Analysis of German Credit Loan Risk Data

Chi-Hua Wang, Junrong Zha, Sirui Wang

Department of Statistics, Purdue University

Department of Earth, Atmospheric, and Planetary Sciences, Purdue University

Analysis of German Credit Loan Risk Data

Introduction

The German credit loan risk data, available at archive.ics.uci.edu /ml /datasets /statlog+(german+credit+data), contains 1000 loan applications, each consist of 1 credit status and 20 profile variables of the applicant. The credit status shows the tendency of applicants to repay their loan, in the sense of good credit (Y = 1) indicating they will repay the loan while bad credit (Y = 0) will not. For the bank operating the loan business, a means of judging applicants with good or bad credit status by their profile variables is crucial. In one hand, the bank can make profit from loan business by accepting the application from an applicant with good credit status. On the other
hand, the bank should reject the bad-credit applicant to avoid potential financial losses.

Two scientific question about German credit loan risk data are addressed in this data analysis. The first one is how does applicants' credit status depends on their own profile variables. We answered this question via a well-established logistic model on the conditional probability of good credit status given profile variables and identified a subset of significant profile variables that majors the mechanism of credit status. The second question is located on the dependencies between the profile variables of an applicant. Method of two-way contingency table analysis was adopted to establish the pairwise independence between categorical profile variables that are statistically significant in the previous logistic model. A

subset of categorical profile variables concerning the capability of an applicant to repay the loan was identified from the adopted method.

Description of Data Set

The dataset contains 1000 applicants for loan business, each applicant has 1 binary response variable indicate the credit status and 20 predictor variables (3 continuous and 17 categorical) describing the profiles of loan applicant. The following is complete list of variables:

Variable	Variable Name	Property
Y	Creditability	Binary Response X_1
	Account.Balance	Categorical variable X_2
	Duration.of.Credit (month)	Continuous Variable X_3
	Payment.Status.of.Previous.Credit	Categorical variable X_4
	Purpose	Categorical variable X_5
	Credit.Amount	Continuous Variable X_6
	Value.Savings.Stocks	Categorical variable X_7
	Length.of.current.employment	Categorical variable X_8
	Instalment.per.cent	Categorical variable X_9
	Marital.Status/Sex	Categorical variable X_{10}
	Guarantors	Categorical variable X_{11}
	Duration.in.Current.address	Categorical variable X_{12}
	Most.valuable.available.asset	Categorical variable X_{13}
	Age(Years)	Continuous Variable X_{14}
	Concurrent.Credits	Categorical variable X_{15}
	Type.of.apartment	Categorical variable X_{16}
	No.of.Credits.at.this.Bank	Categorical variable X_{17}
	Occupation	Categorical variable X_{18}
	No.of.dependents	Categorical variable X_{19}
	Telephone	Categorical variable X_{20}
	Foreign.Worker	Categorical variable

Exploratory Data Analysis

In this section, we perform exploratory data analysis to search potential association between credit status and profiles variables. For categorical profile variables, mosaic plot of proportion of each possible combination of credit status and profile variable are adopted to observe whether there is some specific level of categorical variable has higher proportion than other levels. For continuous profile variables, simple logistic regression was fitted and the estimated conditional probability of getting good credit status were shown to observe the positive or negative influence of continuous profile variables in credit status. The mosaic plot and simple logistic regression plot of each profile variables are given in Figure 1 20 and the main observation are summarized in the following table:

Variable	Variable Name	Property
Y	Creditability	
X_1	Account.Balance	**Higher account balance, better credit.** X_2
Duration.of.Credit (month)		**Shorter Duration of Loan, Better Credits** X_3
Payment.Status.of.Previous.Credit		**Pay Back to get Good Credit.**
X_4	Purpose	**Top 3 Good Credit: New Car, Furniture, Business**
X_5	Credit.Amount	**Lower amount of Loan, Better Credits** X_6
Value.Savings.Stocks		**More Savings or Stocks, Better Credit.** X_7
Length.of.current.employment		**Stable Job, Good Credit**
X_8	Instalment.per.cent	**Higher percentage of installment income, better credit.** X_9
Marital.Status/Sex	**Married/ Widowed Male > Divorced/ Living apart Male** X_{10}	Guarantors
Have Guarantor, better Credit		
X_{11}	Duration.in.Current.address	**Nothing Special.**
X_{12}	Most.valuable.available.asset	**No Asset, Better Credict.**
X_{13}	Age(Years)	**Older, Better Credit.**
X_{14}	Concurrent.Credits	**No further running credits, better credits.**
X_{15}	Type.of.apartment	**Rented Flat, Better Credits**
X_{16}	No.of.Credits.at.this.Bank	**More previous Credits, Better Credits**
X_{17}	Occupation	**Nothing Special**
X_{18}	No.of.dependents	**Nothing Special**
X_{19}	Telephone	**Have Telephone, Better Credits**
X_{20}	Foreign.Worker	**NOT Foreign worker, Better Credits**

Data Analysis and Results

Our data analysis focus on two scientific problems: how does applicants' status of credit depends on their own profile and what are the dependencies between the subjects in applicants' profile. First problem is answered by the result of logistic regression model on the probability of a loan applicant getting good credit given his or her current profile value. Next, focusing on the set of statistically significant profile variables in the above logistic regression, we perform Fisher's exact test and Pearson's Chi-Square test to explore the possible pairwise independence between categorical profile variables.

