Mid-Term Data Analysis Project Template - Report

Vinh Do

10/23/2022

# 1. Introduction.

The concept of credit can be considered a trust that allows an organization to provide assets or money to other parties. These parties then have to pay back at a later date. A scoring system can identify an organization’s credit rating in “bad” or “good”. In this scenario, the company avoids doing business with “bad credit” customers who present a greater risk to work with. To achieve this goal, The company is eager to understand the reliable factors that can indicate a “bad” or “good” credit customer. As a result, the company can apply it as a risk management strategy before entering into any contracts with upcoming customers to reduce unexpected expenses.

The aim of this analysis is to analyze and present the risks posed by “bad” credit accounts, thereby enabling the company to avoid engaging in high-risk transactions with these customers. To accomplish this, we analyzed the German credit data set with 1000 observations. Statistical methods were also used to pinpoint six factors that may contribute to a “bad” credit rating.

Six factors will be applied in the analysis, including employment status, age, education levels, total credit amount owned, and credit duration. All the methods used, analysis and results have been discussed and shown in the data and exploration section. Finally, the conclusion section will include insights and recommendations that may apply to company risk management to avoid”bad” credit transactions in the future.

# 2. Data and Exploratory Analysis

This section is a detailed analysis of the factors discussed in the introduction, supported by data extracted from the employee data set.

## 2.1 The effect of the employment year

### \* QUESTION: Is there a relationship between employment years and credit rating processing?

### Data - Table 3 and 4 | Figure 1

The variables RESPONSE and EMPLOYMENT from the dataset (cd) were used to respond to this question. The employment factor (EMPLOYMENT) is a factor type with five minor variables representing experience and total employment years, including less than 1 year of employment, 1-3 years of employment, 4-6 years of employment, 7 or more years of employment, and unemployment. On the other hand, RESPONSE shows the account rating with “bad” or “good” values.

### Methods

To visualize this, two summary tables were used to compare the two variables’ percentages. In addition, a pie chart was used to visualize the two summary tables and assist in analysing.

### Analysis

There are five variables across bad and good credit ratings in the EMPLOYMENT variable. According to Tables 3 and 4, an employee with less than 1-year of employment has an 8% higher rate of having a bad credit situation. In addition, People with more than 4 years of work usually have a higher rate of owning a good credit account. The percentage difference is 12%, and those with four or more years of experience having bad credit account for 34%. In contrast, people with four or more years of experience having good credit account for 46%. The contrast between the “red” and “green” colours of the two pie graphs in figure 1 indicates the result more clearly.

### Results

Based on the analysis, there is a relationship between Job Levels and credit rating. Accounts with a long employment history are more likely to have a good credit rating.

## 2.2 Effect off Education

### \* QUESTION: Is there a relationship between the account’s credit rating and the education level? Do people with higher education have better credit ratings?

### Data - Table 1

Data from table 1 shows the number and percentage of education on two credit rating factors that assist in answer the question. The data is organized into four sections, including “bad credit rating with no education,” “bad credit rating with no education,” “good credit rating with no education,” and“good credit rating with no education.” The distribution of each section will be counted and displayed in the summary table.

### Methods

The method used in the study was a table summary of the number of credit rating accounts (counts) and the educational level of the applicants.

### Analysis

In the summary (table 1), education factors are outlined, including “yes” and “no”. By examining this summary, we can see there is an imbalance between two elements in education. Moreover, only 28 of the 50 educated people (56%) got a good credit rating. In contrast, 71% of less educated people get good credit ratings. The total gap is 15%.

### Results

The analysis helps conclude that education and credit rating has a correlation. People with higher education stand a higher chance of having a bad credit rating. As a result, the null hypothesis is rejected in this case.

## 2.3 Effect of Saving Account

### \* QUESTION: Is there a relationship between the Saving Account and the credit rating?

### Data - Table 2

The saving account can be a representation of assets and the ability to repay credits. So, to see the relationship between the Saving Account and the credit rating, the RESPONSE, and SAV\_ACCT variables are combined into a new dataset for calculating the total number and analysis.

