Predicting Credit Card Behavior in the United States

Rizwan Mushtaq

Abstract

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 order to do so I will utilize machine learning techniques such as Logistic regression, K-mean clustering and visualization.

Motivation

The motivation behind this idea is to predict the credit card behavior among US citizens based on their characteristics such as demographic (state of residence, age, gender, marital status) education, profession etc. By doing so one can determine the credit card risk by the US citizen based on these mentioned features. The results will be useful for several Governmental and private sector financial institutions to assess the credit card behavior.

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.

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.

Dataset(cont.)

- 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.

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,

Data Preparation and Cleaning (cont.)

- Data preparation and cleaning took a lot of time as I repeatedly found error ‘could not convert string to float’.
- 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.

Research Question(s)

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

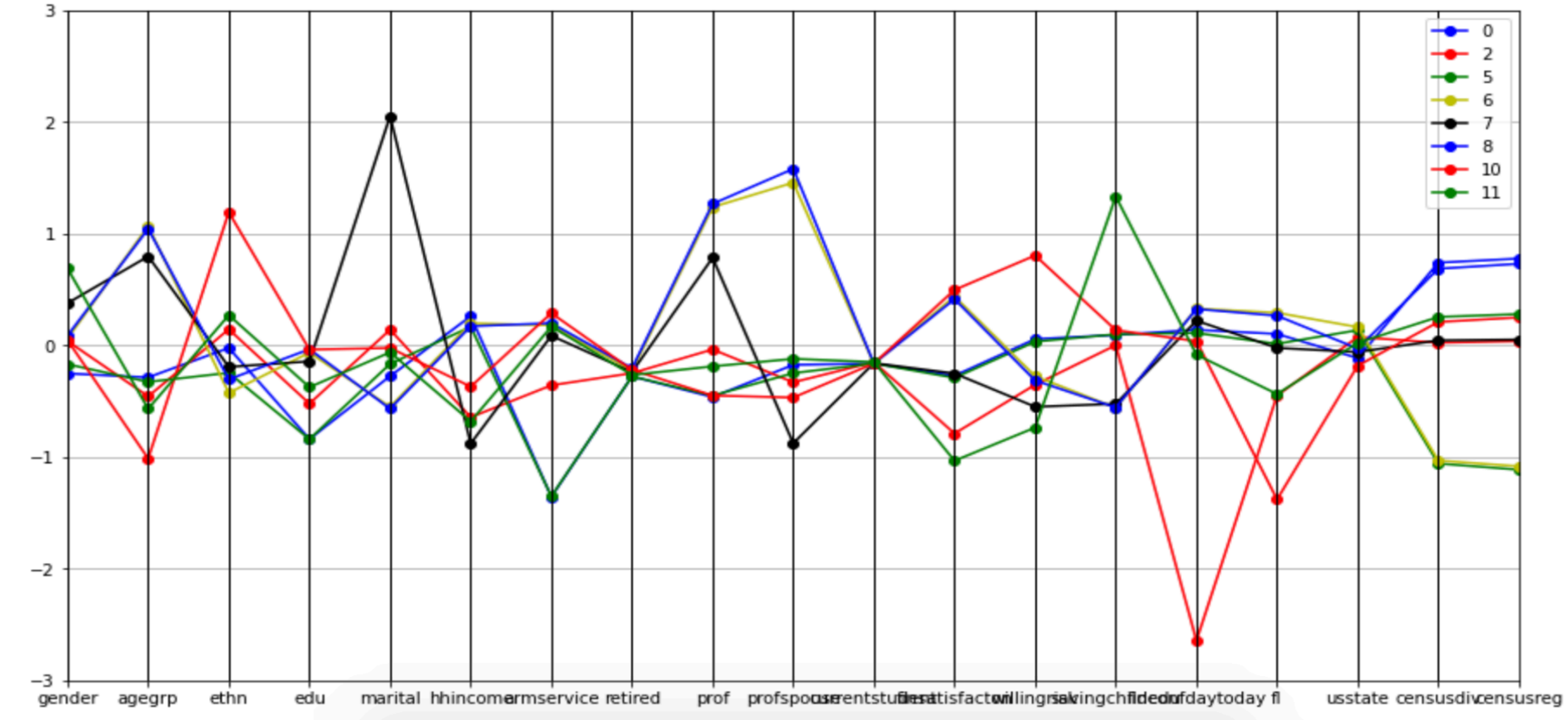
- Can we predict the credit card behavior based on demographic and household characteristics in the US ?

