.159  0.87345
Saving.accountsmoderate      -1.212e-01  3.070e-01  -0.395  0.69288
Saving.accountsquite rich    -7.095e-01  5.315e-01  -1.335  0.18191
Saving.accountsrich          -1.345e+00  5.855e-01  -2.297  0.02161 *
Checking.accountmoderate     -1.882e-01  2.152e-01  -0.874  0.38192
Checking.accountrich         -1.029e+00  3.734e-01  -2.756  0.00585 **
Credit.amount                6.004e-06  4.681e-05   0.128  0.89793
Duration                     5.104e-02  1.077e-02   4.737 2.16e-06 ***
Purposecar                   3.597e-01  3.627e-01   0.992  0.32129
Purposedomestic appliances    7.246e-01  9.419e-01   0.769  0.44172
Purposeeducation             1.097e+00  5.473e-01   2.005  0.04499 *
Purposefurniture/equipment    7.653e-02  3.889e-01   0.197  0.84399
Purposeradio/TV              9.924e-02  3.749e-01   0.265  0.79125
Purposerepairs               3.798e-01  6.635e-01   0.572  0.56701
Purposevacation/others       -3.488e-01  8.302e-01  -0.420  0.67437
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 716.73  on 521  degrees of freedom
Residual deviance: 635.68  on 500  degrees of freedom
AIC: 679.68

Number of Fisher Scoring iterations: 4
```