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

Amy Tan
BUS2- 194A Sec01
11/30/22

AGENDA

- Introduction
  - Background about project topic (Credit cards)
  - What our project is about and why we chose it
- Problem Statement
  - What is the problem we are addressing/ looking to solve?
- Objectives
  - Our goals and objectives going into our project
- Methods
  - Our analytical approach / analysis method
  - How we collected and processed our data
  - Descriptive statistics (numbers and visual charts)
- Results
  - Explain our findings
  - Logistic regression model
  - Coefficient interpretations
  - Model Fit, Model predictions and accuracy
- Conclusion & Recommendation
  - Conclusions from our study
  - Business recommendations/ Future Scope
- Lessons Learned
  - Lessons we learned from our project
- Challenges
  - Challenges we faced working as a team and how we solved our problems

INTRODUCTION

Based on research from 2021, the average credit card debt for Americans is $5,221 with a low default rate and delinquency rate of 1.81%, yet only 40% of credit card applicants get approved

Analysis of credit card applications before approving or issuing a credit card to a consumer is a prevalent risk management strategy in the financial sector.

Determining whether a person has the ability to return their loan is crucial for banks/ financial institutions since if consumers don't pay back credit, the lender suffers a financial loss

Another potential risk is a business loss that results from incorrectly denying good candidates

PROBLEM STATEMENT

There are currently too many applicants getting rejected due to factors such as credit score, age, insufficient income, etc.

Our goal is to look at past applications and their approval status to determine why certain individuals were approved and why others were denied.

We want to identify correlations between variables associated with credit card approval/denial to help potential credit card applicants increase their odds of getting approved under the circumstances that they do not pose a potential risk to the bank.

OBJECTIVES

The goal of our project is to use analytical techniques to predict a company/financial institution's decision to approve or deny credit card applications.

In order to forecast future application acceptance or rejection, we will test for relationships between different variables and credit card approval status using data of past credit applications from UCI.

Using our findings, we aim to build a model that predicts credit card approval given information from an individual's credit card application.

METHODS

- **Binomial Logistic Regression**
  - Our analysis involves predicting a binary dependent variable based on multiple continuous or nominal independent variables.

Data collection

Import UCI
Reserve clean data
from Kaggle

Preliminary Exploration & Data Pre-Processing

- Descriptive Statistics

Feature Engineering & Selection

- Correlation analysis
- Drop independent variables that are correlated
- Split data into training and testing sets
- Create dummy variables to account for independent variables with multiple categories

Training

- 70% of data set
- Call logistic regression model
- Compute prediction regression equation

Testing

- Make predictions using test data (30% of data set)
- Model evaluation
  - Pseudo – R squared
  - Accuracy score

Results

- Model conclusions
- Business recommendations

DESCRIPTIVE STATISTICS

```
      Age              Debt           CreditScore
 Min.   :13.75   Min.   : 0.000   Min.   : 0.0
 1st Qu.:22.67   1st Qu.: 1.000   1st Qu.: 0.0
 Median :28.46   Median : 2.750   Median : 0.0
 Mean   :31.51   Mean   : 4.759   Mean   : 2.4
 3rd Qu.:37.71   3rd Qu.: 7.207   3rd Qu.: 3.0
 Max.   :80.25   Max.   :28.000   Max.   :67.0
```