Logistic Regression of Probability of getting Good Credit

We describe the relationship of binary response variable Y with profile variables X_1, X_2, , X_{20} via modeling the conditional probability of getting good credit status (Y = 1) given the profile variables X_1, X_2, , X_{20} as

$$P\ (\text{Good Credit} X_1, \quad , X_{20}) = P\ (Y = 1 X_1, \quad , X_{20}) = \frac{e^{\beta_0+\beta_1 X_1 + + \beta_{20} X_{20}}}{1 + e^{\beta_0+\beta_1 X_1 + + \beta_{20} X_{20}}}.$$

In the framework of logistic regression, the log odds of getting good credit status given profile variables are determined by the linear combination of profile variables, i.e.,

$$\text{Logit}(P\ (\text{Good Credit} X_1, \quad , X_{20})) = \beta_0 + \beta_1 X_1 + \quad + \beta_{20} X_{20}.$$

Since we have bunch of profile variables that can explain the mechanism of getting good credit status, we pursue the most simplified model to enhance the meaning of resulting regression coefficient. The variable selection procedure we adopted in the data analysis is backward elimination until all remaining profile variable are statistically significant under α level 0.05. The resulting model had drop 9 and keep 11 profile variable. The dropped profile variables are presented in the following table:

Variable	Variable Name
X_4	Purpose
X_{11}	Duration.in.Current.address X_{12}
Most.valuable.available.asset	X_{13}
Age in years	
X_{15}	Type.of.apartment
X_{16}	No.of.Credits.at.this.Bank
X_{17}	Occupation X_{18}
No.of.dependents X_{19}	
Telephone	

Next, the regression coefficient of statistically significant profile variables ordered by its significance are reported in the following:

	Variable	Estimated Coefficient
$\hat{\beta}_0$	Intercept	-4.46
$\hat{\beta}_{20}$	Foreign.Worker	1.13
$\hat{\beta}_1$	Account Balance	0.59
$\hat{\beta}_{10}$	Guarantors	0.36
$\hat{\beta}_3$	Previous Payment Status	0.35 $\hat{\beta}_2$
	Duration of Credit	-0.32 $\hat{\beta}_8$
	Installment.per.cent	-0.27 $\hat{\beta}_9$
	Sex.Martial.Status	0.27 $\hat{\beta}_{14}$
	Concurrent.Credits	0.24 $\hat{\beta}_6$
	Value.Savings.Stocks	0.24 $\hat{\beta}_5$
	Credit Amount	-0.23 $\hat{\beta}_7$
	Length of Employment	0.16

One can observe the top five significant profile variables are: Foreign.Worker, Account Balance, Guarantors, Previous Payment Status and Duration of Credit. In words, the people who are local resident, has higher account balance, one or two guarantors, paid out the previous loan and asked for short duration of loan are moss preferable by the bank in loan business.

Pairwise Independence between Significant Categorical Profile Variables

Followed from the above logistic regression, the 11 statistically significant profile variables have 8 categorical variable: X_1 : Account Balance, X_3 : Previous Payment Status, X_6 : Value Savings Stocks, X_7 : Length of Employment, X_9 : Sex Martial Status, X_{10} : Guarantors, X_{14} : Concurrent.Credits, X_{20} : Foreign.Worker.

Our subsequent analysis is explore the pairwise independence between these 8 significant categorical profile variables.There are total 28 pairs between these 8 variables. Fisher's exact test and Pearson's Chi-sq test gave us p-values of testing on pairwise independence. Under significance level 0.05, we accept null hypothesis when p-value is larger than 0.05, meaning two variables are independent. The p-values and results of pairwise independence test are summarized by the following table:

Tested Pair	p-value	Indep	Tested Pair	p-value	Indep
(X_1, X_3)	< 0.001		(X_6, X_9)	0.072	V
(X_1, X_6)	< 0.001		(X_6, X_{10})	0.010	
(X_1, X_7)	0.009		(X_6, X_{14})	0.997	V
(X_1, X_9)	0.327	V	(X_6, X_{20})	0.886	V
(X_1, X_{10})	< 0.001		(X_7, X_9)	< 0.001	
(X_1, X_{14})	0.339	V	(X_7, X_{10})	0.064	V
(X_1, X_{20})	0.04		(X_7, X_{14})	0.752	V
(X_3, X_6)	0.126	V	(X_7, X_{20})	0.203	V
(X_3, X_7)	0.002		(X_9, X_{10})	0.878	V
(X_3, X_9)	0.050	V	(X_9, X_{14})	0.527	V
(X_3, X_{10})	0.014		(X_9, X_{20})	0.062	V
(X_3, X_{14})	< 0.001		(X_{10}, X_{14})	0.401	V
(X_3, X_{20})	0.677	V	(X_{10}, X_{20})	0.002	
(X_6, X_7)	0.026		(X_{14}, X_{20})	0.922	V