Methods

- After data preparation and cleaning I moved towards data analysis, as my DV is a binary variable therefore, I choose to use two machine learning approaches, that are rightly suitable for my dataset.
- Firstly, I applied K-means clustering and then I applied Logistic regression. Since, by dependent variable is binary (converted for that purpose) I can not proceed with regression as it requires DV to be continous.
- I also used mosaic plot to show the probability of the variables, as they are useful when we have categorical data.

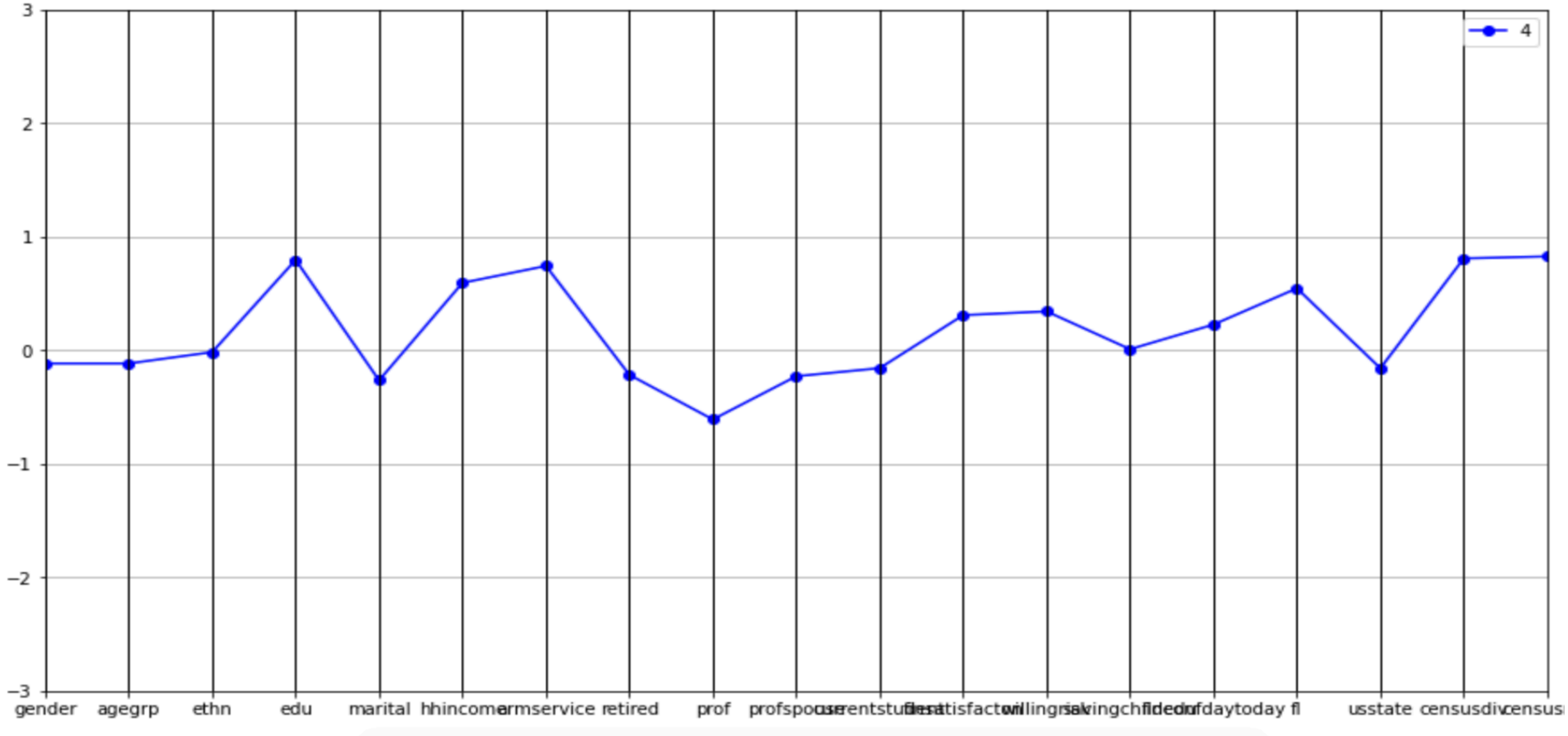
Findings- K-means clustering

```
parallel_plot(P[P['edu'] < 0.4])
```



Findings- K-means clustering

```
parallel_plot(P[P['fl'] > 0.5])
```



Findings- K-means clustering

- Figur above in previous slide shows a graph of K-means clustering by using dataset defined above. I divided dataset as 80% for traning and 20% for testing purpose.