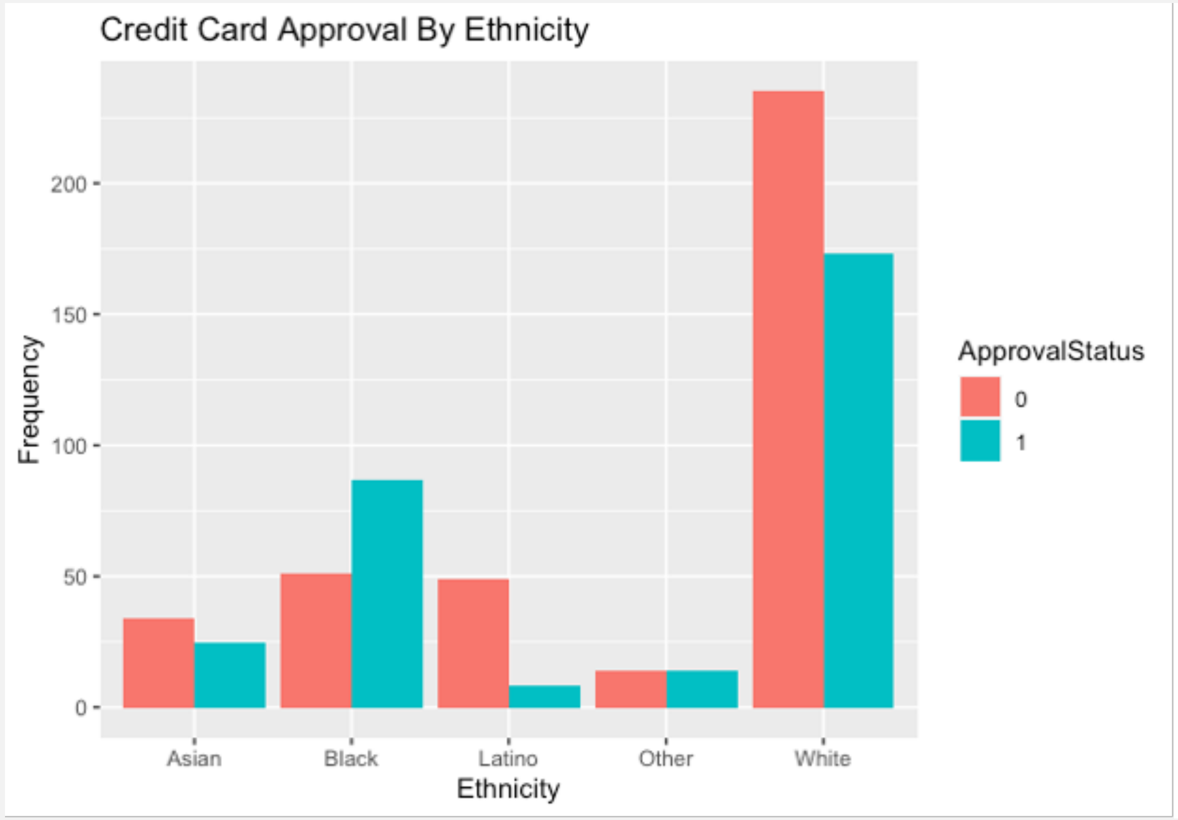
```
     Income
 Min.   :      0.0
 1st Qu.:      0.0
 Median :      5.0
 Mean   :   1017.4
 3rd Qu.:    395.5
 Max.   :100000.0
```

