BUS 3320 Term Project

Credit Card Default Payment Prediction Using Multiple Linear Regression

Riley Sweeting

2024-04-25

Contents

Introduction

Credit card payment defaults pose significant harm for both banks and credit card holders. A default happens when a cardholder neglects to meet the minimum required payment on their card. This failure usually leads to the suspension of the user's card, a decrease in their credit score, and in extreme cases, legal action from the bank to recover the debt. This can be severely costly for the bank, which highlights the importance and necessity of predicting and preventing credit card defaults.

Research Question

The goal of this project is to determine if the credit limit of credit card holders influences the probability of defaults on their credit card payments.

Hypotheses

It is hypothesized that the credit limit of credit card holders negatively influences the probability of defaulting. In other words, as the credit limit of the card holder increases, the likelihood of them defaulting decreases. The null hypothesis states that there is either no significant relationship between the credit limit of card holders and the probability of defaulting, or that the relationship is positive.

Hypotheses:

$H_0 : \beta_{Limit} \geq 0$
$H_a : \beta_{Limit} < 0$

Dataset Description

The dataset used is sourced from the University of California, Irvine's Machine Learning Repository, and contains 6 months of credit information of Taiwanese credit card holders. The time period spans from April 2005 to September 2005. This repository is open to the public, free of charge, and accessible to everyone. The dataset is titled "Default of Credit Card Clients", and was published in 2016.

The dataset contains 30,000 samples, each with 23 features, some of which are aggregated in the analysis. All samples are free of missing values, but some samples appear to contain missing values which will be detailed later. The features used in the regression are as follows:

- `Limit` - The quantitative measure of the credit limit of the card holder.
- `Sex` - The categorical gender of the card holder.
  - 0 → Male
  - 1 → Female
- `Education` - The categorical education status of the card holder.
  - 1 → Graduate School
  - 2 → University
  - 3 → Highschool
  - 4 → None/Other
- `Marriage` - The categorical marriage status of the card holder.
  - 1 → Married
  - 2 → Single
  - 3 → Other
- `Age` - The quantitative measure of the age of the card holder.
- `Average Bill` - The quantitative calculation of the average bill of the card holder over the 6 months.
- `Average Payment` - The quantitative calculation of the average monthly payment of the card holder over the 6 months.
- `Consistence` - The categorical representation of the monthly payment status of the card holder over the 6 months.
  - 0 → Payed duly every month
  - 1 → Did not pay duly every month
- `Missed Payment` - The categorical representation of whether the card holder failed to a payment anytime during the 6 months.
  - 0 → Did not miss a payment
  - 1 → Missed a payment

Data Preprocessing

Given that the dataset contains categorical data, it is essential to correctly format and process the data so that it is properly interpreted as categorical data during regression. Failure to format the data can lead to inaccurate results. This section details the pre-processing and formatting methods used.

Import Excel Data

Import the Excel file containing the data from its location. Limit the number of rows to the number of samples, being 30,000.

```
# Open package
library(readxl)

# Import data from Excel
df <- read_excel("C:/Users/riley/OneDrive/Documents/School/College/Semester 4/Business
↪  Stats/Assignments/Final/credit_data.xlsx", n_max = 30000)
```

Filter Abnormal Values

Some of the samples in the dataset have feature values outside of the defined range. For example, the categorical feature `Education` ranges from 1 to 4, representing graduate, university, highschool, and other education. However, the value 0 appears in the dataset, so samples with an `Education` value of 0 are removed. The same applies to the feature `Marriage`.

```
# Filter and eliminate samples
df = df[df$Education != "0", ]
df = df[df$Marriage != "0", ]
```

What is Factoring?

Categorical features of the dataframe must be factored, or in other words, interpreted as nominal data. If not factored, the regression function will interpret the categorical data as quantitative data, and the results of the regression will be inaccurate. For example, if an unfactored feature `Color` has values $red, blue, green$ represented by $0, 1, 2$, the regression function will interpret the values as measured numbers rather than categories.