From above table, we founded 16 pairs of independence and 12 pairs of dependence. Based on the results, we plot the adjacency between these 8 variable in Figure 21 and divide them into two groups as

Group I

- X_1 : Account Balance
- X_3 : Previous Payment Status
- X_6 : Value.Savings.Stocks
- X_7 : Length of Employment
- X_{10} : Guarantors

Group II

- X_9 : Sex.Martial.Status
- X_{14} : Concurrent.Credits
- X_{20} : Foreign.Worker

As one can observe, the profile variables included in Group I concerned the loan applicant's capability of paying back the loan. The profile variables included in Group II are not directly related to the loan applicant's capability of paying back the loan, but related to the profile variables in Group I. Thus, the profile variables in Group II are minor but important variables in loan business.

Model Diagnostics

In this section, we perform model diagnostic to access the appropriateness of fitted logistic model. The residuals versus fitted plot, the normal probability plot, the scale-location plot and the Cook's distance are reported in Figure. 22. The residuals versus fitted plot and the scale-location plot appears as a common situation when one using logistic regression. The normal probability plot suggests residuals slight deviated from normal distribution. All applicant has small Cook's distance so we need not to concern outlier problem. The 1.03 estimated dispersion parameter value suggests our fitting does not subject to overdispersion issue. The 0.64 p-value based on chi-square statistic gives an evidence of the good git of the model to the data. In conclusion, although residual has some weak signal of deviated from assumption, the fitting is overall a good fit and can effectively explain how the profile variables affects the status of good credit.

Discussion of the Results

At the stage of logistic regression model, we establish a model containing 11 profile variables and the model diagnostic suggest the residuals may slight deviated from normal distribution. One possible remedy is using the generalized linear model methodology to explore more general variance function to resolve this situation. Besides, is the result of variable selection will remain same if we adjust the significance level used in backward elimination? The robustness of variable selection result is important to the subsequent pairwise independence testing and the classification of categorical profile variables. After we performed the pairwise independence testing between significant profile variables, can we use the result of classification of profile variable to explore possible dependency in three-way contingency table? As above mentioned, there are still many interesting issue worth to look at in this approach of analysis credit loan risk data.

Conclusion

For German Credit Loan Risk Data, which contains 1 binary response variable and 20 profile variables, we have build a logistic regression of the probability of loan applicants getting good credit in account of all their profile variables. The resulting logistic model possess 11 profile variables to describe the probability of loan applicants getting good credit, where top 5 profile variables are the status of foreign worker, the amount of account balance, the number of guarantors, the status of previous loan payment and the duration of the applicant's loan. Furthermore, by analyzing the pairwise independence between the categorical profile variables, we classify the categorical profile variables which are statistically significant in the logistic regression model into two groups, one is a group of profile variables concerned applicants' capability of paying back their loan, and second is a group of minor but important profile variable in loan business. We wish this data analysis have gives a better understanding about the mechanism of loan business in bank and serve as a guidance for the loan-applying people to enhance their probability of getting loan.