### Methods

The method used is a table summary of the number (counts) of credit rating accounts and saving accounts. In the table, the saving account number for each section and percentage is calculated by summarising and mutating the function.

### Analysis

Table 2 shows that people with poor credit ratings have less money in their savings accounts. 72% of people with bad credit accounts hold less than 100DM in their savings accounts compared to 52% from good credit accounts.

### Results

The saving account contributes to credit rating decisions based on the two characteristics above. People with more money will be less likely to fall into bad credit situations.

## 2.4 The effect of age on the credit level

### \* QUESTION: Is there a relationship between age and the credit score level? Do older people tend to have good credit ratings?

### Data - Figure 2

The information, including age and credit situation, is needed to visualize and analyze this question. The AGE variable store numerical values that show the age of each account owner. Therefore, AGE and RESPONSE have been used to show the difference between older and younger consumers and their credit rating.

### Methods

The method used is a box plot visualization representing the spread of data, the standard deviation, and the median age and credit rating. There are two boxes. The bad credit rating spread is shown on one and has a red colour, and the good credit rating spread is on the other side with green colour.

### Analysis

Figure 2 illustrates different contributions with bad and good credit ratings based on age. The interquartile range is higher for good credit rating observations that start at 27 and end around 42. Compared to a bad credit score, it begins at about 24 and ends at 40. Moreover, the average good credit rating consumer is 4 years older than a bad credit rating consumer.

### Results

Based on the analysis, credit rating is significantly impacted by age. Older people tend to have a stable income source and manage their finances better. Thus, we can accept the hypothesis that older people tend to have better credit ratings.

## 2.5 Effect of Duration

### \* QUESTION: Is there a relationship between the duration and the credit score?

### Data - Figure 3

In figure 3, the violin plot type has been used to answer the Hypothesis. The Horizontal (x) axis is the rating rate value from RESPONSE. Additionally, DURATION variables are vertical value axes. DURATION stores the length of time a person needs to pay for the loan.

### Methods

A violin plot is one of the great plots to compare a numerical variable and a categorical variable. The plot will show the distribution of the duration of credit in months by two types of credit ratings,

### Analysis

The range of a bad credit rating is wider, starting from 7 months up to 70 months. In contrast, a good credit rating consumer often pays fully between 5 to 60 months. So, accounts with long-term credit debt have a higher chance of being bad credit rating accounts.

### Results

Finally, the plot demonstrates the relationship between bad and good credit ratings by the spread of the distribution. Overall, a bad credit rating usually has a longer duration of months.

## 2.6 The effect of Instalment Rates on the credit level

### \* QUESTION: Is there a relationship between the instalment rate and the credit rating? Does bad credit tend to have higher interest rates?

### Data - Figure 4

A high instalment rate can also negatively affect people when the monthly payment exceeds their income. Two variables are a good fit to answer this question INSTALL\_RATE and RESPONSE. These variables were combined to visualize a bar graph that INSTALL\_RATE is a horizontal axis (x) and the bar represents the “Yes” and “No” observations from RESPONSE.

### Methods

For this purpose, a bar plot with a full-fill position will be used; this will allow the comparison of two credit ratings to be more efficient and the audience to see the trend more clearly.

### Analysis

The bar chart visualization (figure 4) shows a clear negative relationship between credit rating and the instalment rate. The good credit rating steadily decreases as the interest rate rises. Good credit ratings have the largest proportion in the 1% instalment rate columns. In the last column (4% instalment rate), the reduction of good credit accounts is apparent compared to 1% or 2% column.

### Results

The analysis shows a strong correlation between credit rating and the instalment rate. The lower the instalment rate is, the higher the chance of a customer having a good credit rating.

# III. Conclusion

Through analysis and report sections, the German Credit dataset offers the company 6 insights into identifying bad and good credit accounts. This will enable them to determine the best solution to avoid or resolve these instances. Specifically, all variables have been used to identify the relationship to credit rating. The instalment rate, amount of credit, duration, and education negatively correlate with “good” credit ratings. Alternatively, there is a positive relationship between “good” credit ratings and age, employment years, and money in a savings account.

