**Do household and personal characteristic influence Credit Card Behavior in the United States**

By

Rizwan Mushtaq

**Do household and personal characteristic influence Credit Card Behavior in the United States**

Executive Summary

This short paper will help to predict the credit card behavior based on demographic and household characteristics in the United States. Demographic indicators such as age, gender, marital status, financial literacy, education and profession along with household characteristics such as household income state and region of residence will be used to predict the credit card behavior. These variables play might play crucial role in determining the credit card behavior of any individual. In addition to the exploratory data analysis I expanded this project by adding regression analysis. I also applied classification technique called confusion matrix for logistic regression, where it shows the predicted probabilities.

1. **Introduction & Motivation**

Individuals are different and exhibit diverse behaviors, however a general prediction model could be used to asses one's behavior. Global financial crisis of 2007-2008 is greatly considered as a result of subprime mortgage loans in the United States. In addition, a huge amount of credit card payments goes pending every month in the US. While credit card debt has increase around 32% in the last five years. Therefore, to avoid future banking crisis, it is important to devise a mechanism that can predict the credit card behavior of the potential user.

In this research I will use an open source dataset 'The National Financial Capability Study' (NFCS) provided by FINRA. This dataset is a primary data collected from the US citizens with an interval of three years. The first round of survey was conducted in the year of 2009 following the Global financial turmoil. Later on, second round was conducted in 2012, third in 2015 and most recent survey was conducted in 2018. More detail of the dataset can be found [here](https://www.usfinancialcapability.org/) and can be downloaded [here](https://www.usfinancialcapability.org/downloads.php). This dataset includes a wide range of financial, demographic and professional indicators of American Citizens from all the regions and states. Since dataset includes regional information, I will use Foursquare API to access Foursquare location data for US to compare the US states based on credit card behavior. I will combine the NFCS dataset with the US location data in order to explore the areas and states with good credit records and bad credit records. This result of this analysis might be useful for banks and other financial institutions as well as for Government to devise credit policy in the different regions based on the credit card behavior of the population.

1. **Problem Statement**

Individuals are different and exhibit diverse behaviors, however a general prediction model could be used to asses one's behavior. Global financial crisis of 2007-2008 is greatly considered as a result of subprime mortgage loans in the United States. In addition, a huge amount of credit card payments goes pending every month in the US. While credit card debt has increase around 32% in the last five years. Therefore, to avoid future banking crisis, it is important to devise a mechanism that can predict the credit card behavior of the potential user.

1. **Research Question and Hypothesis**
2. **Research Question**

The research question I want to answer by using this dataset as defined earlier is:

* Can we predict the credit card behavior based on demographic and household characteristics in the US?

1. **Hypothesis**

Personal characteristics do not determine credit card behavior?

Demographic indicators do not influence credit card behavior.

1. **Dataset(s)**

In this research I will use an open source dataset 'The National Financial Capability Study' (NFCS) provided by FINRA. This dataset is a primary data collected from the US citizens with an interval of three years. The first round of survey was conducted in the year of 2009 following the Global financial turmoil. Later on, second round was conducted in 2012, third in 2015 and most recent survey was conducted in 2018. More detail of the dataset can be found here and can be downloaded here.

This dataset includes a wide range of financial, demographic and professional indicators of American Citizens from all the regions and states. Since dataset includes regional information, I will use Foursquare API to access Foursquare location data for US to compare the US states based on credit card behavior. I will combine the NFCS dataset with the US location data in order to explore the areas and states with good credit records and bad credit records. This result of this analysis might be useful for banks and other financial institutions as well as for Government to devise credit policy in the different regions based on the credit card behavior of the population.

1. **Data Preparation and Cleaning**

I combined four survey datasets from 2009 to 2018. In the first step of data preparation and cleaning I renamed the columns and selected most relevant ones to advance further. Then I removed missing values and converted string variables to integers where it was necessary.

Secondly, for the purpose of analysis I took credit card record as dependent variable (DV), this variable coded in four categories from very bad to very good. To create a binary variable for the purpose of analysis, I took the average of DV and gave 1 if an individual scored above average zero otherwise.

* I converted independent (X) variables from string to numeric with the command pd.to\_numeric.
* Finally, I made two sets of variables to analyze further, First set of indicators include X variables that are believed to be predictors.
* Then as defined above I assigned dummy variable (DV) as defined above as Y.

1. **Methodology**

After data preparation and cleaning I moved towards data analysis, as my DV is a binary variable therefore, I choose to use logistic regression, that is rightly suitable for my dataset. I firstly used mosaic plot to show the probability of the variables, as they are useful when we have categorical data. As a first step mosaic plots are shown here that exhibit the descriptive as well as the association between different variables. In the second step I used logit model, since my dependent variable is transformed to binary variables. Hence, the most appropriate method applicable is logit model. Results are presented and interpreted in the following section.

1. **Results**

Figure: 3 shows mosaic plot, here, I use mosaic plot to examine the association between my DV and other independent variables such as financial literacy. Financial literacy ranges from 0-6, as we can see higher financial literacy falls in category 1 meaning that the respondants with higher financial knowledge tend to possess good credit card behavior and vice versa.

Figure: 1 Mosaic Plot-Financial literacy | Credit record

A screenshot of a cell phone

Description automatically generated

Figure 2 shows mosaic plot of age group and credit card record. Here, I use mosaic plot to examine the association between my DV and other independent variables such as age group of the respondants. Age group ranges from 1-6, as we can see middle age group from 4-6 have better credit record as compared to young group 1-3. Keeping other things constant, we cany fairly conclude that credit record improves with age.