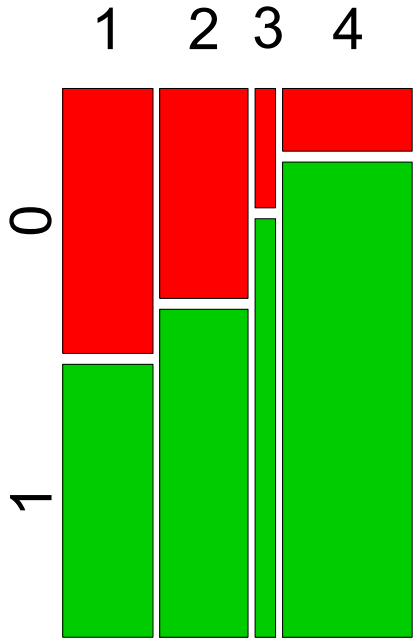

Figure 1 . Mosaic Plot of X_1: Account Balance

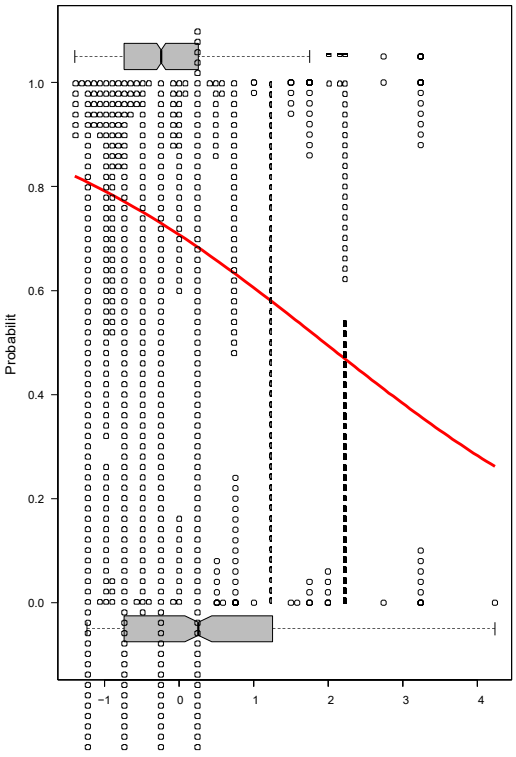

Figure 2 . Mosaic Plot of X_2: Duration of Loan in month

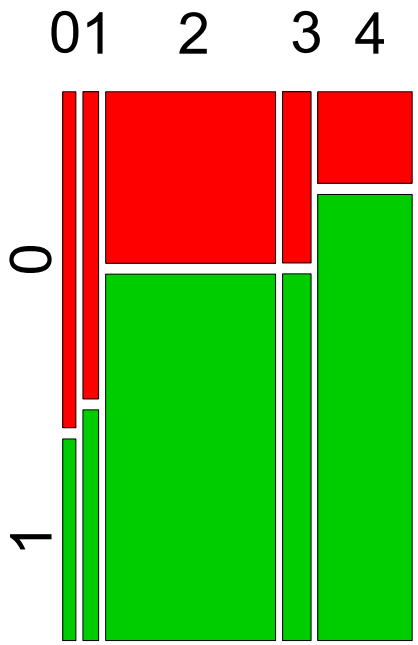

Figure 3 . Logistic Regression Plot of X_3: Payment Status of Previous Credit

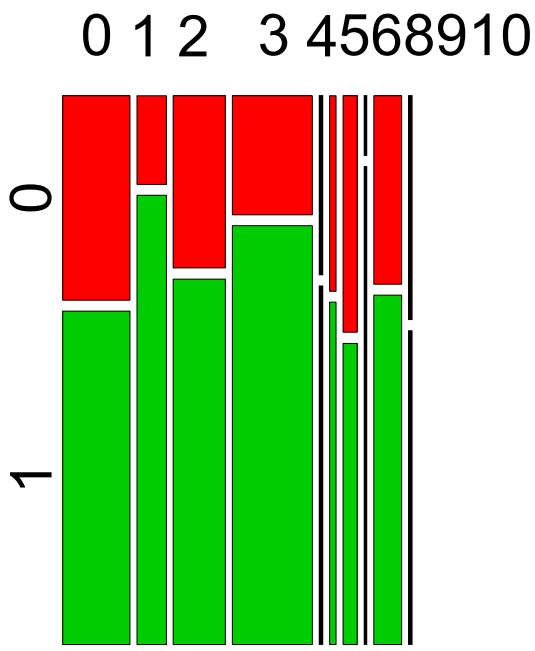

Figure 4 . Mosaic Plot of X_4: Purpose

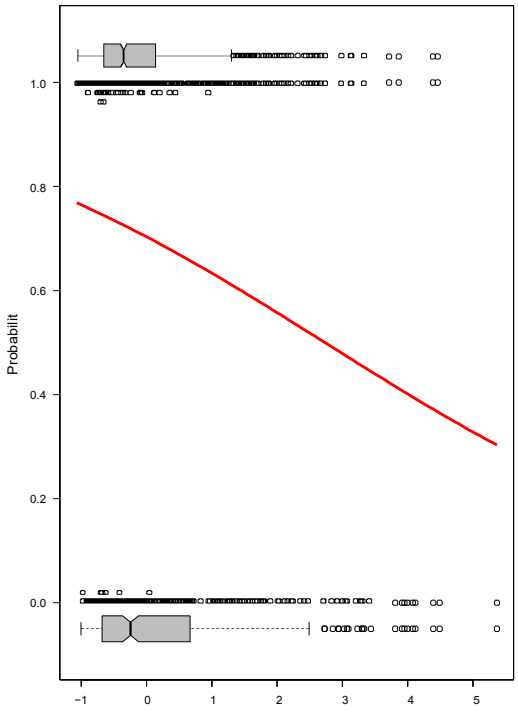

