Predicting Default Probability on Credit Card Payments

Pamela De Vera

20/12/2020

Abstract

Knowing a borrower's credit risk is important to financial institutions as it can help determine how much of a premium to charge clients. The data was retrieved from the *UCI Machine Learning Repository* to develop a logistic model predicting a credit card client's probability of defaulting on a payment. The model predicted this probability based on the client's age, sex, education, their latest bill statement, and the number of months they were past due on their latest payment. A goodness-of-fit test was then performed to see how well the model fit the data. There was no evidence to indicate that the model was a bad fit and all predictors were found to be significant. That is, age and the number of months payment was made past its due date has a positive relationship with the log probability of defaulting. Being male also increases this probability. Furthermore, the level of education and the bill statement amount has a negative relationship with the log probability.

Keywords: Logistic regression, goodness-of-fit, loan default, credit risk

Code and data used in this analysis is found at: https://github.com/pameladv/Predicting-Loan-Default-Risk-from-Personal-Factors-and-Current-Loan-Status-

I. Introduction

When banks decide whether to approve a loan, they look at an individual's income, assets, and debts (Government of Canada, 2019). However, after they've approved the loan, the bank still runs the risk of an individual failing to pay their loan, or defaulting, otherwise known as credit risk (Berk et al., 2018). To accommodate for the risk of default they acquire, banks will often add premiums to loans varying on each loan's default risk (Canada Mortgage and Housing Corporation, n.d.).

Logistic regression is used to predict the logistic probability of a binary response variable using one or more predictor variables, both numerical and categorical (Sheather, 2009). For this paper, I will investigate the relationship between defaulting on a loan with an individual's age, sex, education, their repayment status after the first term, and their bill statement after the first term.

This paper will be using a dataset to determine a possible, valid multilevel model predicting the probability that an individual may have heart disease. In the Methodology section, I will be explaining the components of the dataset and the regression model. Next, the results of the model and diagnostics will be laid out in the Results section. Lastly, conclusions, inferences, weaknesses, and next steps will be found in the Discussion section.

II. Methodology

II-i. Data

The data we will use in this analysis is provided by the *UCI Machine Learning Repository* through *kaggle*. It includes the sex, age, education, and marital status of 30,000 credit card clients in Taiwan from 2005, along with the principal amount of the loan made by each client, their previous payments, bill statements, and whether they have defaulted on their next payment. Unfortunately, the method in which the data was collected is unknown, so, we are unable to determine our frame population. Our sample population is then the 29,532 credit card clients remaining after filtering of the data discussed later on in this section. Our target population is all credit cardholders in Taiwan.

We will specifically be looking at a client's age, sex, education level, as well as how late their most recent payment was and their latest bill statement. We have chosen to look at their age as generally, one would expect younger people to have the most default risk from lack of assets and experience. However, it has been shown that younger people typically default on loan repayment the least of any other age group (Debbaut et al., 2013). Also, women tend to perform better in loan repayments than men (Goodman et al., 2016). We also note that for student loans in California, the students with parents of higher levels of education had a lower default risk than those with parents of lower levels of education (Fuinhas et al., 2019), justifies the need to investigate how education can predict default risk.

In the dataset, the delay in payment is listed as values from 1 to 8 for the number of months past the due date and -2 to 0 if the payment was made appropriately. We noticed in the dataset that the values -2 and 0 represent the same value in terms of months past due, the data was mutated such that only one value, 0, represents the payment made on time. For education, only four levels were listed: high school, university, graduate school, and 'other'. Since 'other' could include many different education levels like elementary school or have completed a Doctorate, we will not include any data entries with 'other' as an education level.

Characteristic	**Sample (n=29532)**	**Not Default (n=22929)**	**Default (n=6603)**
Age			
Mean	35.4750779	35.4065158	35.7131607
Median	34	34	34
Sex			
Mean	0.3967899	0.3863666	0.432985
Median	0	0	0
Education			
Mean	1.8080726	1.7876488	1.8789944
Median	2	2	2
Delay on Latest Payment (months)			
Mean	0.3592036	0.1964325	0.9244283
Median	0	0	1
Latest Bill Statement			
Mean	5.0886254×10^4	5.1636357×10^4	4.8281513×10^4
Median	2.22785×10^4	5.1636357×10^4	4.8281513×10^4

Table 1: Baseline Characteristics

Figure 1 displays the distribution of our variables between the clients who have defaulted and the clients who have not. When looking at the latest bill statement, age, and the number of months the last payment was made past the due date, we can see from their corresponding graphs that they are right-skewed. This is also evident in Table 1 because the mean for each is greater than the median. When comparing the distributions in the groups who have and have not defaulted on their next payment, we can see that the age distribution is more spread out for those who have defaulted, while there is a peak in the younger ages for those who did not default. We can also see a similar trend in the number of months payment was made past the due date. Furthermore, when looking at sex, we note that females have a higher proportion of clients who haven't defaulted on their next payment relative to males.

