Credit Card's Fate: A Statistical Analysis on the Factors that Influence Client's Credit Approval

BUS2- 194A
Amy Tan

Introduction

Following the invention of credit cards in the 1950s, they have become an integral part of Americans' daily life due to their convenience and versatility. A credit card is a financial tool offered by banks or other financial institutions as a type of loan, allowing cardholders to borrow funds to pay for goods and services. Credit cards can be used to make purchases, both online and in-person, without needing to carry around cash or use a checkbook. Additionally, credit cards often offer rewards for using them, such as cash back, travel points, and more. Finally, credit cards offer a level of security that cash does not, protecting against theft, fraud, and other financial losses. As of 2021, more than 70% of Americans have one or more credit cards, and 14% of Americans have ten or more [1].

With an increasing number of credit card applicants each year, banks are met with an overwhelming amount of applications. Determining whether a person has the means or ability to return their loan is crucial for banks, since if consumers don't pay back their credit, as lenders, the banks suffer from financial losses. On the other side of the risk spectrum, the banks face risks of a business loss that results from incorrectly denying good candidates. Thus, proper analysis of credit card applications before approving or issuing credit cards to consumers has become a prevalent risk management strategy in the financial sector.

Problem Statement

As the market for credit cards continues to increase, credit card companies are enforcing stricter lending regulations and often use credit scoring systems to determine a consumer's creditworthiness. Those with a lower credit score or with no credit history may not meet necessary criteria for approval. In addition, consumers that struggle with lower income and high debt from previous stages are also often denied due to their potential risk. In either case, applicants may be equally capable of being liable for a credit card as other "good" candidates, but face increased odds of denial. We hope to look at past applicants and their approval status to identify correlations between variables associated with credit approval to help potential applicants increase their odds of being approved under the circumstance that they do not pose a potential risk to the bank.

Objectives

Algorithms are increasingly being used to automate the decision-making process of credit applications. These algorithms use various data points, such as credit history, financial status, and other relevant information, to determine the likelihood of the applicant's ability to pay back loans. This automated process can provide a more efficient and objective decision-making process, while also reducing the amount of manual labor required. In this project, our team collected data from Kaggle, which provided a clean version of the Credit Approval dataset available in the archives of machine learning repository of UCI [2].

The main goal of this project is to explore the impact of different factors like Age, Income, Credit Score, Years Employed, Prior Default, etc. , on the approval of credit cards for applicants within our dataset. This project applies statistical knowledge and analytical tools (R programming language) to identity significant variables that determine a company's decision to approve or deny credit card applications. These variables

are then used to develop a model that can predict whether or not an applicant should be approved or rejected for credit for future applications.

Methods

A. Statistical Analysis Method

For our project, we used Binomial Logistic Regression because our analysis involves predicting a binary dependent variable (whether an applicant was approved or not) based on multiple continuous (Age, Income, Debt, Credit Score, Employed, Years Employed) or nominal independent variables (Gender, Married, Bank Customer, Prior Default, Drivers License, Ethnicity, Citizen, Industry). Figure 1 shows our analytical process beginning with collecting our data followed by exploring our raw data through descriptive statistics, processing and selecting our predictor variables, and splitting our data into training and testing sets for analysis. For our model, we assumed an alpha level of significance of 0.05, independence of observations, and little to no multicollinearity among independent variables (Figure 1).

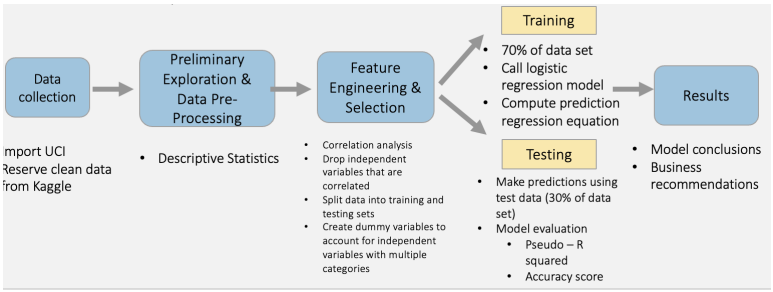

Figure 1. Prediction framework for credit approval

B. Preliminary Exploration - Descriptive Statistics

To get a general understanding of our data, we summarized the statistics of our data, which shows a five number summary for each of the independent variables in our data set. Some of our preliminary insights include a mean applicant age of 31.51, mean monthly income of $1017.4, and mean credit score of 240. Statistics for other variables are included in Figure 1.