- This figure gives us the centroids with normalized data, we can observe a significant number of identical clusters.

- For instance if we see, red, green, black blue are the identical clusters and preditcted by the same features i.e., demographic and professional.

- Train set Accuracy: 0.9624878320122375
  Test set Accuracy: 0.9502224694104561

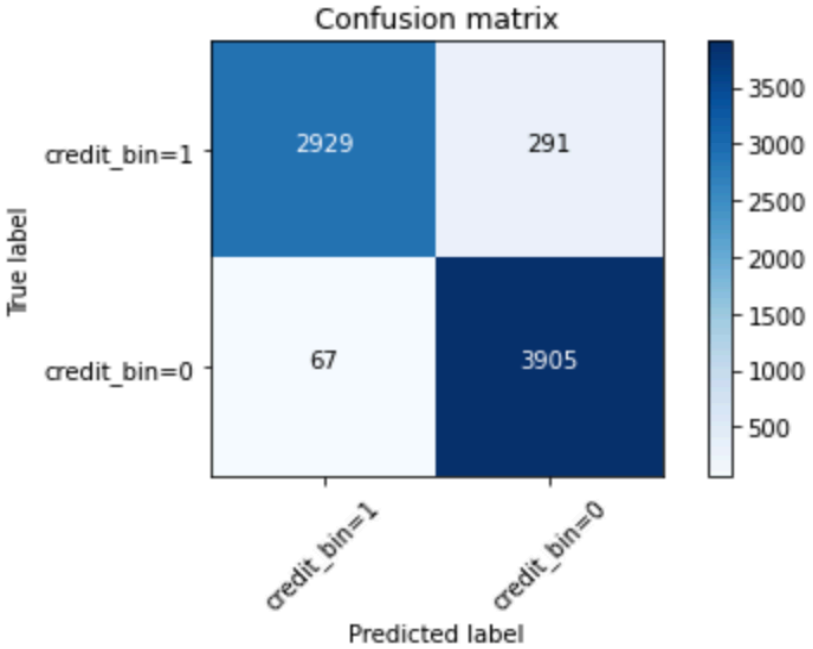
- The accuracy of the model is quiet ok, as the accuracy for training set is around 96% and the accuracy for testing set is 95% that is acceptable according to standards.

Findings- Confusion Matrix Logistic Regression

- 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.

```
Confusion matrix, without normalization
[[2929  291]
 [  67 3905]]
```



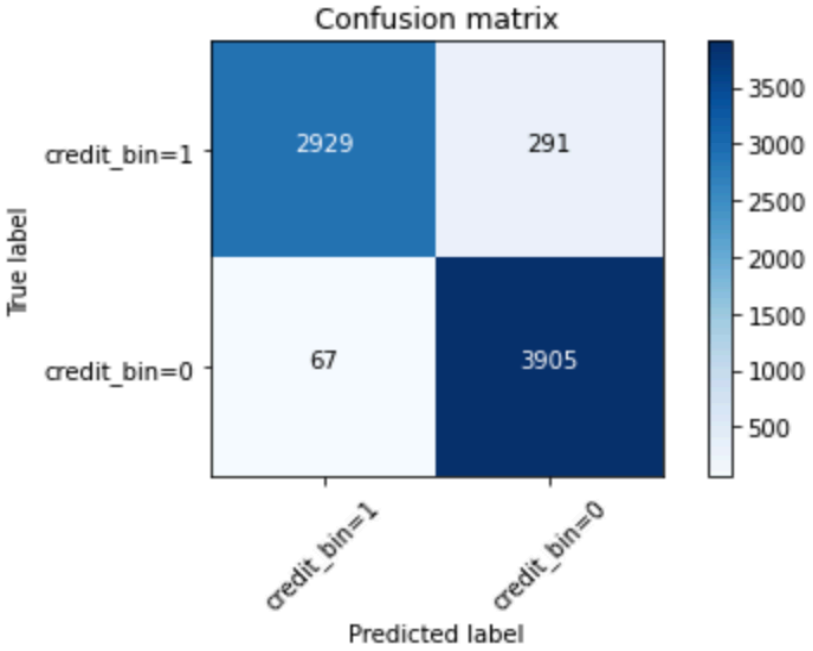
Findings- Confusion Matrix Logistic Regression(cont.)