In order to identify a “good” credit customer and avoid risks associated with “bad” credit transactions, the following factors must be considered as the company’s strategy.

* To limit transactions involving “bad” credit accounts, the company must target customers with **Long Employment Years** with stable incomes. People with **4 years of employment or more are a top choice**.
* Moreover, the age range for good credit ratings starts from **27 to 44**. Thus, **Older People** will be safer when providing the service.
* **Short duration** is another factor to consider. Customers with good credit often choose **payment plans of 12 or 24 months**.
* A **lower instalment rate** is another factor that determines a good credit rating. The company should target customers with an **instalment rate of 1% to 2%**.
* Higher-educated people have a higher chance of accumulating bad credit accounts. the company should be **careful to engage in all transactions regardless of education background**.
* The credit rating can be impacted by having a **High-Savings Account**. To avoid “bad” credit, the company must engage consumers holding **more than 100 German Deutsche Mark in a savings account**.

In conclusion, the company can use the above factors as a guide to avoid risk transactions with “bad” credit customers and decrease unexpected losses by deciding potential and target customers based on the above factors.

# IV. Appendix/Appendices.

## Code

### table 1 - The Number and Proportion of Credit Rating by Education

| EDUCATION | RESPONSE | N | Percentage |
| --- | --- | --- | --- |
| No | Bad Credit Rating | 278 | 29.263 |
| No | Good Credit Rating | 672 | 70.737 |
| Yes | Bad Credit Rating | 22 | 44.000 |
| Yes | Good Credit Rating | 28 | 56.000 |

### table 2 The Number and Proportion of Saving Account by Credit Rating

| RESPONSE | SAV\_ACCT | savingacct | Percentage |
| --- | --- | --- | --- |
| Bad Credit Rating | <100 DM | 217 | 72.333 |
| Bad Credit Rating | >1000 DM | 6 | 2.000 |
| Bad Credit Rating | 101−500 DM | 34 | 11.333 |
| Bad Credit Rating | 501−1000 DM | 11 | 3.667 |
| Bad Credit Rating | No saving account | 32 | 10.667 |
| Good Credit Rating | <100 DM | 386 | 55.143 |
| Good Credit Rating | >1000 DM | 42 | 6.000 |
| Good Credit Rating | 101−500 DM | 69 | 9.857 |
| Good Credit Rating | 501−1000 DM | 52 | 7.429 |
| Good Credit Rating | No saving account | 151 | 21.571 |

### table 3 - Bad Credit Rating Account by Employment Years

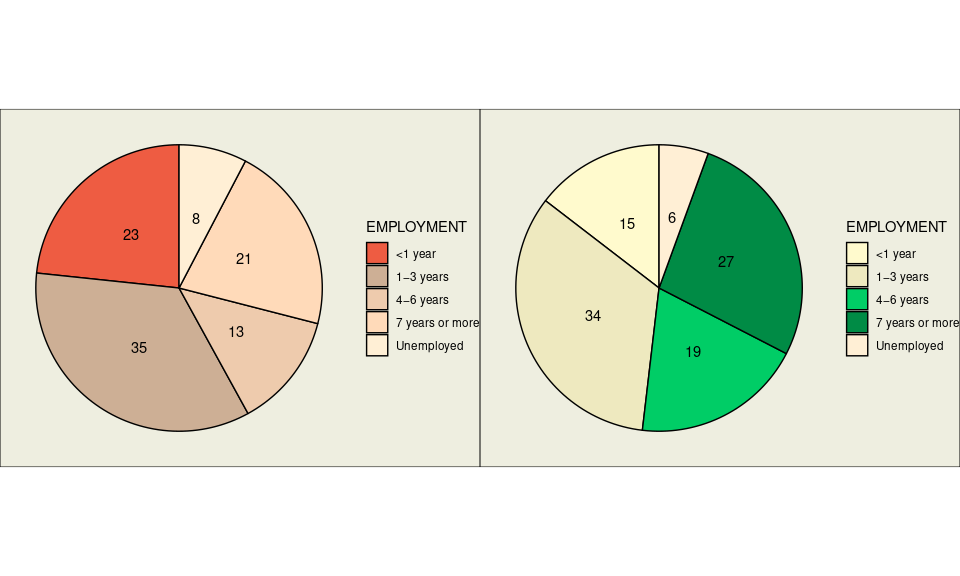
| RESPONSE | EMPLOYMENT | Number | Percentage |
| --- | --- | --- | --- |
| Bad Credit Rating | <1 year | 70 | 23.333 |
| Bad Credit Rating | 1−3 years | 104 | 34.667 |
| Bad Credit Rating | 4−6 years | 39 | 13.000 |
| Bad Credit Rating | 7 years or more | 64 | 21.333 |
| Bad Credit Rating | Unemployed | 23 | 7.667 |

