Predicting credit card behavior of US citizens using Machine Learning approaches

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 addition to the exploratory data analysis . I used survey data sets provided by FINRA from 2009 to 2018. I used all the surveys conducted so far that include (108310) number of households in the US. I will examine whether we can predict credit card behavior of the participant based on their demographic, education and profession ? In order to do so, I utilized supervised machine learning techniques such as OLS Regression & Logistic regression, Unsupervised such as K-mean clustering and Deep learning such as neural network and random forest models within the framework of logistic regression to dive deeper in the analysis.

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.). Individuals are different and exhibit diverse behaviors, however a general prediction model could be used to assess one's behavior.

Problem Statement

Individuals are different and exhibit diverse behavior, however a general prediction model could be used to assess 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.

Problem Statement

Individuals are different and exhibit diverse behavior, however a general prediction model could be used to assess 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.

Research Question and Hypothesis

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?

Hypothesis

- Personal characteristics do not determine credit card behavior?
- Demographic indicators do not influence credit card behavior.

Dataset(s)

In this study I will use National Financial Capability Study dataset by FINRA. This study interviews nearly 27000 per survey year. The most recent survey was concluded in 2018 and the initial survey was started in 2009. With an interval of three years we have a total of 4 surveys for the years of 2009, 2012, 2015 and 2018. In this study I choose the most recent survey i.e., National Financial Capability Survey 2018.

This is an open source data and is available to everyone at the following address: National Financial Capability Survey Data. You simply need to fill in a declaration form to download the datasets.

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.

- Then I summed all the dimensions of DV and converted it to a binary variable to analyze further, I took the average 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.

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.
- Supervised Machine Learning
    - OLS Regression and Classification (Logistic)
- Un-Supervised Machine Learning
    - K-means clustering
- Deep Learning
    - Neural Networks

Methods

- I begin with the data exploration followed by Supervised, unsupervised and deep learning techniques discussed above.
- Firstly, I applied regression and Logistic regression then, K-means clustering. Since, my dependent variable is binary (converted for that purpose) I can not proceed with regression as the latter requires DV to be continuous.
- In the third step of data analysis I used two deep learning models such as neural network and random forest models. Since, my data is cross sectional and I’m using the NN in the logistic regression framework, here the most appropriate machine learning approach is neural network.

Findings –Mosaic Plot

Question: Are financial literacy and credit card behavior associated ?