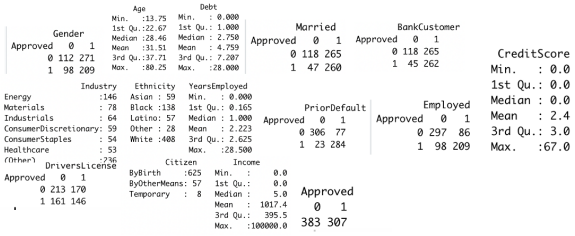

Figure 1. Summary Statistics of Data

To visualize some of the statistics of the applicants in our dataset, we explored variables that we assumed would play a role in determining credit approval. Prior Default, Employment, and Credit Score are common predictors banks use to asses an applicant's risk for credit. Ethnicity was included because we were curious whether one ethnicity had an advantage over another for credit approval. Looking at the stacked bar chart for prior default (Figure 2a), we can see that there is a significant difference in approval rates depending on whether an applicant has a Prior default (1) or not (0). Almost 90% of applicants are denied credit for the former, while the latter sees a much higher approval percentage and lower denial percentage. Figure 2b shows that applicants who are unemployed (0) face a much higher denial rate for credit card application compared to those who are employed (1). Figure 2c shows that applicants who have a lower credit score has a definite higher rate of being denied (0) than those

with higher credit scores. Contrary to our assumption, ethnicity doesn't appear to affect credit approval. Figure 2d shows that the credit approval rates for each ethnicity seem to be fairly balanced– each with similar frequencies of approval and denial.

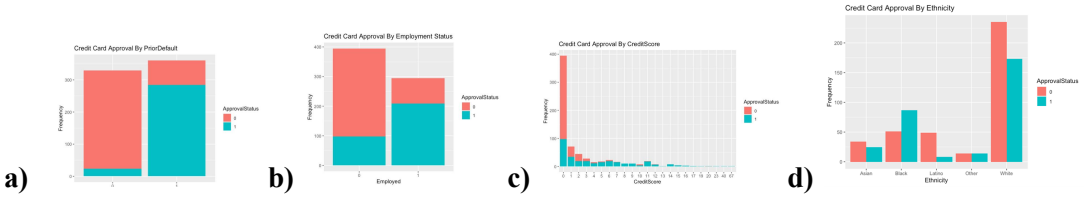

Figure 2. Credit approval visualization with prior default, employment status, ethnicity, and credit score. a) stacked bar chart showing approval outcomes based on prior default, b) stacked bar chart showing approval outcomes based on employment status, c) stacked bar chart showing approval outcomes based on credit score, d) stacked bar chart showing approval outcomes based on ethnicity

C. Feature Engineering (Data Pre-Processing)

The dataset for this project consisted of independent variables that had multiple categories (Ethnicity, Citizen, Industry), which had to be transformed into numbers so they could be run properly against our dependent variable. We created a unique identifiers for each category under each variable. For example, the variable Citizen has 3 categories: By Birth, By Other Means, and Temporary, so we assigned the value of 1 to represent " By Birth", otherwise output 0, and 2 to represent " By other Means," otherwise output 0. Only (k-1) indicator variables, where k= number of categories, was used since we will know if an applicant is a " Temporary" citizen if their output is 0. Ethnicity and Industry variables were transformed in the same way.

D. Feature Selection

In the beginning, we removed the " Zip Code" column from our dataset as it is not relevant in determining credit approval. To account for other conflicting variables, we first did a correlation test, using a threshold of r= 0.7 to test for multicollinearity among independent variables. Figure 3 depicts the r values for the independent variables in our dataset. We can see that only the variables " Married" and " Bank Customer" are highly correlated, with r = 0.99, which is greater than 0.7. To risk potential errors in our model that come from having two highly correlated variables, we dropped the column " Married" from our dataset.