- 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.

```
Confusion matrix, without normalization
[[2929  291]
 [  67 3905]]
```



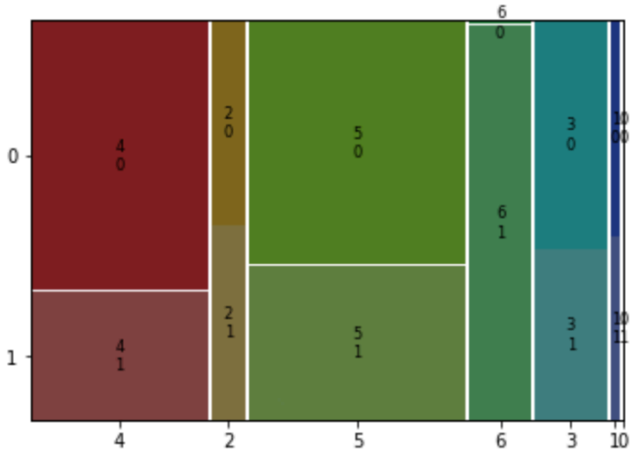
Findings-Mosaic plot

- 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.

```
from statsmodels.graphics.mosaicplot import mosaic
import matplotlib.pyplot as plt
import pandas

#gender = ['male', 'male', 'male', 'female', 'female', 'female']
#pet = ['cat', 'dog', 'dog', 'cat', 'dog', 'cat']
#data = pandas.DataFrame({'gender': gender, 'pet': pet})
mosaic(df3, ['fl', 'credit_bin'])
plt.show()
```



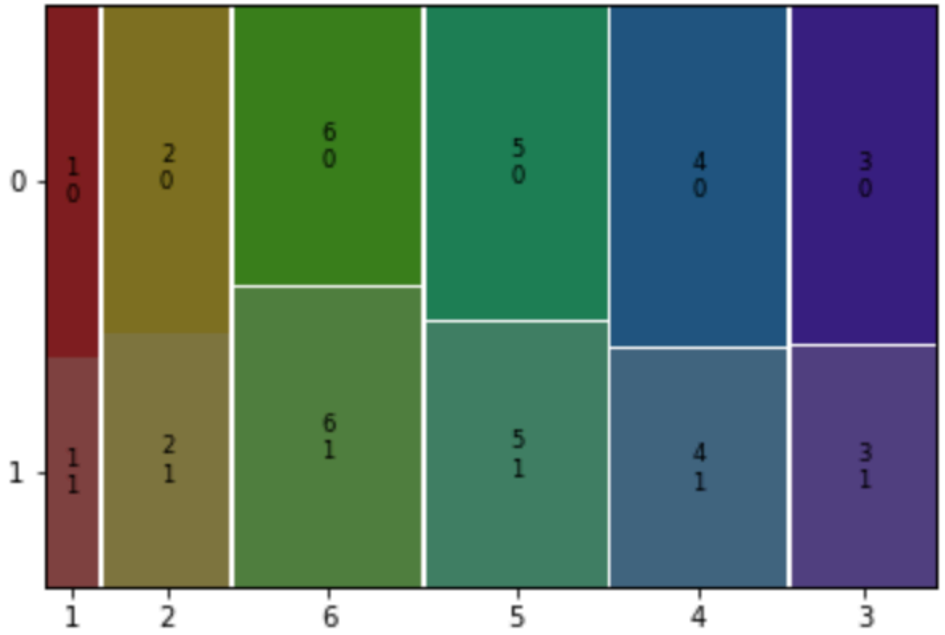
Findings-Mosaic plot (cont.)

- 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.

```
mosaic(df3, [ 'agegrp', 'credit_bin'])
plt.show()
```



Limitations

- These results are limited to the United states, we can not generalize them to the other countries.

- We used the most relevant techniques however; more advanced techniques could have applied to dive in further.

- We could have used dependent variables as different classes in addition to converting it to a binary variable.

Conclusions

- 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.

Acknowledgements

Where did you get your data? Did you use other informal analysis to inform your work? Did you get feedback on your work by friends or colleagues? Etc. If you had no one give you feedback and you collected the data yourself, say so.

References

If applicable, report any references you used in your work. For example, you may have used a research paper from X to help guide your analysis. You should cite that work here. If you did all the work on your own, please state this.