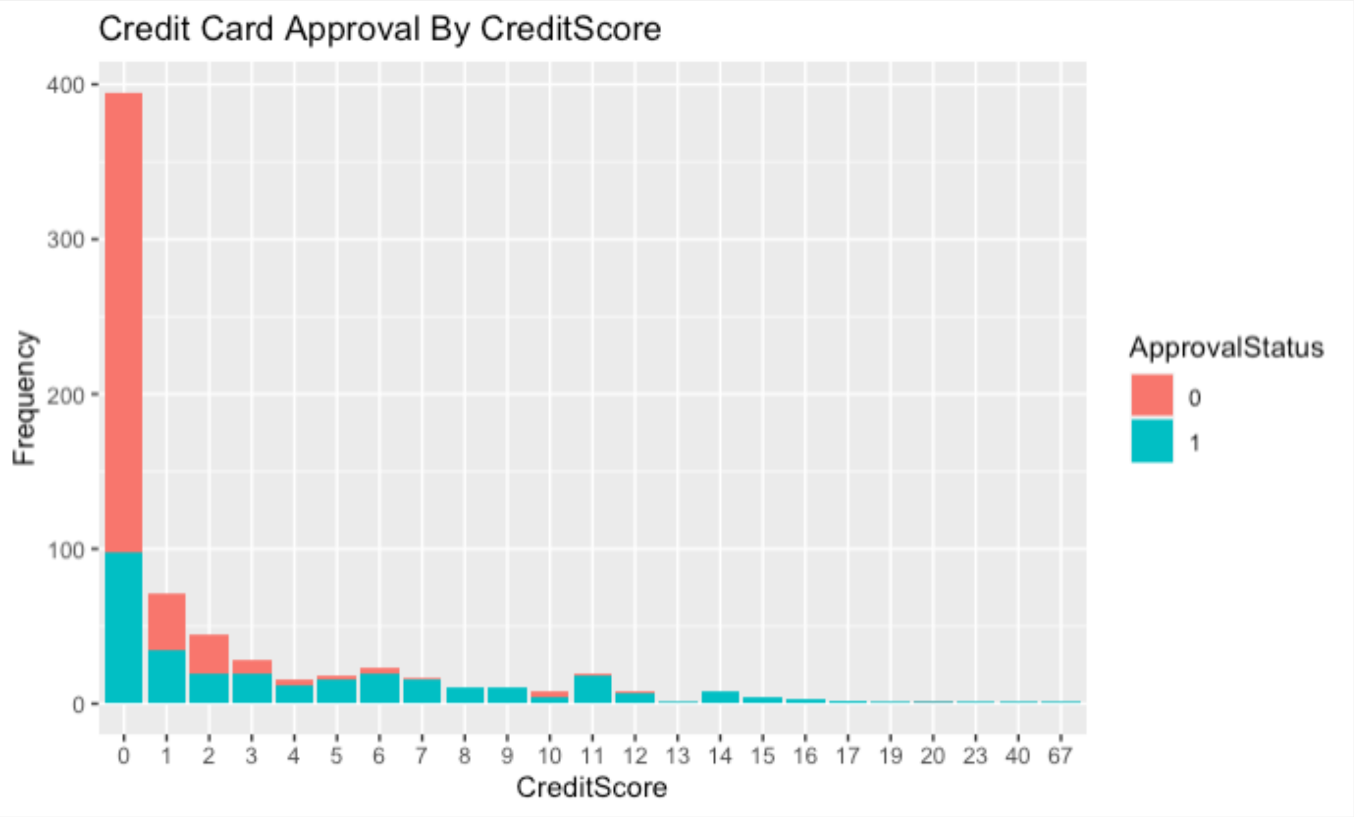
```
         PriorDefault
Approved   0   1
       0 306  77
       1  23 284
```

```
         Employed
Approved   0   1
       0 297  86
       1  98 209
```