	Gender	Age	Debt	Married	BankCustomer	Industry	Ethnicity	YearsEmployed
Gender	1.00	0.04	-0.04	-0.07	-0.07	-0.02	-0.10	0.09
Age	0.04	1.00	0.20	0.11	0.10	0.12	0.04	0.39
Debt	-0.04	0.20	1.00	0.07	0.08	0.08	0.00	0.30
Married	-0.07	0.11	0.07	1.00	0.99	-0.08	-0.06	0.07
BankCustomer	-0.07	0.10	0.08	0.99	1.00	-0.08	-0.06	0.08
Industry	-0.02	0.12	0.08	-0.08	-0.08	1.00	0.17	-0.04
Ethnicity	-0.10	0.04	0.00	-0.06	-0.06	0.17	1.00	-0.01
YearsEmployed	0.09	0.39	0.30	0.07	0.08	-0.04	-0.01	1.00
PriorDefault	-0.03	0.20	0.24	0.15	0.14	-0.15	-0.03	0.35
Employed	-0.08	0.09	0.17	0.18	0.17	-0.13	0.05	0.22
CreditScore	-0.02	0.19	0.27	0.11	0.11	-0.08	-0.01	0.32
DriversLicense	0.05	0.05	-0.01	-0.01	0.00	-0.08	-0.09	0.14
Citizen	0.07	-0.02	-0.09	0.00	0.00	-0.04	-0.03	0.02
Income	0.00	0.02	0.12	-0.01	0.06	0.03	-0.10	0.05

	PriorDefault	Employed	CreditScore	DriversLicense	Citizen	Income
Gender	-0.03	-0.08	-0.02	0.05	0.07	0.00
Age	0.20	0.09	0.19	0.05	-0.02	0.02
Debt	0.24	0.17	0.27	-0.01	-0.09	0.12
Married	0.15	0.18	0.11	-0.01	0.00	-0.01
BankCustomer	0.14	0.17	0.11	0.00	0.00	0.06
Industry	-0.15	-0.13	-0.08	-0.08	-0.04	0.03
Ethnicity	-0.03	0.05	-0.01	-0.09	-0.03	-0.10
YearsEmployed	0.35	0.22	0.32	0.14	0.02	0.05
PriorDefault	1.00	0.43	0.38	0.09	-0.05	0.09
Employed	0.43	1.00	0.57	0.02	-0.18	0.08
CreditScore	0.38	0.57	1.00	0.01	-0.10	0.06
DriversLicense	0.09	0.02	0.01	1.00	0.04	0.02
Citizen	-0.05	-0.18	-0.10	0.04	1.00	-0.14
Income	0.09	0.08	0.06	0.02	-0.14	1.00

Figure 3. Correlation between independent variables for multicollinearity assumption check

Before further analysis, we split our data into training and testing sets, the former containing 70% of our data and the latter with 30%. The training data set was used to analysis our data and build our model.

First, we called a logistic regression model using all our predictor variables from our dataset, which we used to identify which of our variables were significant at an alpha level of 0.05 (p-value less than 0.05). The significant variables came to be whether or not an applicant has a prior default, their credit score, monthly income, citizenship status, and industry of employment. Using those variables, we created a second logistic regression model which focuses on just our significant variables and is shown as Figure 4. The variables that were not selected by our model were dropped from our data set.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0583777  0.6226009  -1.700 0.089144 .
PriorDefault  3.5890748  0.3587495  10.004  < 2e-16 ***
CreditScore   0.1805077  0.0544140   3.317 0.000909 ***
Income        0.0004588  0.0001611   2.847 0.004408 **
Citizen      -1.3699147  0.4727591  -2.898 0.003759 **
Industry     -0.0902874  0.0375665  -2.403 0.016243 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4. Logistic Regression Model for Credit Approval. Contains only significant independent variables.

Results

A. Prediction Model

Figure 5 shows our predictive model/ estimated regression equation. The intercept and values of predictor coefficients required to form an estimated regression equation was taken from Figure 4.

$$\hat{y} = \frac{e^{-1.06+\ 3.59x_1+0.18x_2+0.0005x_3-1.37x_4-0.09x_5}}{1+\ e^{-1.06+\ 3.59x_1+0.18x_2+0.0005x_3-1.37x_4-0.09x_5}}$$

Figure 5. Estimated Logistic Regression Equation for Credit Approval. x1= Prior Default, x2= Credit Score, x3=Income, x4=Citizen, x5= Industry