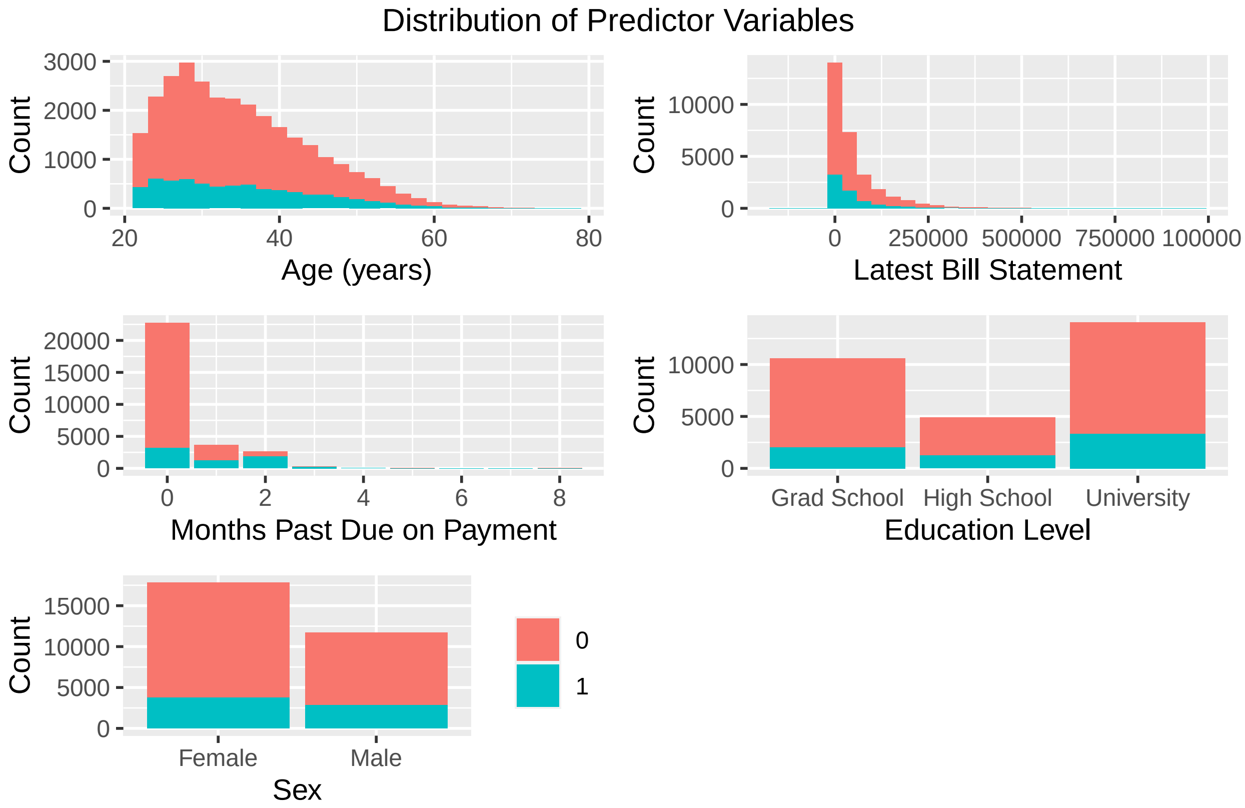

Figure 1: 0 = Not default, 1 = Default

II-ii. Model

For the analysis, we will be performing a logistic regression model with the *lme4* package in R. The logistic model will give us the log probability of our binary response variable (a client defaulting on their next payment), given certain values of our predictor variables. It is expressed in the form of the equation below, where p is the probability of default, β_0 is the log probability of defaulting when all predictor values are 0, and each remaining β_i is the change in log probability when x_i increases by 1, holding all other x's constant. We will be using this model because we will be able to use our predicted probability as a measure of default or credit risk.