Figure 5 . Logistic Regression Plot of X_5: Loan Amount

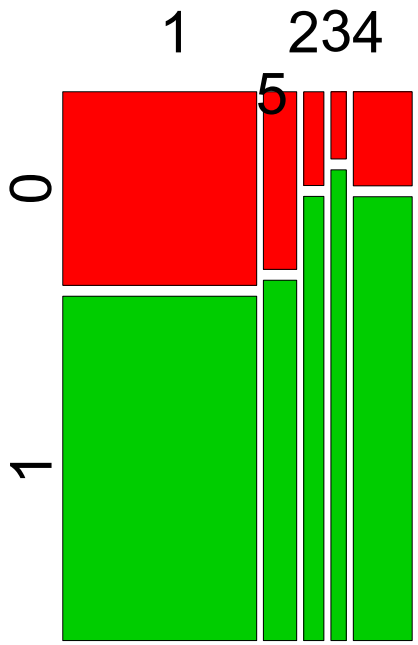

Figure 6 . Mosaic Plot of X_6: Value Savings Stocks

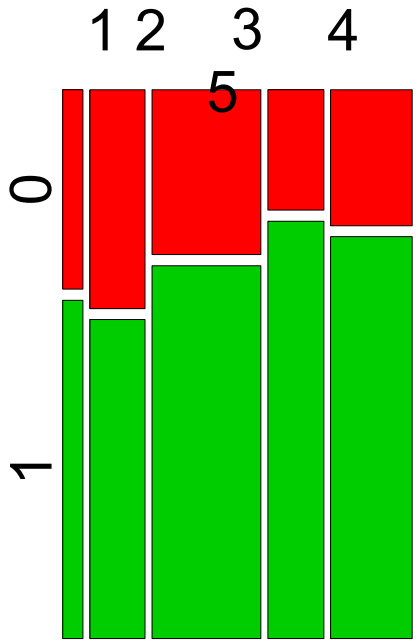

Figure 7 . Mosaic Plot of X_7: Length of Current Employment

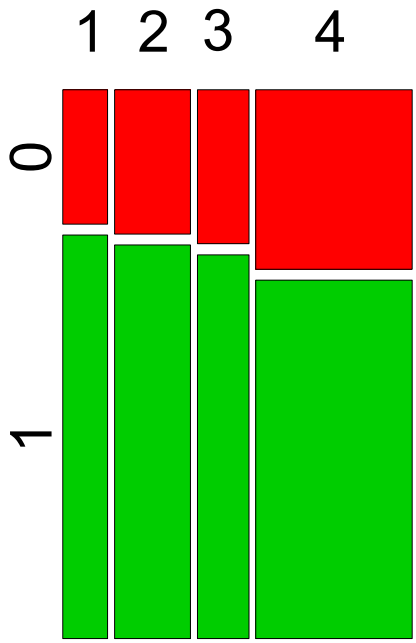

Figure 8 . Mosaic Plot of X_8: Installment percent

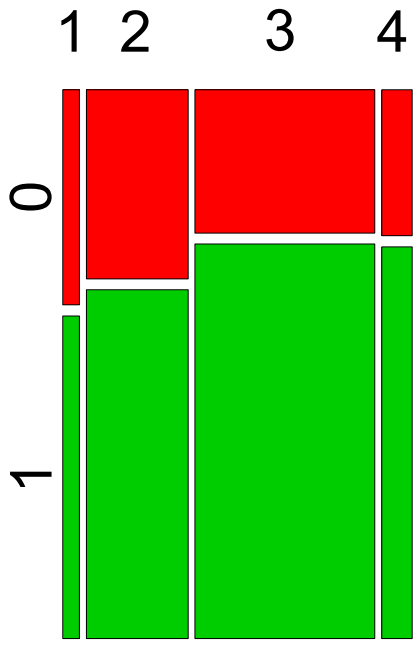

Figure 9 . Mosaic Plot of X_9: Sex and Martital Status

