Final Project Report - Credit Scoring Model Using Weight of Evidence and Logistic Regression

Siheng Huang, Yicong Li, Yunhao Li

April 30, 2023

Abstract

This report presents a credit scoring model developed using the Weight of Evidence (WOE) and logistic regression techniques. The model assesses the credit risk of individuals by evaluating their financial and personal characteristics and assigning scores. We demonstrate the effectiveness of this approach in creating an efficient credit risk assessment tool.

1 Introduction

Credit scoring is a critical aspect of the financial industry, where lending institutions need to assess the risk associated with providing loans to borrowers. The problem we aim to address is creating a model that can accurately evaluate the credit risk of individuals based on their financial and personal characteristics. An effective credit scoring model will help lending institutions make informed decisions, minimize the risk of default, and optimize their lending portfolios.

2 Materials and Methods

To develop a credit scoring model, we used a dataset (Credit Fusion, Will Cukierski, 2011) containing various financial and personal characteristics of individuals. Our methodology involved the Weight of Evidence (WOE) method for variable transformation and logistic regression for model building and scorecard creation.

2.1 Data Description

The dataset consists of several financial and personal characteristics of individuals, such as age, monthly income, debt ratio, the number of times an individual has been late on payments (30-59 days, 60-89 days, and 90+ days), the number of real estate loans, and the number of dependents. The target variable is the credit risk, which indicates whether an individual is at risk of defaulting on their loans (See Appendix 5.1 Table 1).

2.2 Data Cleaning

The first step in our analysis was to clean the data by removing duplicates, handling missing values, and addressing outliers. We used various techniques such as imputation and truncation to ensure the data quality and consistency.

Missing value analysis revealed that the MonthlyIncome column had 29,731 missing values, and the NumberOfDependents column had 3,924 missing values. Since the proportion of missing values was high, directly removing them would have resulted in a significant loss of observations, which is not the most suitable method. We used the regression method to fill in the missing values.

The regression method considers the relationships between variables in the dataset, providing more accurate estimates for the missing values compared to other methods, such as mean or median imputation. By leveraging the existing relationships between variables, the regression method can better preserve the overall structure of the data.

Outlier analysis discovered that the age variable included a value of 0, which is an evident outlier that needed to be removed from the dataset. Otherwise, retaining this value could compromise the accuracy and reliability of our credit scoring model. By eliminating this outlier, we enhanced the dataset quality and minimized the likelihood of skewed results in our subsequent analysis and modeling.

2.3 Exploratory Data Analysis

After cleaning the data, we conducted an exploratory data analysis (EDA) to understand the data's distribution, identify relationships between variables, and detect potential patterns. This process involved creating histograms, density plots, scatter plots, and box plots to visualize the data and computing summary statistics and correlation coefficients to gain insights into the relationships between variables.

We focused on the variable distributions of age and monthly income, which are key factors in determining credit risk. Their distributions provided insights into the overall characteristics of the borrowers in the dataset. In Figure 1, the plots suggested that the age variable and the monthly income variable in the dataset appear to follow a normal distribution, which is desirable for statistical analysis.

Figure 2 shows that the correlation between variables is minimal. Logistic regression requires checking for multicollinearity issues, but in this case, since the correlation between variables is small, it can be preliminarily concluded that there is no multicollinearity problem. Of course, after modeling, we can still use the VIF (variance inflation factor) to check for multicollinearity problems. If multicollinearity exists, meaning that two variables may be highly correlated, dimensionality reduction or removal may be required to address the issue and improve the model's performance.