76]:
```
##Mosaic plots
from statsmodels.graphics.mosaicplot import mosaic
import matplotlib.pyplot as plt

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

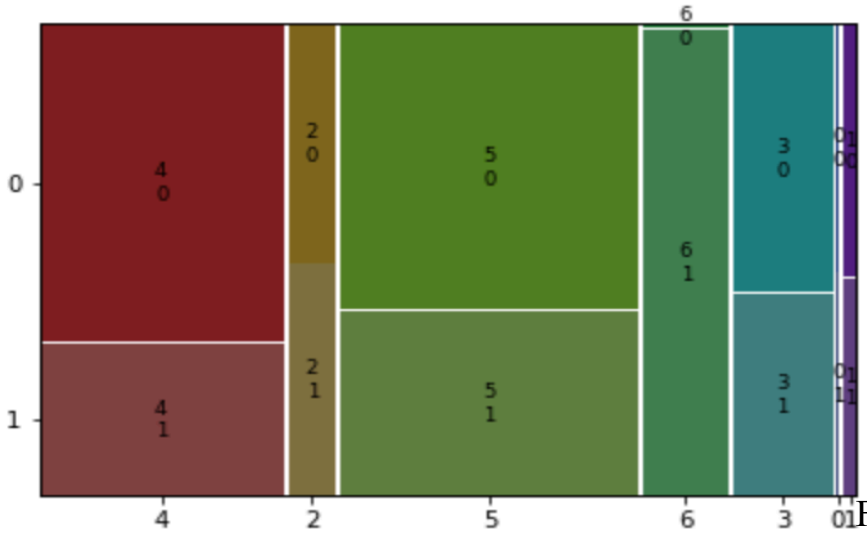


Figure: shows mosaic plot, here, I use mosaic plot to examine the association between my DV and other independent variables such as financial literacy. Financial literacy ranges from 0-6, as we can see higher financial literacy falls in category 1 meaning that the respondants with higher financial knowledge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

Figure: 1 Mosaic Plot-Financial literacy | Credit record

Findings –Mosaic Plot

Question: Does age group of the individual determine credit card behavior ?

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

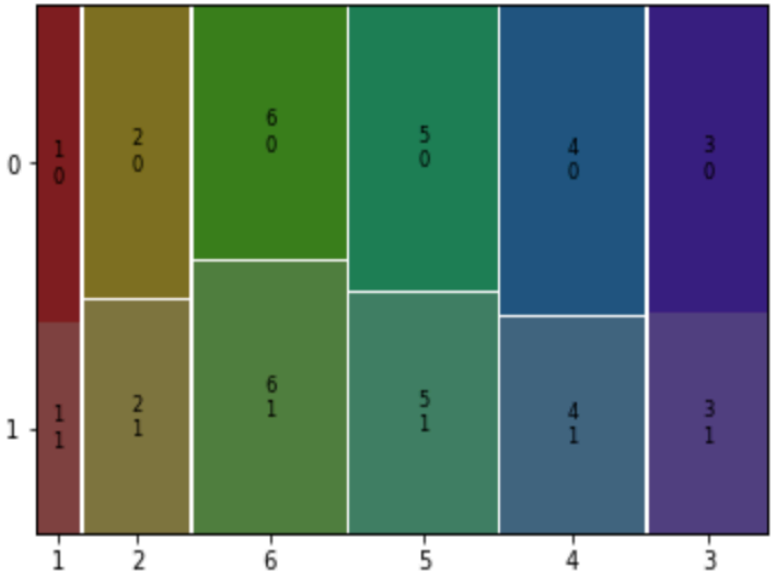


Figure 2 shows mosaic plot of age group and credit card record. Here, I use mosaic plot to examine the association between my DV and other independent variables such as age group of the respondants. Age group ranges from 1-6, as we can see middle age group from 4-6 have better credit record as compared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

Figure: Mosaic Plot-Age Group| Credit record

Logistic Regression (Supervised Learning)

[31]:
```
##OLS REGRESSION
y1 = df3['credit'].values
model = sm.OLS(y1, X)
```

[32]:
```
results = model.fit()
print(results.summary())
```

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.465
Model:	OLS	Adj. R-squared (uncentered):	0.465
Method:	Least Squares	F-statistic:	2116.
Date:	Tue, 16 Feb 2021	Prob (F-statistic):	0.00
Time:	11:26:04	Log-Likelihood:	-80350.
No. Observations:	36562	AIC:	1.607e+05
Df Residuals:	36547	BIC:	1.609e+05
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.0364	0.012	3.051	0.002	0.013	0.060
x2	0.0476	0.015	3.087	0.002	0.017	0.078
x3	-0.0505	0.012	-4.231	0.000	-0.074	-0.027
x4	1.9090	0.013	151.664	0.000	1.884	1.934
x5	-0.0317	0.015	-2.173	0.030	-0.060	-0.003
x6	-0.0833	0.014	-5.785	0.000	-0.112	-0.055
x7	0.1069	0.012	8.745	0.000	0.083	0.131
x8	0.0222	0.014	1.537	0.124	-0.006	0.051
x9	0.0227	0.015	1.477	0.140	-0.007	0.053
x10	0.1393	0.012	11.736	0.000	0.116	0.163
x11	0.2156	0.014	15.897	0.000	0.189	0.242
x12	-0.0060	0.013	-0.465	0.642	-0.031	0.019
x13	-0.0259	0.012	-2.114	0.035	-0.050	-0.002
x14	0.0647	0.012	5.394	0.000	0.041	0.088
x15	0.0341	0.013	2.649	0.008	0.009	0.059

Omnibus:	5561.893	Durbin-Watson:	0.334
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17784.900