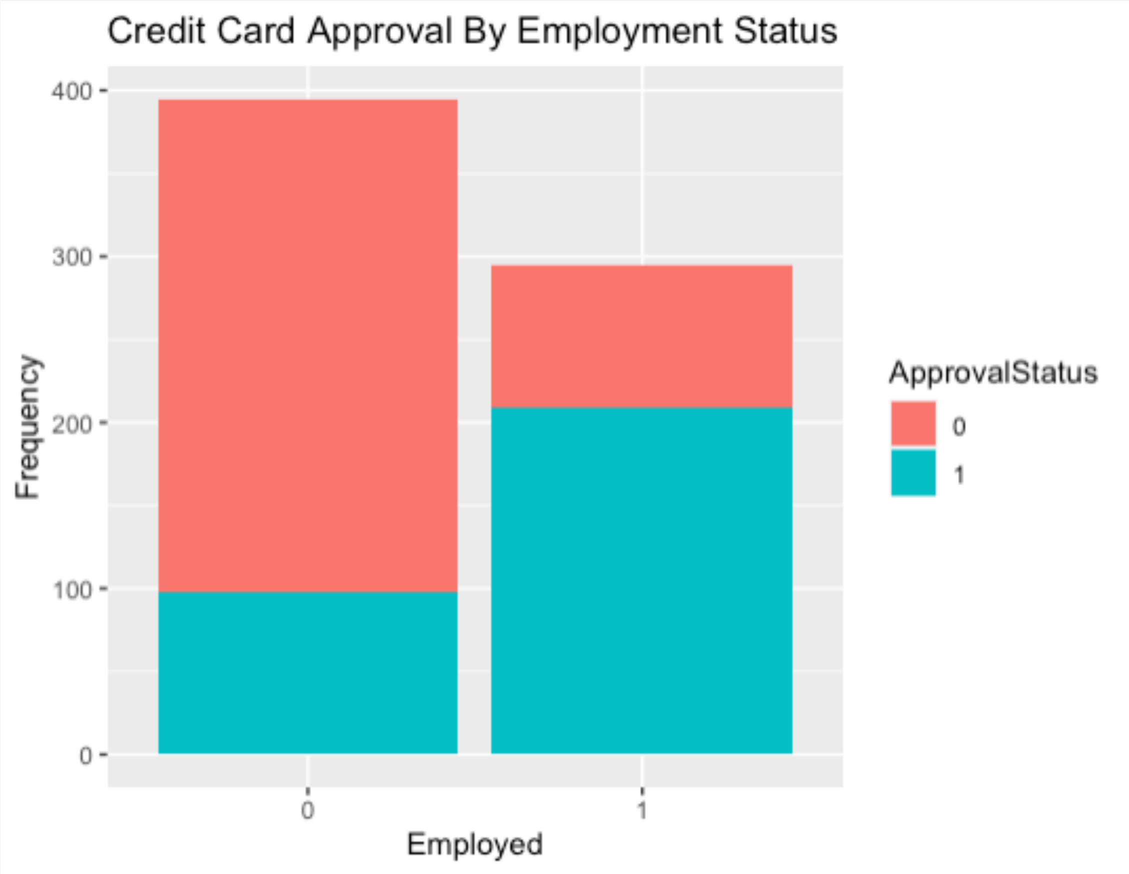
Independent Variable	Descriptive statistics
Age	Mean age is 31.51
Income	Mean monthly income is $1017.4
Debt	Mean debt is $4,759
Credit Score	Mean credit sore is 240, min= 0, max =670
Employed	295/690=43% are employed
Prior Default	361/690 = 52% have a prior default