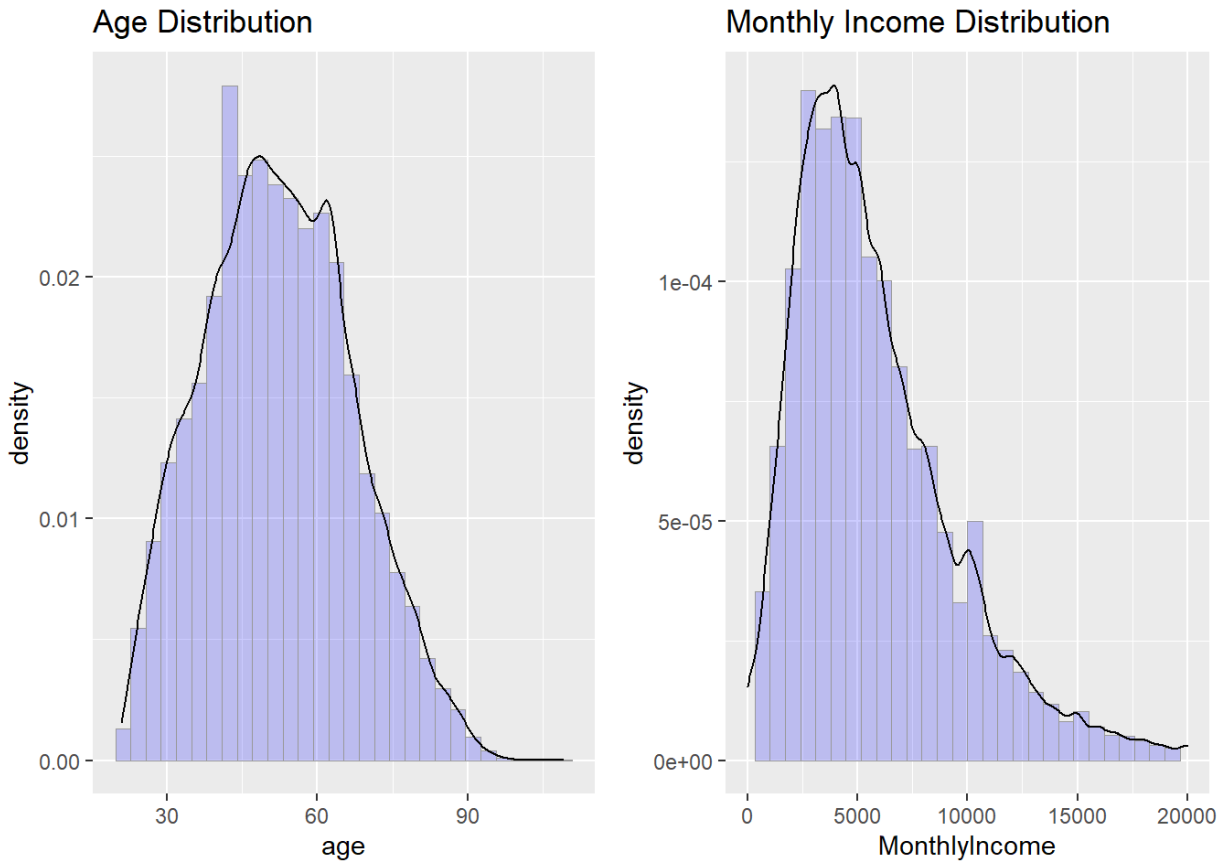

Figure 1: Univariate Analysis: Variable Distribution

2.4 Logistic Regression

Logistic regression is a fundamental component in the development of credit scoring cards. With the integration of Weight of Evidence (WOE) transformation applied to independent variables, the outcomes of logistic regression can be directly translated into a summary table, known as the standard scoring card format.

Initially, we performed a p-value test for variable selection. Two variables, NumberOfOpenCreditLinesAndLoans and RevolvingUtilizationOfUnsecuredLines, did not pass the p-value test. After removing these variables, all remaining variables in the second regression model passed the test, and the Akaike Information Criterion (AIC) value decreased, indicating a better model fit.

We evaluated the model's performance using ROC analysis. As shown in Figure 3, the AUC value of 0.692 signifies a fairly high level of accuracy.