Skew:	0.783	Prob(JB):	0.00
Kurtosis:	6.036	Cond. No.	2.85

Logistic Resgression (Supervised Learning)

[73]:
```
#LOGIT REGRESSION
import statsmodels.api as sm
logit_model=sm.Logit(y,Features)
result=logit_model.fit()
print(result.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.102938
         Iterations 9
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.850
Dependent Variable: y                AIC:              7557.2099
Date:               2021-02-16 11:47 BIC:              7684.8113
No. Observations:   36562            Log-Likelihood:   -3763.6
Df Model:           14               LL-Null:          -25140.
Df Residuals:       36547            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          -0.1469    0.0636  -2.3110 0.0208 -0.2714 -0.0223
agegrp          -0.0403    0.0279  -1.4467 0.1480 -0.0950  0.0143
ethn            -1.0825    0.0773 -14.0113 0.0000 -1.2339 -0.9311
edu              2.0926    0.0281  74.4324 0.0000  2.0375  2.1477
marital         -0.2047    0.0357  -5.7401 0.0000 -0.2747 -0.1348
hhincome        -0.2042    0.0201 -10.1455 0.0000 -0.2436 -0.1647
retired          0.7317    0.0965   7.5841 0.0000  0.5426  0.9208
prof            -0.0105    0.0154  -0.6798 0.4966 -0.0406  0.0197
profspouse      -0.0020    0.0158  -0.1272 0.8988 -0.0329  0.0289
currentstudent  -0.1148    0.0868  -1.3225 0.1860 -0.2850  0.0553
finsatisfacton   0.1619    0.0134  12.0593 0.0000  0.1356  0.1882
willingrisk     -0.1057    0.0132  -8.0304 0.0000 -0.1315 -0.0799
savingchildedu  -0.2236    0.0436  -5.1259 0.0000 -0.3091 -0.1381
finconfdaytoday -0.0720    0.0214  -3.3661 0.0008 -0.1138 -0.0301
fl              -0.1482    0.0275  -5.3937 0.0000 -0.2020 -0.0943
==================================================================
```

Regression (Supervised Learning)

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.

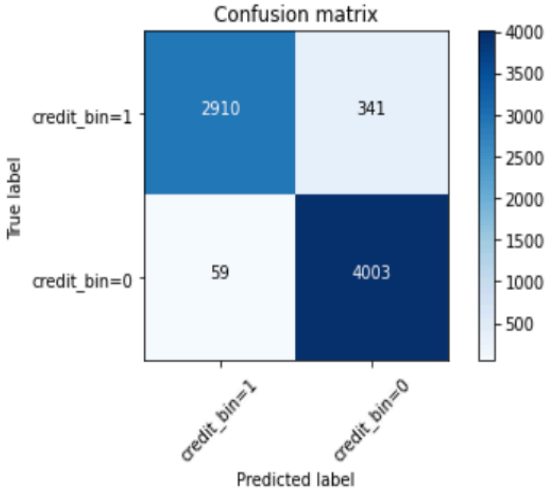
The binary variable 1 represents good behavior while 0 represents bad credit card behavior, gender 1 male 0 otherwise, age group is also a binary variable along with the list of independent variables explained earlier. I applied logit regression model by using Python package statsmodel.api. Results are 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.

Findings- Confusion Matrix Logistic Regression

[52]:
```
# Compute confusion matrix
cnf_matrix = confusion_matrix(y_test, yhat, labels=[1,0])
np.set_printoptions(precision=2)