To interpret our intercept and coefficients with context of our data, we computed the odds ratio for each along with their respective confidence intervals (Figure 6) to make sure our ratios are significant.

```
                    OR       2.5 %     97.5 %
(Intercept)   0.3470183  0.10015779  1.1615678
PriorDefault 36.2005675 18.57498635 76.4238667
CreditScore   1.1978253  1.08564215  1.3445313
Income        1.0004589  1.00016372  1.0007842
Citizen       0.2541286  0.09785539  0.6309999
Industry      0.9136686  0.84783942  0.9828086
```

Figure 6. Odds ratio Output for Coefficient Interpretation of Estimated Regression Equation

B. Model Evaluation

We applied measures of model fit and model accuracy to evaluate our model. To check how well our model fits our data, we computed McFadden’s Pseudo- R2, which is one of the multiple goodness of fit measures identified for logistic regression. This value is computed by comparing the model we are using with a null model, one containing no predictor variables. Figure 7 shows our output pseudo- r2 value of 0.52. As pseudo r2 values tend to be relatively low, it is said that values ranging from 0.2 to 0.4 already indicate good model fit[]. Thus, with a pseudo r2 of 0.52, we can assume that our model is a great fit. We can also interpret this as our model having a 52% increase in model fit compared to a null model.

```
fitting null model for pseudo-r2
        llh      llhNull           G2     McFadden         r2ML         r2CU
-160.2443444 -331.4188835  342.3490782    0.5164900    0.5077641    0.6802061
```

Figure 7. McFadden’s Pseudo- R2 to Assess Model Fit

To calculate our model accuracy, we made predictions using our Test data. First predictions of credit approval were made based on our predictive model for applicants in our Test data. Those predicted

probabilities were then used to categorize respective applicants to be either approved (indicated with the number 1) or not approved (indicated with 0). The classification for applicants was made using a default probability threshold of 0.5 where applications with a predicted probability greater than 0.5 are classified as being “approved” and those with a predicted probability less than 0.5 are classified as “not approved.” Our model accuracy of 86% shown in Figure 8 reflects the proportion of applications that were correctly classified by our predictor model, which was calculated by comparing the predicted approval outcomes to actual outcomes. With 86% accuracy, our classification error, the proportion of observations that have been misclassified, is 14%, which is relatively low.

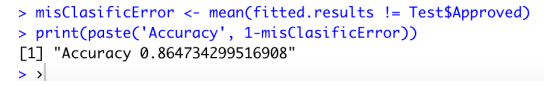

Figure 8. Computation of Model Accuracy

Conclusion & Recommendations

In conclusion, using past data of a collective of client variables ranging from Age, Income, Employment, Credit Score, etc. against credit approval, we identified the variables Prior Default, Credit Score, Income, Citizen, and Industry to have a significant impact in determining applicants’ credit card approval. We used binomial logistic regression to analyze our data and trained a predictive model to help analysts automate the credit approval processes among different industries. However, there are variances in requirements for credit approval– different companies have different thresholds of credit score and other variables that they consider to make a “good” applicant–the model should only serve as a reference. Companies should adjust and account for credit requirements for their respective companies in their predictive models. Some other recommendations include categorizing applicants into categories of low, medium, or high risk based on past client data to see if applicant’s category matches the approval decision made by a predictive model. Also, businesses should take into consideration exceptions that cause the predictive model to yield different results than actual approval. These steps will help prevent wrongly denying or wrongly accepting applicants which is associated with potential financial risks. These results can also serve as a source of information for consumers – having a better idea of which factors are most significant in credit approval outcome may give them a better idea of their chances of being approved for credit and how they can improve their application. Under the situation where companies that offer credit cards are transparent with their requirements, consumers should choose to apply to the credit card that they think they have the best chance at getting approved for– based on their understanding of their personal statistics– to avoid being rejected more times than necessary, which lowers their credit score.

Bibliography

[1] Credit Card Statistics [Internet]. Shift Credit Card Processing. 2021. Available from: https://shiftprocessing.com/credit-card/.

[2] Cortinhas S. Credit Card Approvals (Clean Data) [Internet]. www.kaggle.com. 2022 [cited 2022 Nov 30]. Available from: https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data.

[3] Anderson DR. Statistics For Business & Economics. New York: Cengage Learning; 2019.