Evaluation of Customer Default Risk in Tianchi Financial risk

Chuhan Su

Abstract:

Based on the PCA(principal component analysis) and logistic regression model, this essay evaluates the default Risk of the borrowers' information in the Tianchi Financial Risk dataset. The research finds that the default rate is the main factor affecting Tianchi Financial Risk. Combining borrowers' credit grades with factors influencing the default rate, the logistic regression analysis is conducted. It is concluded that individuals with a step above D have a high risk of default, whereas those with a grade below D have low-risk defaults.

Keywords: default risk, PCA, logistic regression model

1. Introduction

With the development of the economy and the substantial improvement of the per capita level, people are committed to pursuing a higher quality of life. In this context, more individuals will choose loans to meet their short-term financial demands. Commercial banks, as one of the most essential sources for individuals seeking loans, play a vital role in assessing the credit risk of borrowers.

Financial risk assessment based on personal loans is quite common, where banks process and analyze data by integrating customers' profiles and financial transactions [1]. In this paper, based on the loan and credit profile information of different borrowers, the most significant factor influencing financial risk is the default risk of borrowers. Through PCA, logistic regression on default risk combines the primary elements and the user's credit grades. Determine whether the user falls into the category of low-risk default or high-risk default.

2. Methodology

2.1 Data selection and processing

Using the Tianchi Financial Risk database available on the Kaggle website, we analyzed the 'testA.csv' in it as our dataset. This dataset comprises people's loan information, including details such as loan amount, loan term, interest rates, repayment amount, etc. And personal credit profile information like employment title. Employment length and property ownership status, etc. Through data preprocessing and analysis, we aim to address financial risk assessment.

During data processing, all dimensions(factors) are processed in numeric values, excluding non-numeric influences such as 'grade,' 'subgrade, employment length,' and others. This will enable us to create a new dataset for subsequent analysis.

2.2 Principal Components Analysis

PCA (Principal component analysis) is a technique for simplifying complex data through dimensionality reduction. In other words, it helps identify the data's most important patterns and trends. PCA aims to find a new set of variables known as PC(principal component), which summarize the most critical information in the original dataset. These main components were selected to explain as much of the variation in the original data as possible.

Since the presence of numerous influencing factors affecting financial risk in this database, using PCA can help us identify r (where r<n) new variables that capture the primary characteristics of the phenomena. At the same time, it can also ensure that each new variable is a linear combination of the original variables, reflecting the comprehensive effect of the actual variables.

After principal components analysis in R studio, we obtained 40 main component factors. Combining the standard deviation of each component, we calculated the variance explained by each central part (as shown in Figure 1) and calculated the cumulative proportion of the overall conflict. By dividing each variance by the total variance interpreted by all 40 PCs, we can determine the proportion of variance for each PC (as illustrated in Figure 2).

As can be seen from Figure 1, it is evident that PC1 has the most significant impact on financial risk, followed by decreasing effects for subsequent components. Simultaneously, as shown in Figure 2, with more choices of principal components, the impact analysis of the influence on financial risk becomes more comprehensive.