# Plot non-normalized confusion matrix
plt.figure()
plot_confusion_matrix(cnf_matrix, classes=['credit_bin=1','credit_bin=0'],normalize= False,  title='Confusion matrix')
```

```
Confusion matrix, without normalization
[[2910  341]
 [  59 4003]]
```



Findings- Confusion Matrix Logistic Regression(cont.)

- This result implies that for 2910 respondants of the survey the actual credit behavior value in the survey is 1 in the test set and that the classifier also predicted them as 1, that is quite significant acheivement of the model.

- However, on the other hand, for 341 the actual value was 1 in the dataset but the classifier model predicted them as 0, which is not good sign, we may say that this error of the model.

- Similar interpretations can be used for the second row.

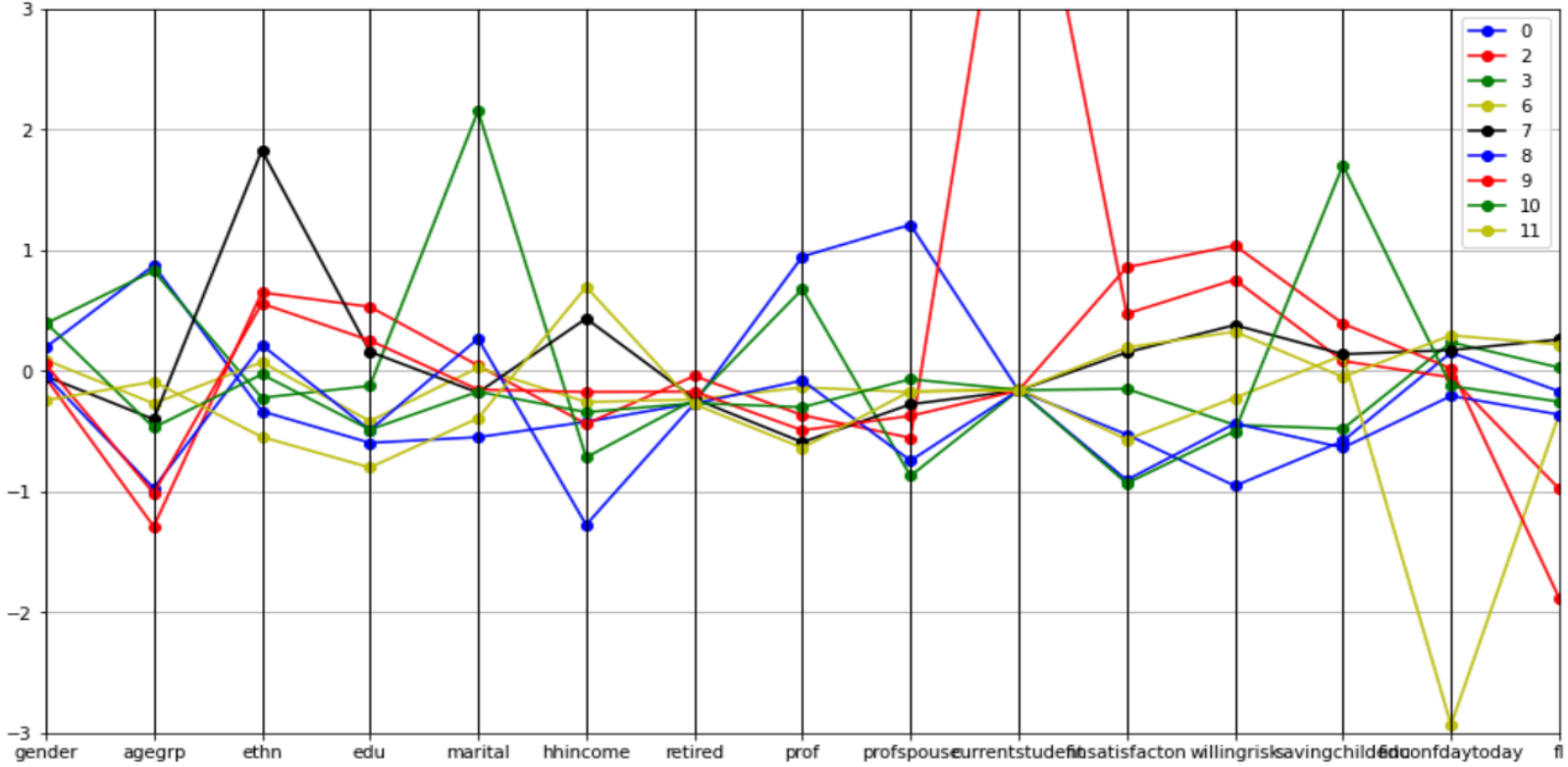
Findings- K-means clustering

[37]:
```
#now we test the accuracy
from sklearn import metrics
print("Train set Accuracy: ", metrics.accuracy_score(y_train, knn.predict(X_train)))
print("Test set Accuracy: ", metrics.accuracy_score(y_test, yhat))
```