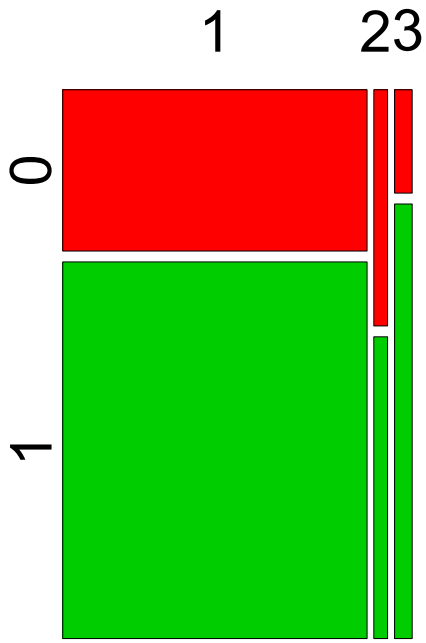

Figure 10 . Mosaic Plot of X_{10}: Guarantors

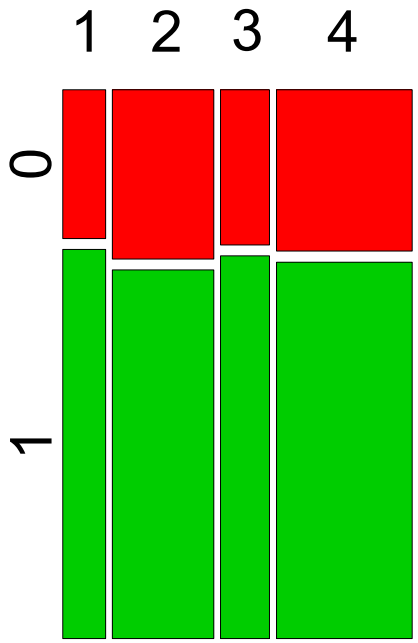

Figure 11 . Mosaic Plot of X_{11}: Duration in Current Address

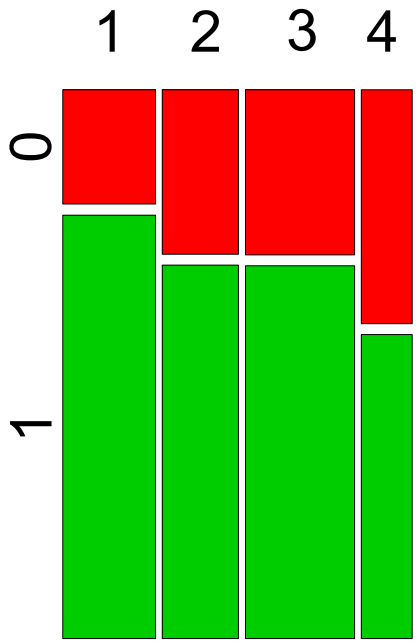

Figure 12 . Mosaic Plot of X_{12}: Most valuable available asset

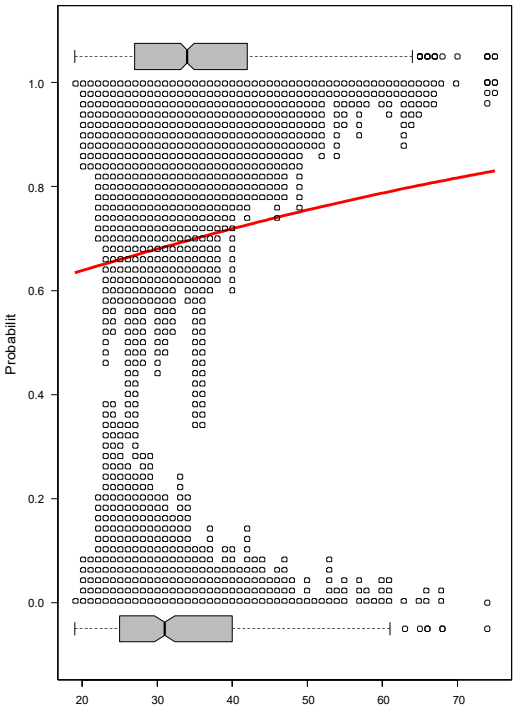

Figure 13 . Logistic Regression Plot of X_{13}: Age in years