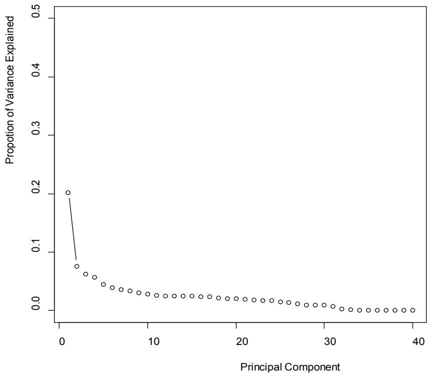

Figure 1

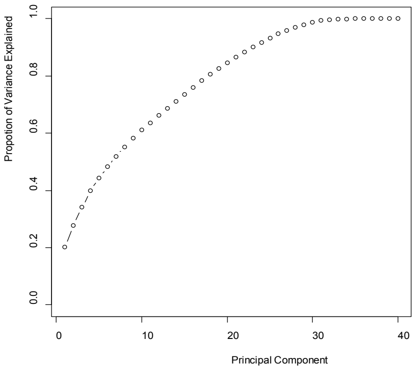

Figure 2

Furthermore, in conjunction with the rotation display in the PCA, the top three factors that have the most significant impact on PC1 are as follows: 'openAcc' (0.3037555620), 'totalAcc' (0.2412781951) and 'loanAmnt' (0.1043420366). 'openAcc' represents the number of outstanding credit lines in the borrower's credit profile, 'total ACC' represents the current credit limit in the borrower's credit profile, and 'loanAmnt' signifies the number of loan.

The classified information of borrowers has varying impacts on the default rate of borrowers. The objective and formal aspects of borrowers' information can be further categorized into borrowers' characteristic and economic information[2]. Considering the rotation data, borrower characteristics information('openAcc' and 'totalAcc'), and borrower financial information ('loanAmnt'), these three primary influencing factors correlate with PC1. This suggests that PC1 may represent the default rate of borrowers, with a higher PC1 indicating higher risk in Tianchi financial risk.

2.3 Logistic regression model on default risk

Through principle components analysis, we have learned that the default rate (PC1) has the most significant impact on Tianchi's financial risk. In addition, since there is a linear combination relationship between the new variables and the original variables in PCA, the primary factors influencing PC1 are 'openAcc,' 'totalAcc,' and 'loanAmnt.' Along with the customer's credit grade assigned by the bank, we determine whether the user falls into the category of a low-risk default or a high-risk default.

Firstly, we need to perform a simple retrieval and processing of the data. We extracted 'grade,' 'loanAmnt,' 'openAcc,' and 'totalAcc' from the entire 'testA' dataset. We scaled 'loanAmnt' (i.e., changed the unit to '000) and converted the variable' Grade. ' Convert the letter grades(A-F) into numerical values(1-7). Assuming that individuals with grade numbers greater than 4 (i.e., grades above 'D') are classified as high default risk (default=1), while the rest are considered low default risk (default=0).

Therefore, we adopt the binary Logistics regression model to investigate the risk of loan default associated with multiple variables. The specific regression model is as follows.

$$g\left(\delta_i\right) = Ln\frac{P\left(Y=1\right)}{P\left(Y=0\right)} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$

Where P (Y=1) represents high-risk default, P (Y=0) represents low-risk default, β_0 is the intercept, β_1 is the coefficient for x_{1i}('loanAmnt'), β_2 is the coefficient for x_{2i} ('openAcc'), β_3 is the coefficient for x_{3i}('totalACC'). When applied to the test dataset, this yields the coefficients for the Logistic Regression model.

$$g\left(\delta_i\right) = Ln\frac{P\left(Y=1\right)}{P\left(Y=0\right)} = -2.801 + 0.045x_{1i} + 0.027x_{2i} - 0.015x_{3i}$$

Since the coefficients for 'loanAmnt' and 'openAcc' are positive, it indicates a positive correlation with default risk. This means that as the number of loans and open credit lines increases, the log odds of high-risk defaults compared to low-risk ruins also increase. On the contrary, 'totalAcc' is negatively correlated. As the number of total accounts credit lines increases, the log odds of high-risk defaults compared to low-risk ruins decrease.

2.4 Training the Logistic Regression Model and Model Accuracy

After successfully establishing the Logistic Regression model, it is essential to test the model and convert the default probabilities into categories (i.e., high-risk default and low-risk default) and then set a certain threshold to measure the model's accuracy.

The whole test A data was divided into training set and test set, in which 70% of the data is used for training the model, and the remaining test data is used to make predictions with our established Logistic Regression model. Since the regression model is a binary logistic regression model, we set the threshold at 0.5.