Factor Categorical Data

Factor the categorical features of the dataset, and if needed, change the levels of the features.

```
# Factor categorical features so they are interpreted as nominal data
df$Sex = factor(df$Sex, levels = c("1", "2"), labels = c("0", "1")) # Change levels to 0
↪  and 1
df$Education = factor(df$Education)
df$Marriage = factor(df$Marriage)
df$Consistence = factor(df$Consistence)
df$Missed_Payment = factor(df$Missed_Payment)
df$Default = factor(df$Default)
```

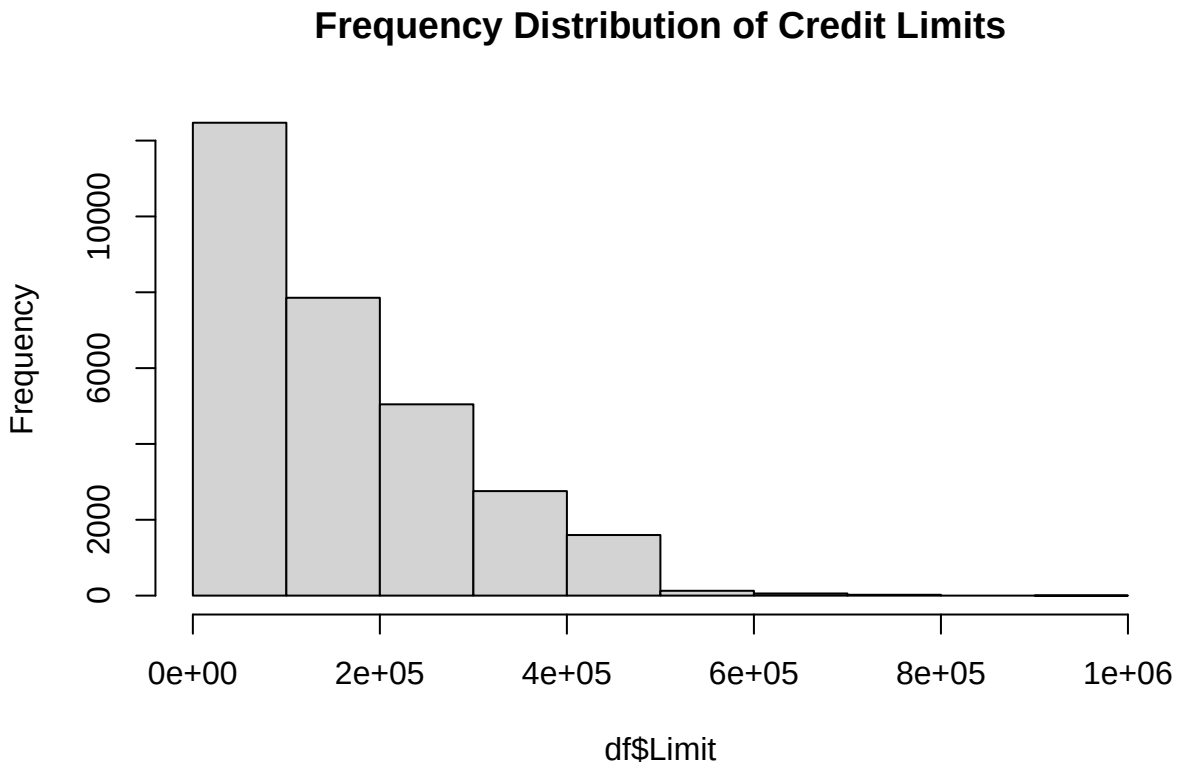
Summary of Data

In this section, we will analyze the descriptive statistics of the independent and dependent variables used in the simple and multiple logistic regressions.