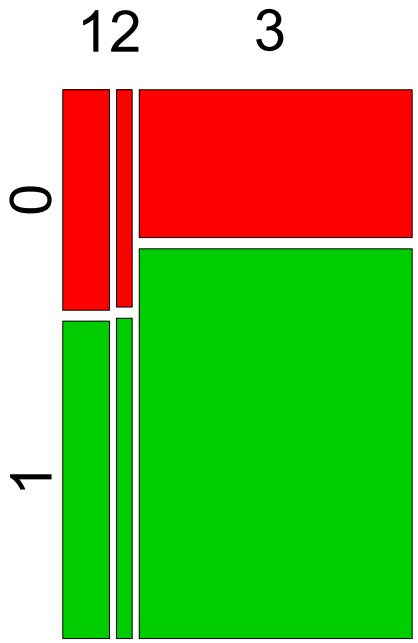

Figure 14 . Mosaic Plot of X_{14}: Concurrent Credits

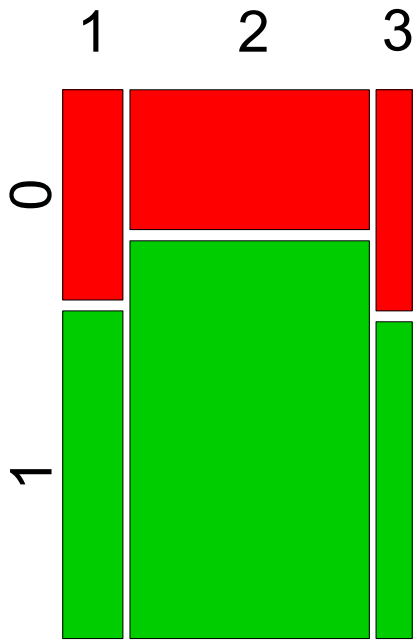

Figure 15 . Mosaic Plot of X_{15}: Type of Apartment

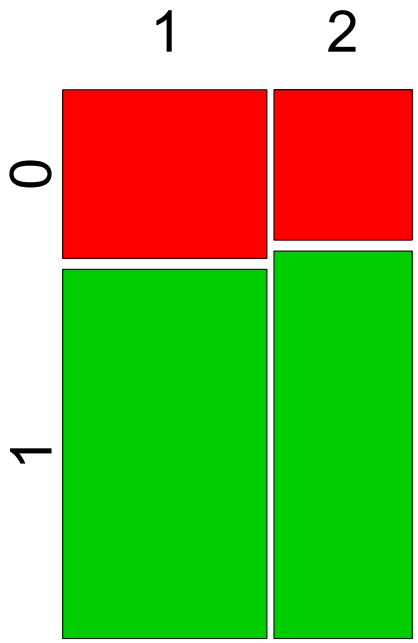

Figure 16 . Mosaic Plot of X_{16}: Number of Previous Credits at this bank

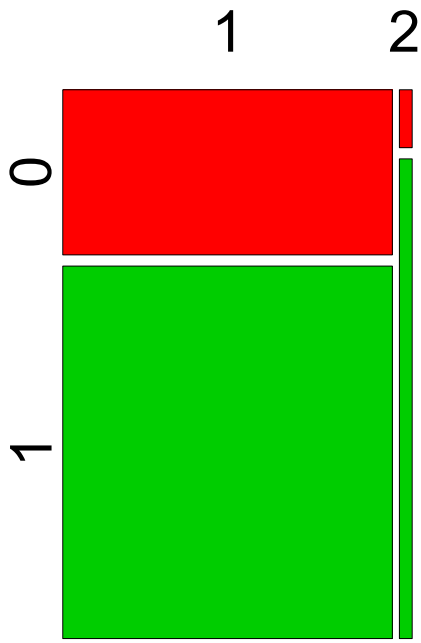

Figure 17 . Mosaic Plot of X_{17}: Occupation

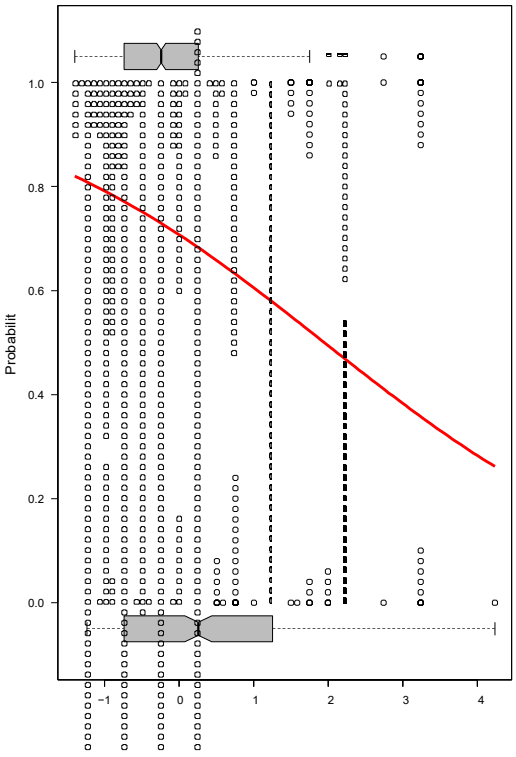