The accuracy function was used to calculate the accuracy of the Logistic Regression model and the threshold-based categorization; the obtained accuracy rate is 89.96%. This demonstrates that the Logistic Regression model holds significance, and it is reasonable to take the user’s credit rating of ‘D’ (the converted number is 4) as the boundary between high-risk and low-risk default.

3. Results & Discussion

Through the establishment of PCA and logistic regression models, this paper demonstrates the significance of the default rate in financial risk assessment. At the same time, combining the main factors affecting default rate and the classification of user credit rating given by banks, we can conclude that users with a grade of D or higher are more likely to pose an increased risk of default. In contrast, those with a degree of below D belong to low-risk default.

4. Limitation & Improve

The analysis method and model adopted in this paper allow for the classification of default risk within the Tianchi Financial Risk dataset; due to the absence of specific values of borrowers’ default rate, we relied solely on PCA to infer that the default rate is the most important influencing factor. Therefore, the logistic regression model cannot accurately quantify the default rate but only divides the default risk according to the users’ level.

Different banks have their risk assessment mechanisms, as well as their systems to measure default rates. Combining the essential characteristics that affect customer credit risk, a pre-loan risk scoring model based on Logistic is established, and the regression results are converted into a scorecard through scoring scaling, which enhances the model’s intuitiveness and interpretability. [3] This approach facilitates the quantified analysis of financial risk.

5. Program Code

```
> library(ggplot2)
> tianchi <- read.csv(“C:\\Tianchi Financial Risk\\testA.csv”,
+                    header = TRUE, sep = “,”)
> numeric_cols <- apply(Bianchi, is.numeric)
> tianchi1 <- tianchi[, numeric_cols]
> non_constant_cols <- apply(tianchi1, 2, function(col) any(col != col[1]))
> tianchi1 <- tianchi1[, non_constant_cols]
> tianchi1 <- na.omit(tianchi1)
> if (col(tianchi1) == 0) {stop(“No valid columns remain for PCA after preprocessing.”)}
> pca_result <- prcomp(tianchi1, scale = TRUE)
> summary(pca_result)
> pca_result$rotation
> # Plot cumulative variance explained
> plot(cumsum(pca_result$sdev^2) / sum(pca_result$sdev^2),
+     xlab = “Number of Principal Components”,
+     ylab = “Cumulative Proportion of Variance Explained”,
+     type = “b”)
> tianchi_2 <- tianchi[,c(6,2,21,26)]
> attach(tianchi_2)
> tianchi_2[,2] <- loanAmnt/1000
> tianchi_data <- data.frame(tianchi_2)
> letter_mapping <- setNames(1:26, LETTERS)
> tianchi_data[,1] <- letter_mapping[tianchi_data$grade]
> tianchi_data$default <- ifelse(tianchi_data$grade > 4, 1, 0)
> set.seed(123)
> sample_indices <- sample(1:nrow(tianchi_data), n * 0.7)
> train_data <- tianchi_data[sample_indices, ]
> test_data <- tianchi_data[-sample_indices, ]
> tianchi_data_glm <- glm(default ~ loanAmnt + openAcc + totalAcc, data = train_data, family = binomial(link=logit))
> summary(tianchi_data_glm)
> predictions <- predict(tianchi_data_glm, new data = test_data, type = “response”)
> predicted_classes <- ifelse(predictions > 0.5, 1, 0)
> accuracy <- sum(predicted_classes == test_data$default) / length(test_data$default)
> cat(“accuracy:”,accuracy,”\n”)
```

References:

[1] Zhang Ruizhi, Yang Guowei and Xu Quan, a classification audit model of bank loan risk based on self-coding clustering algorithm. Audit Observation, 2022(03): pp. 77-81.

[2]Gu Huiying and Yao Zheng, A study on the Influencing Factors of Borrower Default Risk in P2P online lending Platform -- A case study of WDW. Shanghai Economic Research, 2015(11): 37-46.

[3] Li Xianhang, Machine Learning-based Explainable Credit Risk Scoring and Default Prediction, 2022, Southwestern University of Finance and Economics. Page 61.