### table 4 - Good Credit Rating Account by Employment Years

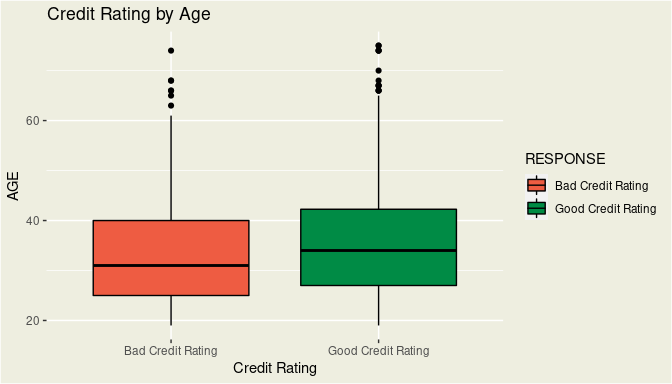
| RESPONSE | EMPLOYMENT | Number | Percentage |
| --- | --- | --- | --- |
| Good Credit Rating | <1 year | 102 | 14.571 |
| Good Credit Rating | 1−3 years | 235 | 33.571 |
| Good Credit Rating | 4−6 years | 135 | 19.286 |
| Good Credit Rating | 7 years or more | 189 | 27.000 |
| Good Credit Rating | Unemployed | 39 | 5.571 |

## Including Plots

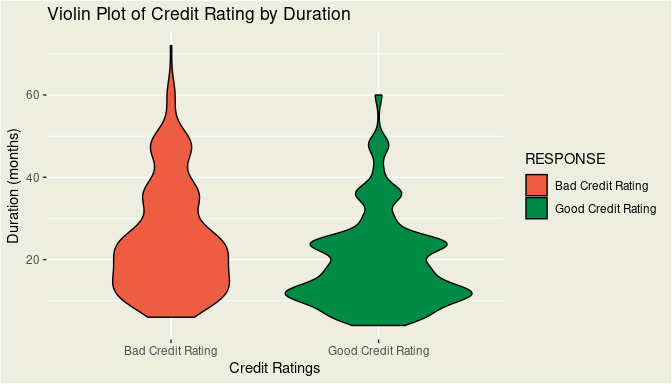
### Figure 1 - Bad Credit Rating Account and Good Credit Rating Account by Employment Years



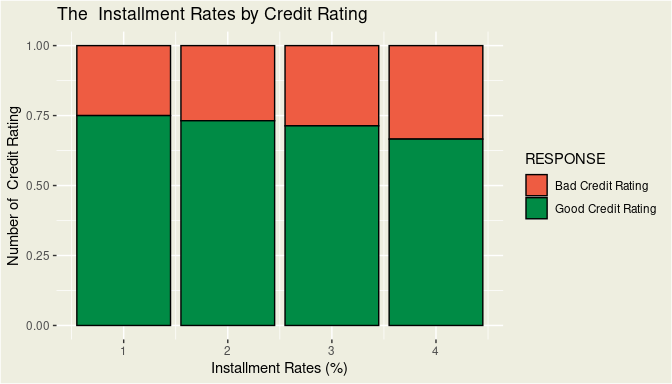
### Figure 2 - The Total Credit Rating by Age



### Figure 3 - Violin Plot of Credit Rating by Duration



### Figure 4 - The Credit Rating Accounts by Installment Rate



— End of Instructions —