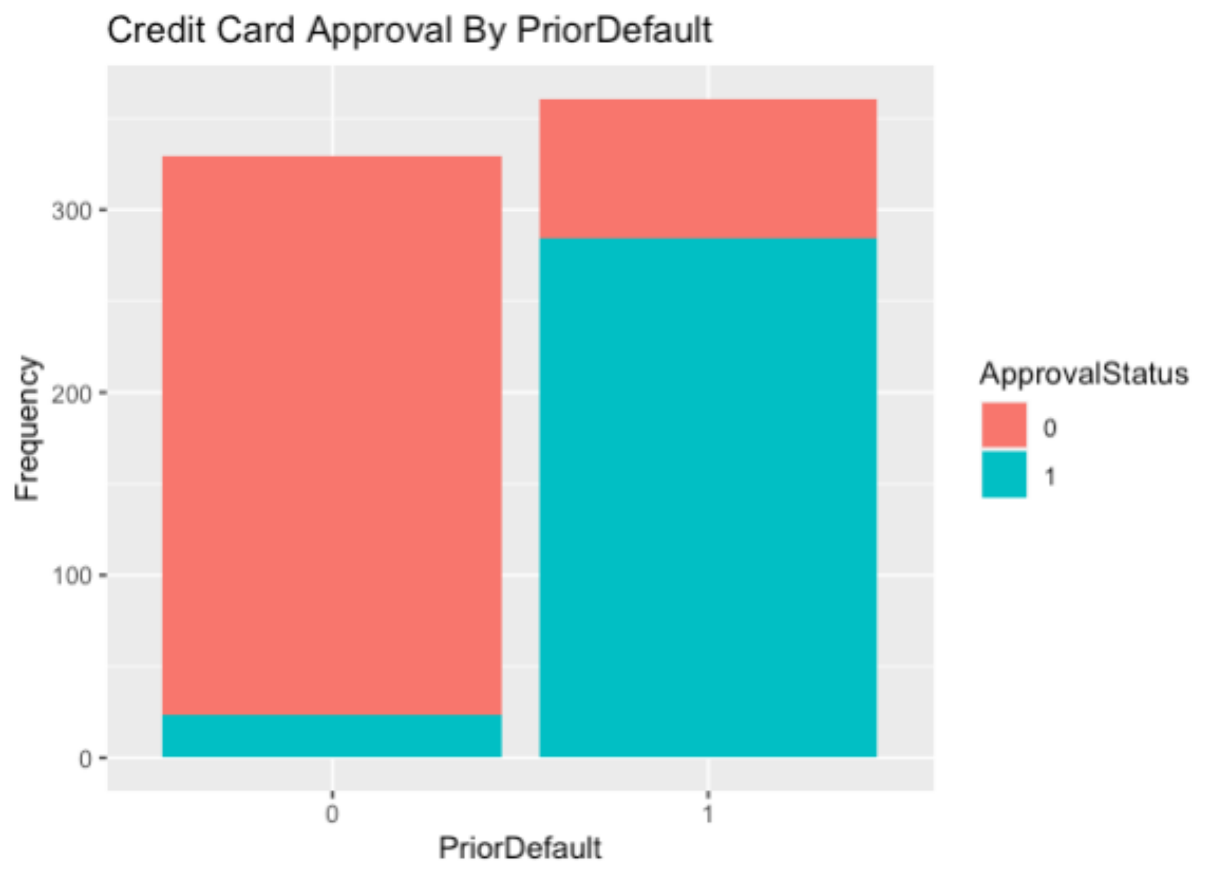
DESCRIPTIVE STATISTICS





DESCRIPTIVE STATISTICS





Assumptions

- Level of significance : $\alpha = 0.05$
- The dependent variable is binary (i.e. a categorical variable with two outcomes)
- The independent and dependent variables are not related linearly
- Observations are independent of each other
- Little to no multicollinearity among independent variables
  - $|r|$ <0.7
  - Drop “Married” column for accuracy

```
> round(corr$r, digits= 2)
              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
>
```

RESULTS

ESTIMATED REGRESSION EQUATION:

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

x1= Prior Default
x2= Credit Score
x3= Income
x4= Citizen
x5= Industry

SIGNIFICANT VARIABLES

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.8479321	0.9554557	-0.887	0.37483	
Gender	0.0385381	0.3265995	0.118	0.90607	
Age	-0.0217344	0.0135745	-1.601	0.10935	
Debt	-0.0077104	0.0303188	-0.254	0.79926	
BankCustomer	0.6678687	0.3718816	1.796	0.07251	.
PriorDefault	3.6739774	0.3943606	9.316	< 2e-16	***
YearsEmployed	0.0846306	0.0514672	1.644	0.10010	
Employed	0.4834080	0.3868830	1.249	0.21148	
CreditScore	0.1283745	0.0651749	1.970	0.04887	*
DriversLicense	-0.5170443	0.3118085	-1.658	0.09727	.
Income	0.0004544	0.0001632	2.784	0.00537	**
Ethnicity	-0.0669836	0.1872788	-0.358	0.72059	
Citizen	-1.4731632	0.4962746	-2.968	0.00299	**
Industry	-0.0774333	0.0389367	-1.989	0.04673	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Only variables with p-value < 0.05 are significant:
Prior Default, Credit Score, Income, Citizen, Industry
-> all other variables will be dropped to improve our model