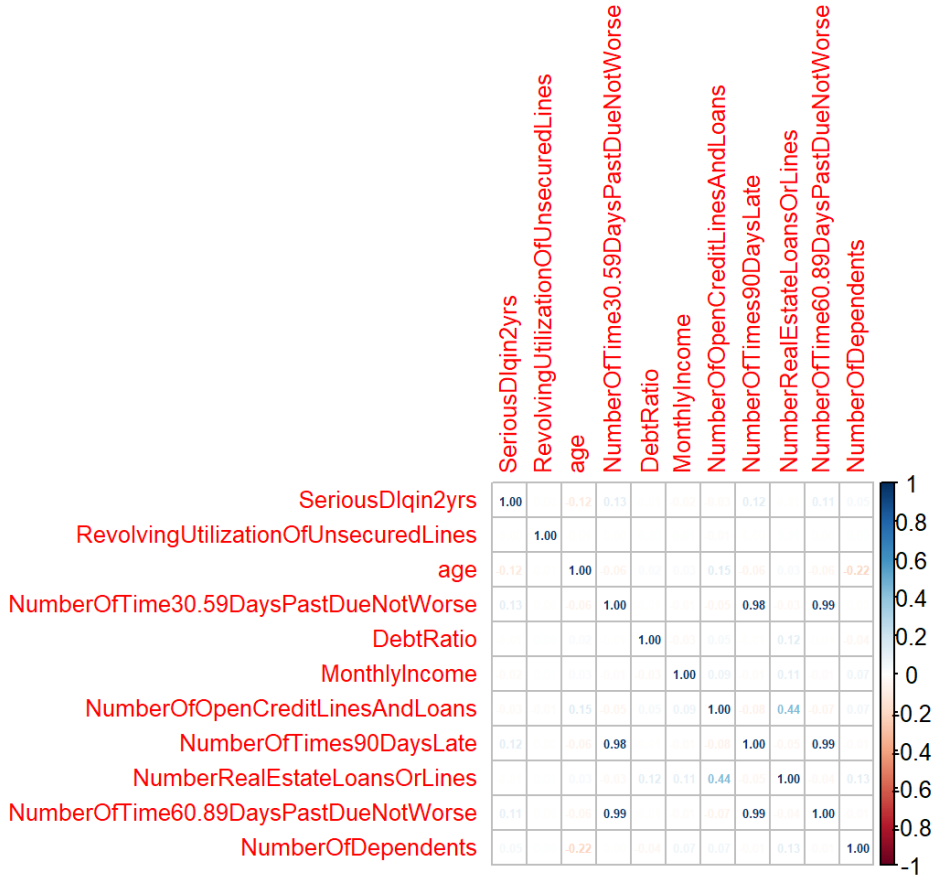

Figure 2: Multivariate Analysis: Variable Correlation

2.5 Weight of Evidence Approach

To better understand the relationship between each variable and credit risk, we employed the Weight of Evidence (WOE) approach. WOE is a statistical measure used primarily in the context of credit scoring and predictive modeling. This technique transforms categorical and continuous variables into a more informative representation for use in logistic regression models. The main goal of WOE is to establish a relationship between a predictor variable and a binary outcome (e.g., good credit risk vs. bad credit risk) (See Appendix 5.2).

The implementation of the WOE method involved the following steps:

1. Binning: Divide each variable into several bins based on their values. For continuous variables, we used equal-width, equal-frequency, or custom binning techniques (See Figure 4).

2. Calculation: For each bin, calculate the WOE value using the formula:

$$\text{WOE} = \log\left(\frac{\text{defaults\%}}{\text{non-defaults\%}}\right) \tag{1}$$