$$
\begin{aligned}
log(\frac{p}{1-p}) = \beta_0 + \beta_{\text{age}}x_{\text{age}} + \beta_{\text{male}}x_{\text{male}} + \beta_{\text{highschool}}x_{\text{highschool}} \\
+\beta_{\text{university}}x_{\text{university}} + \beta_{\text{pmt}}x_{\text{pmt}} + \beta_{\text{bill}}x_{\text{bill}}
\end{aligned}
$$

When determining the validity of the model, we will look at the residual deviance of the model with $29532 - 6 - 1 = 29525$ degrees of freedom to perform a goodness-of-fit test. This test will tell us if our model is appropriate for our sample. It has a null hypothesis, H_0, and an alternative hypothesis, H_a, shown below. We calculate the p-value, which is the probability that the true deviance is greater than the residual deviance of the model, taken on a chi-squared distribution of 29525 degrees of freedom, with R. If our calculated p-value is less than 0.05, then we have enough evidence to reject the null hypothesis, indicating our model may not be a good fit (Sheather, 2009).

$$H_0 : \text{logistic model is appropriate}$$

$$H_a : \text{logistic model is not appropriate}$$

III. Results

Coefficient	**Estimate**	**Standard Error**	**z-value**	**p-value**
Intercept	-2.0322055	0.0654335	-31.0575742	$9.015231 \times 10^{-212}$
Age	0.0033822	0.0016935	1.9971598	0.0458078
Sex	0.1581998	0.0311876	5.0725262	3.9256912×10^{-7}
Latest Payment Delay	1.1550559	0.0196441	58.799136	0
Latest Bill Statement	$-9.1881311 \times 10^{-7}$	2.2158067×10^{-7}	-4.1466303	3.3740421×10^{-5}
Education				
Highschool	0.164184	0.0461919	3.5543898	3.7885727×10^{-4}
University	0.1609277	0.034372	4.6819493	2.8415983×10^{-6}

Table 2: Model Summary

In Table 2, the summary values of our model is displayed. So, we can write our model equation for the log probability below.

$$\begin{aligned} log(\frac{p}{1-p}) &= -2.0322055 + 0.0033822x_{\text{age}} + 0.1581998x_{\text{male}} \\ &+0.164184x_{\text{Highschool}} + 0.1609277x_{\text{University}} \\ &+1.1550559x_{\text{pmt}} + -9.1881311 \times 10^{-7}x_{\text{bill}} \end{aligned}$$

From this, we can directly calculate the probability of default based on the predictors as:

$$p = \frac{exp(r)}{1+exp(r)},$$

$$\begin{aligned} \text{where } r &= -2.0322055 + 0.0033822x_{\text{age}} + 0.1581998x_{\text{male}} + 0.164184x_{\text{Highschool}} \\ &+0.1609277x_{\text{University}} + 1.1550559x_{\text{pmt}} + -9.1881311 \times 10^{-7}x_{\text{bill}} \end{aligned}$$

We can note that as age and the delay in payment increases, the probability of defaulting also increases from the positive estimate value. Furthermore, if the client has completed high school or university, the probability of defaulting also increases, which means that of the three education levels classified in the data, a client who has completed Graduate School will have the lowest probability of default. We can also note that the probability of defaulting increases in men. Looking upon the predictors with negative estimate values, we see that the probability of defaulting decreases as the latest bill statement decreases.

Looking at the p-values of each coefficient estimate, we discern that all estimates are significant to a significance level of $\alpha = 0.05$.

For our goodness of fit test, our residual and null deviance with their corresponding degrees of freedom are shown below. Our null and alternative hypotheses remain the same as seen in Section II-ii. The p-value for this test is the probability that true deviance, G^2 is greater than our residual deviance, $P(G^2 > 27150)$, on a chi-squared distribution of 29525 degrees of freedom, χ^2_{29525}. We calculate this value as 1-$4.7879719 \times 10^{-24}$, or, approximately 1. Since our p-value is much greater than 0.05, we fail to reject the null hypothesis, indicating our model may be a good fit for our data.

	Deviance	**Degrees of Freedom**
Null	3.1387305×10^4	29531
Residual	2.7149938×10^4	29525

Table 3: Null and Residual Deviance

IV. Discussion

IV-i. Summary

We performed logistic regression on the data to find a model which could predict the probability that credit card clients in Vietnam would default on the next repayment of their loan. A goodness-of-fit test was conducted to make sure the model was a good fit to the data. We found that all estimate values in the model were significant in predicting the log probability of default and that the model is an appropriate fit to the data.