Figure 18 . Mosaic Plot of X_{18}: Number of Dependents

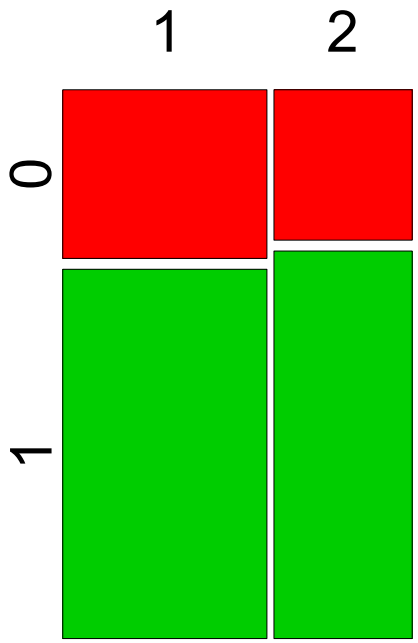

Figure 19 . Mosaic Plot of X_{19}: Telephone

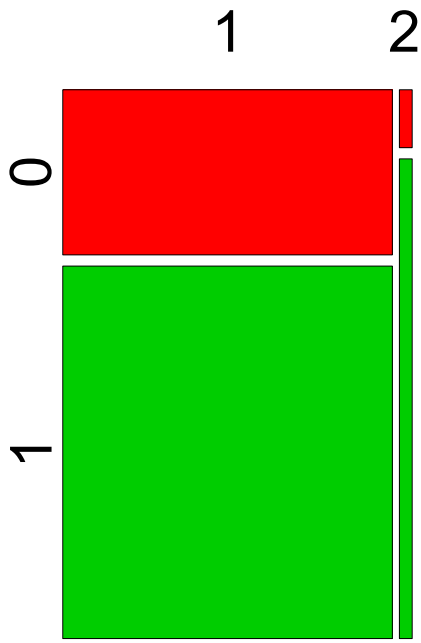

Figure 20 . Mosaic Plot of X_{20}: Foreign Worker

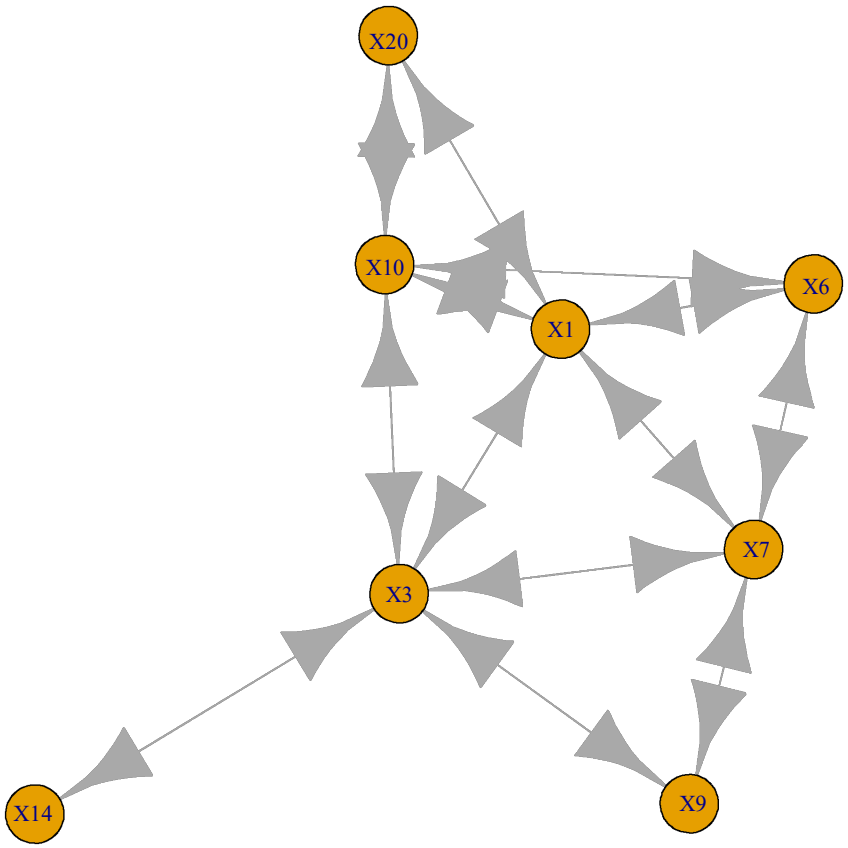

Figure 21 . Adjacency plot of dependency between 8 categorical variable

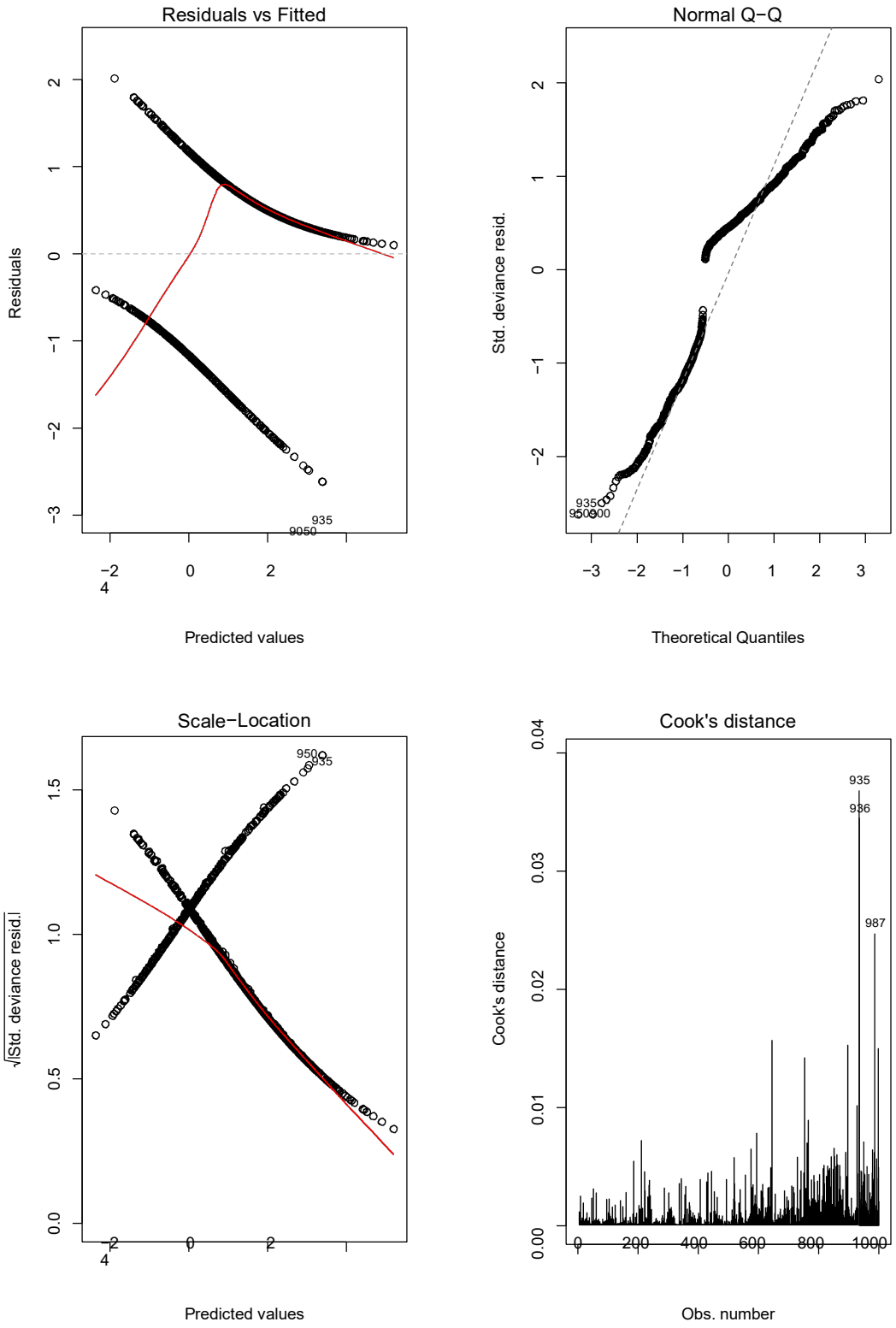

Figure 22 . Diagnostic Plot of Fitted Logistic Model