Summary of Independent Credit Limit

From the descriptive summary statistics below, we see that the distribution of credit limits is positively skewed (to the right), indicating a tendency towards lower values. 25% of card holders have a credit limit in the $50,000 bracket, and 50% in the $140,000 bracket. This makes sense given the context of the data, as the majority of bank users start with a lower limit, and then once the bank deems them trustworthy and responsible are they given higher limits.

```
# Plot a histogram of credit limit data
hist(df$Limit, breaks = 10, main = "Frequency Distribution of Credit Limits")
```



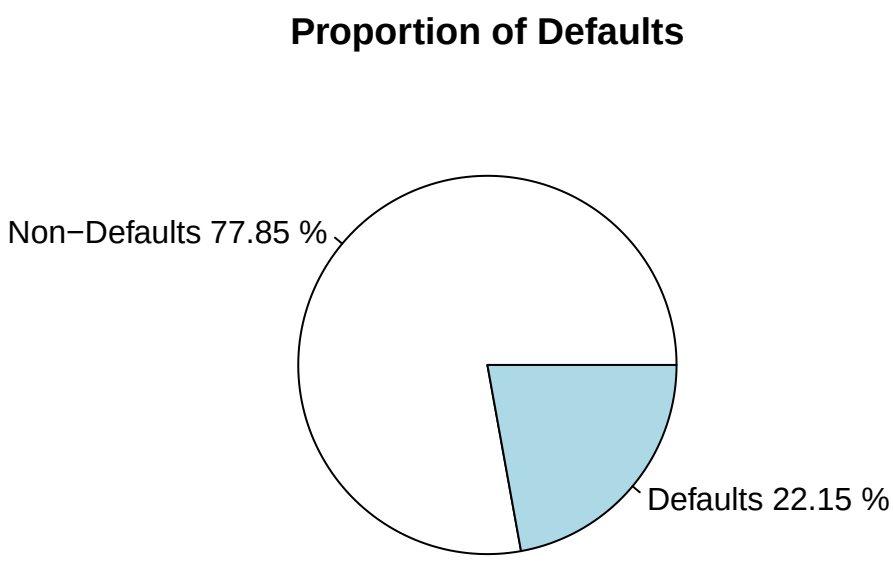

```
# Descriptive summary statistics
summary(df$Limit)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10000   50000  140000  167523  240000 1000000
```

Summary of Dependent Default

From the descriptive summary statistics below, we see that close to a quarter of all card holders defaulted on a payment, 6,631 out of 29,932 to be exact (22.15%). This is higher than expected, meaning it is possible that the creators of the dataset introduced extra samples with defaults into the dataset, meaning that the proportions of the data may not reflect the true distribution.

```
# Plot a pie chart of default data
pie(table(df$Default), labels = c(paste("Non-Defaults", round(table(df$Default)[1] /
↪  length(df$Limit) * 100, 2), "%", sep = " "), paste("Defaults",
↪  round(table(df$Default)[2] / length(df$Limit) * 100, 2), "%", sep = " ")), main =
↪  "Proportion of Defaults")
```




```
# Descriptive summary statistics
summary(df$Default)
```

```
##     0     1
## 23301  6631
```

Compute Regressions

In this section, we will perform simple and multiple logistic regression to determine the regression coefficients of the card holder features. These coefficients will be used in hypothesis testing to evaluate the research question.

Perform Simple Logistic Regression

The research question aims to test whether the credit limit of card holders influences the probability of defaulting. In this simple regression, we are using the feature `Credit Limit` as the sole independent variable, and the feature `Default` as the dependent variable.

```
# Open MFX Package (For Logistic Regression)
suppressMessages(library(mfx))

# Compute simple logistic regression using Logitmfx
simple_model = logitmfx(Default ~ Limit, data = df)
```

```r
# Display Model
simple_model
```

```
## Call:
## logitmfx(formula = Default ~ Limit, data = df)
##
## Marginal Effects:
##             dF/dx   Std. Err.       z     P>|z|
## Limit -5.5365e-07  2.0488e-08 -27.023 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Perform Multiple Logistic Regression

Similar to the simple logistic model, we have `Default` as the dependent variable and `Credit Limit` as the main independent variable. However, we are now including other features of the dataset so we can control for these other features.