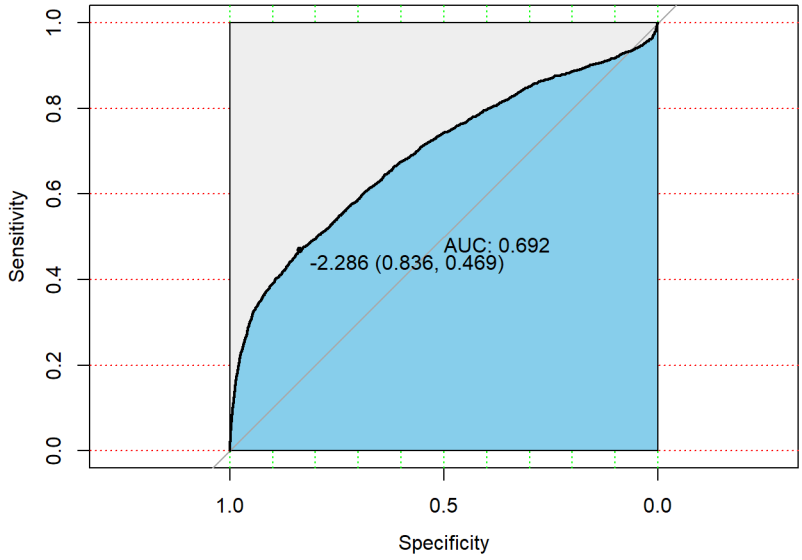

Figure 3: ROC Evaluation

3. Transformation: Replace the original values of each variable with their corresponding WOE values.

4. Model Development: Use the transformed variables to build a logistic regression model, which predicts the probability of an individual being at credit risk.

By using the WOE approach, we can better understand the relationship between each variable and credit risk, which helps improve the accuracy and interpretability of the logistic regression model and the resulting scorecard.

2.6 Scorecard Creation

The process of creating a scorecard involves assigning scores to each bin of the transformed variables based on the logistic regression model's coefficients. The scorecard translates the logistic regression model's output into an easily interpretable format that helps lending institutions assess an individual's credit risk. Here's a detailed description of the scorecard creation process:

1. Determine the Points-to-Double (P) and Base Points (Q): First, we need to determine the number of points required to double the odds (P) and the base points (Q). In this case, we assume that a bad-to-good ratio of 15 corresponds to 600 points, and for every additional 20 points, the bad-to-good ratio is halved. These values are then used to calculate P and Q using the following formulas:

$$P = -\frac{20}{\log(2)} \tag{2}$$

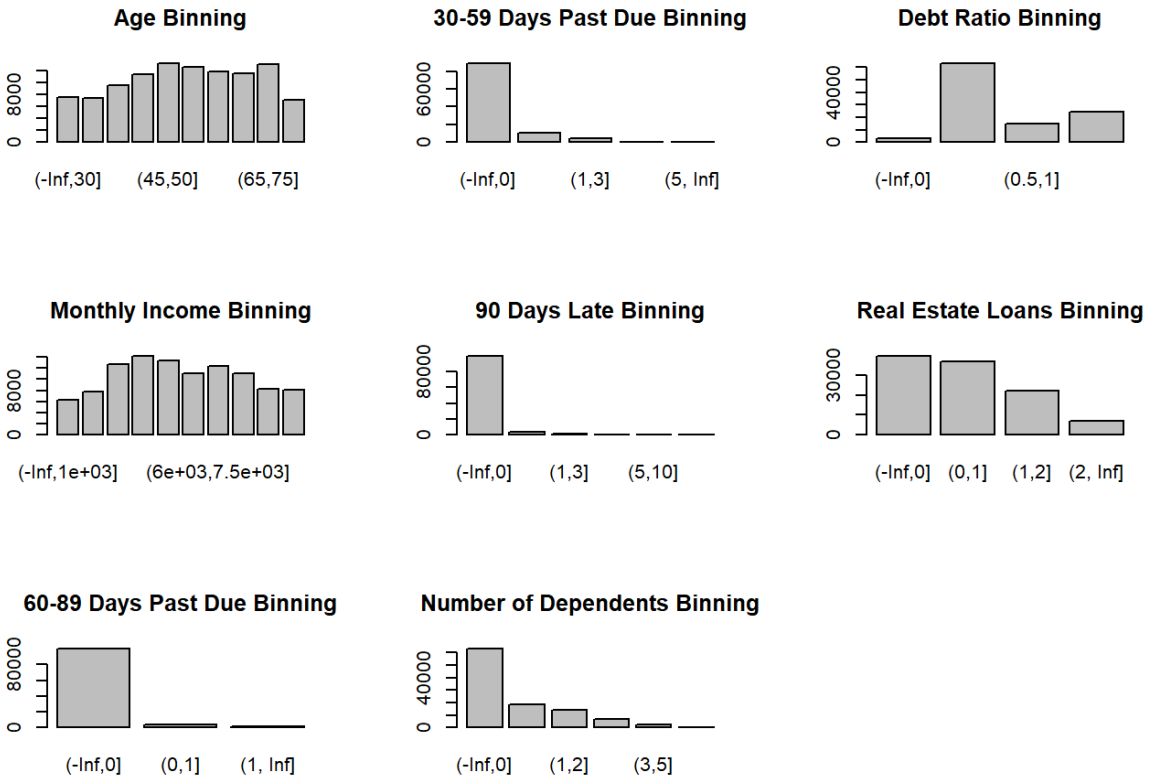

Figure 4: Binning

$$Q = 600 + \frac{20 \cdot \log(15)}{\log(2)} \tag{3}$$

2. Calculate Scores for Each Variable: Next, we calculate the score for each transformed variable using the logistic regression model's coefficients and the P and Q values obtained in step 1. The formula to calculate the score for each variable is:

$$Score = Q + P \cdot \log(odds) \tag{4}$$

   Here, the odds are calculated using the logistic regression model's coefficients and the WOE-transformed variables.

3. Calculate Component Scores: For each variable, we compute the component scores by multiplying the logistic regression model's coefficients with the corresponding WOE values and the P value. The component scores represent the contribution of each variable to the overall credit score.

4. Compute Scores for Each Bin: For each bin of the transformed variables, we calculate the bin scores by applying the logistic regression model's coefficients and the WOE

values. This is done using a custom function that takes the index of the variable and the WOE value as inputs and returns the rounded score.

5. Create the Scorecard: Finally, we create the scorecard by organizing the calculated scores for each bin of each variable in a tabular format. The scorecard displays the scores for each bin of the transformed variables, making it easy for lending institutions to evaluate an individual's credit risk based on their financial and personal characteristics.

3 Results

The analysis revealed that the WOE transformation effectively captures the relationship between the input variables and the target variable, making it suitable for logistic regression. The logistic regression model was able to identify the contribution of each variable to credit risk and assign appropriate scores. The resulting scorecard (See Figure 5) can be used to evaluate the credit risk of individuals based on their characteristics.

age	range	**<=30**	**[30,35)**	**[35,40)**	**[40,45)**	**[45,50)**	**[50,55)**	**[55,60)**	**[60,65)**	**[65,75)**	**>=75**
	score	207	172	117	94	57	18	-85	-165	-343	-428
NumberOfTime30.59DaysPastDueNotWorse	range	**<=0**	**[0,1)**	**[1,3)**	**[3,5)**	**>=5**					
	score	-178	299	577	744	934					
DebtRatio	range	**<=0**	**[0,0.5)**	**[0.5,1)**	**>=1**						
	score	130	-88	301	-25						
MonthlyIncome	range	**<=1000**	**[1000,2000)**	**[2000,3000)**	**[3000,4000)**	**[4000,5000)**	**[5000,6000)**	**[6000,7500)**	**[7500,9500)**	**[9500,12000)**	**>=12000**
	score	72	114	68	41	2	-15	-43	-82	-143	-117
NumberOfTimes90DaysLate	range	**<=0**	**[0,1)**	**[1,3)**	**[3,5)**	**[5,10)**	**>=10**				
	score	-134	683	954	1157	1266	1023				
NumberRealEstateLoansOrLines	range	**<=0**	**[0,1)**	**[1,2)**	**>=2**						
	score	108	-119	-81	115						
NumberOfTime60.89DaysPastDueNotWorse	range	**<=0**	**[0,1)**	**>=1**							
	score	-45	293	437							
NumberOfDependents	range	**<=0**	**[0,1)**	**[1,2)**	**[2,3)**	**[3,5)**	**>=5**				
	score	-38	20	56	71	120	203				

Figure 5: Score Table

4 Discussion

The credit scoring model developed using WOE and logistic regression provides a valuable tool for assessing credit risk. Lending institutions can use this scorecard to make informed decisions about granting loans to borrowers. By identifying high-risk borrowers, they can minimize the risk of default and optimize their lending portfolios.