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

Coefficient Interpretation

b0= -2.35

- If there is no information provided for a credit card applicant, the log odds of being approved for credit is -1.06
- The odds for credit card approval is e^-1.06/ 1+e^-1.06 = 0.35 assuming no information provided about applicant.

b1= 3.44

- Applicants with prior default (1) versus applicants with no prior default (0), changes the log odds of credit card approval by 3.59
- The odds of credit approval for an applicant with no prior default is 36.20 times greater than for an applicant with prior default

b2= 0.19

- For every one unit change in credit score, the log odds of credit approval increases by 0.18
- If log odds increases by 0.18, this means odds will increase by [exp(0.18) =1.20]. This is a 20% increase in the odds of getting approved for credit (assuming other predictor variables remain fixed)

b3 = 0.0005

- For every one unit change in income, the log odds of being approved for credit increases by 0.0005
- This means odds of credit approval will increase by 1.0005 for each unit increase in income.

b4 = -1.37

- Being a citizen " by other means" (2) vs. being a citizen "by birth" (1) changes the log odds of credit approval by -1.37
- The odds of credit approval for an applicant who is a citizen " by other means" is 0.25 times less (a 75% decrease) likely compared to an applicant who is a citizen " by birth"

b5= -0.09

- An applicant who work in "Materials" industry (2) versus an applicant who works in the "Industrials" industry (1) changes the log odds of credit approval by -0.09
- The odds of credit approval for an applicant in the "Materials" industry is 0.91 times less likely (9% decrease) than for an applicant in the "Industrials" industry.

```
> model2$coefficients
  (Intercept)  PriorDefault   CreditScore
-1.0583777339  3.5890747953  0.1805077041
>
```

```
       Income       Citizen      Industry
0.0004588138 -1.3699147151 -0.0902873622
```

Odds Ratio Output:

```
> exp(cbind(OR=coef(model2),confint(model2)))
Waiting for profiling to be done...
                    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
```

McFadden's Pseudo R- squared:

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

Our pseudo r-squared is 0.52, which is good -- (0.2 – 0.4) considered good model fit

Making Predictions using Test Data:

1. Predict credit approval based on predictor variables

```
> fitted.results <- predict(model2, Test,type='response')
> head(fitted.results)
        1         7         9        10        12        18
0.7774734 0.9999997 0.6820320 0.8496814 0.6503399 0.9477323
```

2. Categorize applicants into approved(1) or not approved (0) based on their predicted probabilities of being approved, with p >0.5 as approved

```
> fitted.results <- ifelse(fitted.results >0.5,1,0)
> head(fitted.results)
 1  7  9 10 12 18
 1  1  1  1  1  1
```

Model Accuracy

```
> misClasificError <- mean(fitted.results != Test$Approved)
> print(paste('Accuracy', 1-misClasificError))
[1] "Accuracy 0.864734299516908"
> >
```

Our model has a prediction accuracy of 86%, which is relatively high

Test data insights :

Predicted approval is 98

Actual was 94

Predicted denial 109

Actual denial 113

```
table(Test$Approved,
 fitted.results
    0  1
 0 97 16
 1 12 82
```

CONCLUSION & RECOMMENDATIONS

- Our model has an 86% accuracy in predicting the approval of credit cards
- Prior Default, Credit Score, Income, Citizen, and Industry are significant predictors of credit approval identified by our model
- The predictive model can be used to help analysts automate the credit approval process
- 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
- Consider categorizing applicants into categories of low, medium, or high risk based on past data and see if applicant’s category matches approval decision made by predictive model to prevent wrongly denying or wrongly accepting applicants
- Consider testing the exceptions in which a credit card application yields different results than the predictive model – main risk is approving credit for someone that should have been denied
- Consider adjusting model with regard to requirements for respective financial institutions as accuracy may vary among banks due to variances in requirements for credit approval