```r
# Open MFX Package (For Logistic Regression)
suppressMessages(library(mfx))

# Compute multiple logistic regression using Logitmfx
multiple_model = logitmfx(Default ~ Limit + Sex + Education + Marriage + Age + Avg_Bill +
↪  Avg_Payment + Consistence + Missed_Payment, data = df)

# Display Model
multiple_model
```

```
## Call:
## logitmfx(formula = Default ~ Limit + Sex + Education + Marriage +
##     Age + Avg_Bill + Avg_Payment + Consistence + Missed_Payment,
##     data = df)
##
## Marginal Effects:
##                       dF/dx   Std. Err.        z     P>|z|
## Limit           -4.8126e-07  2.3565e-08 -20.4228 < 2.2e-16 ***
## Sex1            -2.6158e-02  4.7362e-03  -5.5230 3.333e-08 ***
## Education2       6.7802e-03  5.3928e-03   1.2573   0.20866
## Education3       3.6093e-03  7.2453e-03   0.4982   0.61837
## Education4      -1.3299e-01  1.1309e-02 -11.7597 < 2.2e-16 ***
## Marriage2       -3.0354e-02  5.2800e-03  -5.7489 8.982e-09 ***
## Marriage3       -7.8707e-03  2.0230e-02  -0.3891   0.69722
## Age              5.6462e-04  2.8151e-04   2.0057   0.04489 *
## Avg_Bill         8.7272e-07  4.3544e-08  20.0422 < 2.2e-16 ***
## Avg_Payment     -5.7912e-06  5.1069e-07 -11.3400 < 2.2e-16 ***
## Consistence1     3.8029e-05  1.0970e-02   0.0035   0.99723
## Missed_Payment1  1.6498e-01  5.0046e-03  32.9659 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "Sex1"            "Education2"      "Education3"      "Education4"
## [5] "Marriage2"       "Marriage3"       "Consistence1"    "Missed_Payment1"
```

Interpret Results

In this section, we will perform hypothesis testing on the regression coefficients of the `Credit Limit` feature resulting from the simple and multiple logistic regressions. The `logitmfx()` function was used to allow for easier interpretation of the categorical dependent variable `Default`.

Simple Regression Coefficient

The coefficient of the quantitative independent variable `Credit Limit` was determined to be -5.5365×10^{-7}. The negative coefficient signifies a negative relationship between the credit limit of card holders and the likelihood they default on their payments. More specifically, per \$1,000 the credit limit increases, the probability of defaulting decreases by 0.00055365 or 0.055 percent on average. These results are statistically significant at the 0.1% significance level, meaning the p-value corresponding to the regression coefficient is less than 0.001.

Multiple Regression Coefficient

The coefficient of the quantitative independent variable `Credit Limit` was determined to be -4.8126×10^{-7}. The negative coefficient signifies a negative relationship between the credit limit of card holders and the likelihood they default on their payments. More specifically, per \$1,000 the credit limit increases, the probability of defaulting decreases by 0.00048126 or 0.048 percent on average, holding all other features constant. These results are statistically significant at the 0.1% significance level, meaning the p-value corresponding to the regression coefficient is less than 0.001.

Conclusion

In both regression results, the p-value corresponding to the regression coefficient was less than the significance level of 0.001, making the relationship hypothesized in the alternative hypothesis highly statistically significant. Thus, we reject the null hypothesis and must accept the alternative hypothesis.

The null hypothesis states that there is either no significant relationship between the credit limit of card holders and the probability of defaulting, or that the relationship is positive. In other words, the regression coefficient of the credit limit of card holders is either 0 or a positive value. Since we rejected the null hypothesis and accepted the alternative hypothesis, we accept that the coefficient is negative, and that there is a significant negative relationship between the 2 variables.

The coefficient is closer to the coefficient of $\beta_{Limit} = -4.8126 \times 10^{-7}$ determined by the multiple logistic regression. The coefficient of the multiple regression is more reflective of the relationship than the simple regression, because the multiple regression takes other features into account. Thus, when we say that the results are statistically significant, we are saying that the credit limit coefficient is statistically less than the value 0 hypothesized in the null hypothesis.