Figure: 2 Mosaic Plot-Age group | Credit record

A screenshot of a cell phone

Description automatically generated

**Regression Analysis**

In the second part of the data analysis I used different regression model suggested in the course. I started with the simple OLS model by taking credit variable in its continuous form along with several features discussed above. The results of ordinary least square regression are presented in Table 1. Where we can see the value of adjusted r-squared quite higher that indicates the overall significance of the model. Furthermore, most of the explanatory variables indicate statistically significant coefficients. In the next step I used logit model by taking binary variable of credit card behavior as dependent variable and personal and household characteristics as dependent variables. The binary variable 1 represents good behavior while 0 represents bad credit card behavior, gender 1 male 0 otherwise, age group is also a binary variable along with the list of independent variables explained earlier. I applied logit regression model by using Python package statsmodel.api. Results are presented in below table 1. These results indicate negative impact on the credit card behavior. As most of the independent variables show negative and significant coefficients. However, category of education, retired, military and spouse profession indicators show positive significant impact on credit card behavior. These, results indicate that personal factors negatively determine credit card behavior of US household. On the contrary, if the respondent is educated, retired, in military and good profession of spouse they tend to have good credit card behavior in the given sample.

Table:1 OLS Regression

Table

Description automatically generated

Table:2 Results: Logit

================================================================

Model: Logit Pseudo R-squared: 0.864

Dependent Variable: y AIC: 6748.9765

Date: 2020-10-13 19:39 BIC: 6910.2875

No. Observations: 35956 Log-Likelihood: -3355.5

Df Model: 18 LL-Null: -24707.

Df Residuals: 35937 LLR p-value: 0.0000

Converged: 1.0000 Scale: 1.0000

No. Iterations: 9.0000

----------------------------------------------------------------

Coef. Std.Err. z P>|z| [0.025 0.975]

----------------------------------------------------------------

gender -0.3688 0.0695 -5.3092 0.0000 -0.5050 -0.2327

agegrp -0.0658 0.0295 -2.2264 0.0260 -0.1236 -0.0079

ethn -1.0931 0.0836 -13.0806 0.0000 -1.2569 -0.9293

edu 1.8921 0.0277 68.1974 0.0000 1.8377 1.9465

marital -0.2211 0.0371 -5.9624 0.0000 -0.2938 -0.1484

hhincome -0.2017 0.0209 -9.6663 0.0000 -0.2427 -0.1608

armservice 0.8679 0.0437 19.8545 0.0000 0.7822 0.9536

retired 1.0670 0.1003 10.6354 0.0000 0.8704 1.2637

prof -0.0180 0.0162 -1.1130 0.2657 -0.0497 0.0137

profspouse 0.0054 0.0165 0.3265 0.7440 -0.0270 0.0378

currentstudent 0.0881 0.0934 0.9431 0.3456 -0.0950 0.2712

finsatisfacton 0.1525 0.0141 10.7786 0.0000 0.1247 0.1802

willingrisk -0.1184 0.0139 -8.5369 0.0000 -0.1456 -0.0912

savingchildedu -0.2234 0.0453 -4.9315 0.0000 -0.3122 -0.1346

finconfdaytoday -0.0642 0.0232 -2.7712 0.0056 -0.1096 -0.0188

fl -0.1872 0.0295 -6.3427 0.0000 -0.2450 -0.1294

usstate -0.0102 0.0023 -4.5003 0.0000 -0.0146 -0.0057

censusdiv 0.2739 0.0519 5.2759 0.0000 0.1722 0.3757

censusreg -0.8790 0.1237 -7.1070 0.0000 -1.1214 -0.6366

================================================================

Figure 2: provides confusion matrix. As defined earlier my DV is a binary variable i.e., credit record behavior, above average and below average (labeled as credit\_bin). Therefore, logistic regression is suitable for this kind of analysis. This confusion matrix shows that in first row the respondents whose actual financial literacy value is 1, as out of (2929+291+67+3905)=7192 respondents, the credit\_bin value for (2929+291) = 3220 is 1. While out of these 3220 my classifier model correctly predicted 2929 as 1 and 291 as 0.

Figure: 3 Confusion Matrix Logistic Regression

A screenshot of a cell phone

Description automatically generated

This result implies that for 2929 respondents of the survey the actual credit card value in the survey is 1 in the test set and that the classifier also predicted them as 1, that is quite significant number in the model prediction ( with around 91% accuracy). However, on the other hand, for 291 the actual value was 1 in the dataset but the classifier model predicted them as 0, which we may say that this is error of the model ( with around 91% accuracy). Similar interpretations can be applied for the second row.

1. **Conclusion**

In this short paper I attempt to predict credit card behavior based on demographic and household characteristics in the United States. For that purpose I used NFCS dataset from 2009-2018, and applied most relevant machine learning approaches. I found that demographic characteristics such as age, gender, marital status etc. are strong predicters of financial literacy of an individual. Keeping other things constant, we cany fairly conclude that credit record improves with age. The results confirm the accuracy of the model as well as acceptable percentage of the correctly predicted values for logistic regression.

Reference:

<https://www.usfinancialcapability.org/downloads.php>

<https://www.usfinancialcapability.org/>