IV-ii. Conclusions

The logistic model suggests that there is a difference in default risk between age groups, sexes, and levels of education completed. That is, a young female who has completed graduate school will have a lower risk of default compared to an older man who has completed high school, assuming that they both have made their last payment within the same number of months and their bill statements are the same. The model also suggests that the later a client takes to make a payment and the lower their bill statement is, the higher their probability of default will be.

More specifically, we can say that the log probability of a client defaulting on their next payment will increase by 0.0033822, 1.1550559, and decrease by $-9.1881311 \times 10^{-7}$ for every year older the client is, every month late the last payment was made, and every dollar increase in the last bill statement, respectively. If the client is male, their log probability will be higher by 0.1581998 relative to females. Moreover, if the client's highest level of education is highschool or university, their log probability will be higher by 0.164184 or 0.1609277, respectively, relative to a client whose highest level of education completed is graduate school.

IV-iii. Weaknesses

Although we have seen that this model may be a good fit for the data, a possible setback was the lack of knowledge on how the data was collected. As a result of being unaware of the data collection process, we are unable to tell if there is any bias from practices such as selecting certain data points to get specific results when performing an analysis.

Additionally, we notice that the estimate for the intercept value has the smallest p-value of $9.015231 \times 10^{-212}$. However, we are unable to make an inference on this because the intercept is the point where all predictors are equal to 0. That is, it would be the log probability of a female who has completed graduate school, has made their most recent payment on time, and has a bill statement of 0. We note that this individual must also be 0 years of age, making it not possible to be a credit card client.

Furthermore, this model would not be able to predict the credit risk of potential clients as two of our predictors require that the client have already been approved for the loan.

IV-iv. Next Steps

Future steps to improve the use of this model would be to create a similar dataset, this time including income, the value of all assets, and the number of dependents a possible client has. By including this information, a model could be created with these as predictors, along with age, sex, and level of education, which could be used to predict credit risk before the loan is approved. Thus, it may help determine if the loan should be approved or not.

Also, the model could be made specific to the different types of loans, such as student, mortgage, vehicle, and personal loans. The default risk of a specific client could change within each of these types of loans. One reason being the maturity period of the loan. In general, there is less risk to banks on shorter-term loans because they have a better insight of their clients' financial situations in the near rather than distant future (Corporate Finance Institute, n.d.). Thus, streamlining the model to focus on specific types of loans may help to increase the accuracy in predicting credit risk.

V. References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Berk, J., DeMarzo, P., Stangelann, D. (2018). *Corporate Finance* (2nd Canadian ed.). Pearson Education.

Canada Mortgage and Housing Corporation. (n.d.). *Mortgage Loan Insurance and Premiums.* CMHC. https://www.cmhc-schl.gc.ca/en/finance-and-investing/mortgage-loan-insurance/the-resource/mortgage-loan-insurance-and-premiums

Corporate Finance Institute. (n.d.) *Short Term Loan.* CFI. https://corporatefinanceinstitute.com/resources/knowledge/finance/short-term-loan/

Debbaut, P., Ghrent, A.C., Kudlyak, M., & Romero, J. (2013). *Economic Brief: How Risky Are Young Borrowers?*. Federal Reserve Bank of Richmond. https://www.richmondfed.org/~/media/richmond fedorg/publications/research/economic_brief/2013/pdf/eb_13-12.pdf

Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository: default of credit card clients Data Set* [Data set]. Irvine, CA: University of California, School of Information and Computer Science. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Fuinhas, J.A., Ferreira Motinho, V.M., Estefano, S. (2019). Delinquency and Default in USA Student Debt as a Proportional Response to Unemployment and Average Debt per Borrower. *Economies*, 7(4):52. doi.org/10.3390/economies7040100

Goodman, L., Zhu, J., Bai, B. (2016). *Women Are Better Than Men At Paying Their Mortgages.* Urban Institute. https://www.urban.org/research/publication/women-are-better-men-paying-their-mortgages

Government of Canada. (2019). *Getting pre-approved and qualifying for a mortgage.* Canada. https://www.canada.ca/en/financial-consumer-agency/services/mortgages/preapproval-qualify-mortgage.html

Sheather, S.J. (2009). *A Modern Approach to Regression with R* (pp. 125-149). Springer Science+Business Media. doi:10.1007/978-0-387-09608-7_8

Wickham, H., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473-2480.