```
Train set Accuracy:  0.9597251188074806
Test set Accuracy:  0.9453028852728018
```

Findings- K-means clustering

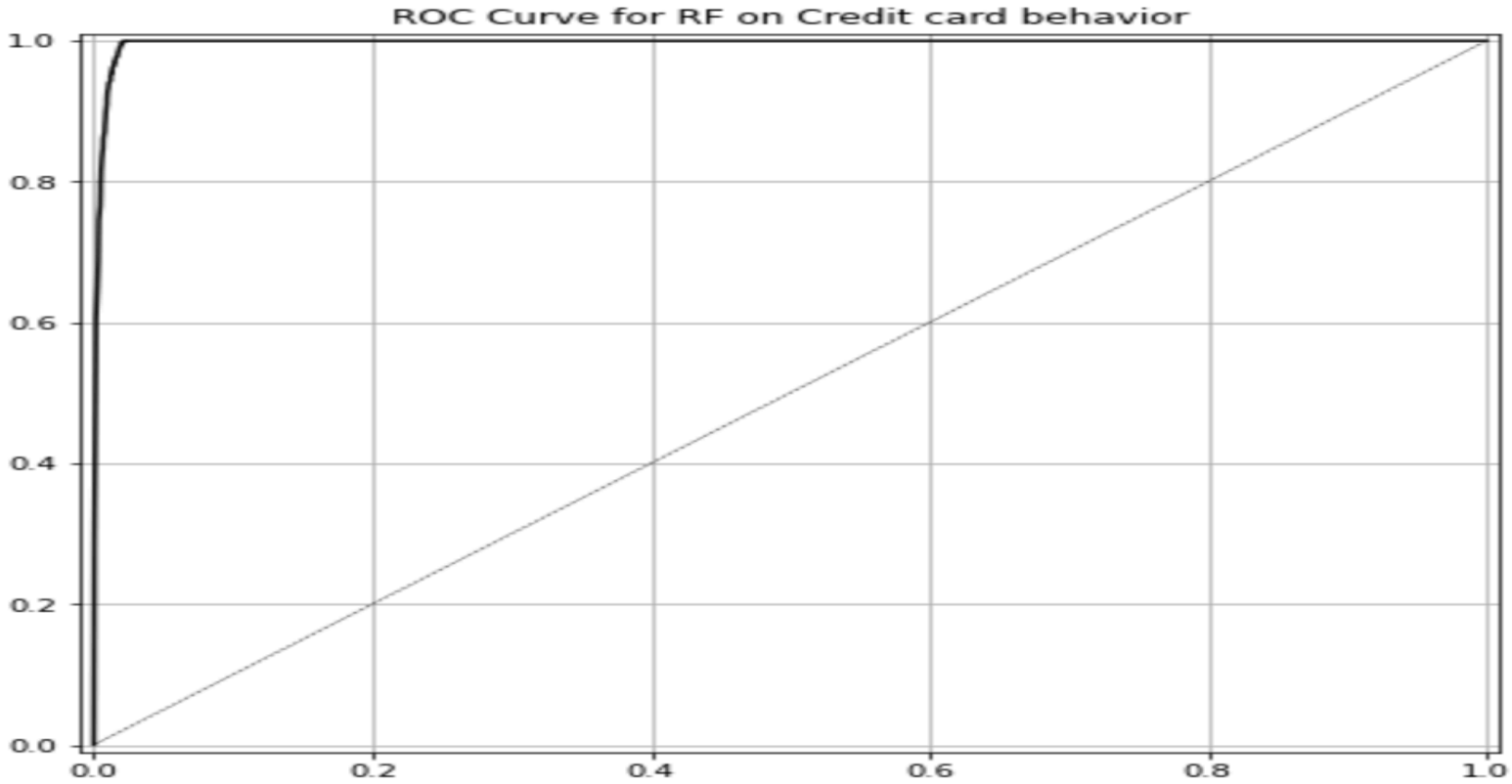
[83]:
```
parallel_plot(P[P['fl'] < 0.4])
#parallel_plot(P[(P['fl'] > 0.5) & (df3['prof'] > 0.4)])
```



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: 0.95
- Test set Accuracy: 0.94
- The accuracy of the model is quiet ok, as the accuracy for training set is around 95% and the accuracy for testing set is 94% that is acceptable according to standards.

Findings-Neural Network (Deep Learning)



accuracy is 0.988: roc-auc is 0.997

Findings-Neural Network

```
from keras.models import Sequential

model_1 = Sequential()
model_1.add(Dense(12,input_shape = (8,),activation = 'sigmoid'))
model_1.add(Dense(1,activation='sigmoid'))
```

[40]:
```
model_1.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 12)                108
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 13
=================================================================
Total params: 121
Trainable params: 121
Non-trainable params: 0
_________________________________________________________________
```

Findings-Neural Network

- As described earlier, I choose neural network from a wide range of deep learning approaches because neural networks are appropriate in prediction problems such as classifications and regressions.

- After doing all the necessary steps These are the main parts of MLP: Weights, Input layer, Hidden Layer, Weights, Net Input, Activation, I plotted the output in ROC curve that shows the sensitivity and specificity tradeoffs.

- In the above figure we can notice the curve is highly tilted towards left side, that indicates better performance of the underlying predictions with the accuracy of 98%.

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.

Reference:

- https://www.usfinancialcapability.org/downloads.php
- https://www.usfinancialcapability.org/