Future work could involve validating this model on a larger dataset, incorporating additional features that might improve its predictive capabilities. Additionally, other machine learning techniques could be explored to compare their performance with the logistic regression model. This will enable further refinement of the credit scoring model, ultimately leading to a more robust and accurate tool for credit risk assessment.

5 Appendix

5.1 Variable descriptions and types

Variable Name	**Description**	**Type**
SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse	Y/N
RevolvingUtilizationOfUnsecured-Lines	Total balance on credit cards and personal lines of credit except real estate and no installment debt	percentage
age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years	integer
DebtRatio	Monthly debt payments, alimony, living costs divided by monthy gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAnd-Loans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

Table 1: Variable descriptions and types

5.2 Weight of Evidence

In order to demonstrate how the WOE transformation helps establish the relationship between predictor variables and the binary outcome. We present the following proof that the log odds of the target variable can be represented as a linear combination of the WOEs of the predictor variables, with the coefficients being the weights for the respective WOEs. This transformation makes the logistic regression model more easily interpretable and aids in feature selection and model evaluation.

Define the Default (D) as "1" and Non-default (N) as "0". The WOE for a specific bin is calculated as the log of the ratio between the percentage of defaults and non-defaults within that bin.

$$\text{WOE(bin)} = \log\left(\frac{D\%}{N\%}\right) = \log\left(\frac{\#D_{\text{bin}}/\#D}{\#N_{\text{bin}}/\#N}\right) \tag{5}$$

As an example, we consider two age bins, age_1 and age_2. The difference between their WOEs can be interpreted as the log of the odds ratio between the two age groups.

$$\begin{aligned}
\text{WOE}(\text{age}_1) - \text{WOE}(\text{age}_2) &= \log\left(\frac{\#D_{\text{age}_1}/\#D}{\#N_{\text{age}_1}/\#N} \Big/ \frac{\#D_{\text{age}_2}/\#D}{\#N_{\text{age}_2}/\#N}\right) \\
&= \log\left(\frac{\#D_{\text{age}_1}}{\#N_{\text{age}_1}} \Big/ \frac{\#D_{\text{age}_2}}{\#N_{\text{age}_2}}\right) \\
&= \log\left(\text{odds ratio between age}_1 \text{ and age}_2\right)
\end{aligned}$$

For any two bins, bin_1 and bin_2, the difference between their WOEs can be calculated as the difference between the log odds of the two bins.

$$\text{WOE}(\text{bin}_1) - \text{WOE}(\text{bin}_2) = \log\left(\frac{P(Y=1|X \in \text{bin}_1)}{P(Y=0|X \in \text{bin}_2)}\right) - \log\left(\frac{P(Y=1|X \in \text{bin}_2)}{P(Y=0|X \in \text{bin}_2)}\right) \tag{6}$$

Since we are using WOE-transformed predictor variables, the log odds of the target variable can be represented as a linear combination of the WOEs of the predictor variables, with the coefficients being the weights for the respective WOEs.

$$\log\left(\text{odds}\right) = \log\left(P(Y=1)/P(Y=0)\right) = \beta^T \text{WOE} = \beta_1 \text{WOE}_1 + \cdots + \beta_n \text{WOE}_n \tag{7}$$

The log odds ratio for two samples can also be expressed as a linear combination of the log odds ratios for the predictor variables, weighted by their respective coefficients.

$$\begin{aligned}
\log\left(\text{odds ratio for 2 samples}\right) = &\beta_1 \log\left(\text{odds ratio of feature 1 for 2 samples}\right) + \ldots \\
&+ \beta_n \log\left(\text{odds ratio of feature n for 2 samples}\right)
\end{aligned}$$

Thus add weights for WOEs of features. A higher coefficient for a specific feature indicates a stronger relationship between that feature and the log odds of the target variable.

6 Reference

1. Credit Fusion, Will Cukierski. (2011, October). Give Me Some Credit, Version 1.0. April 1, 2023 from `https://kaggle.com/competitions/GiveMeSomeCredit`.