



CREDIT CARD DEFAULT PREDICTIONS



By: Jose Antonio Villegas
Coding Dojo Student
03/25/2022





Project Description

The purpose or goal of this project is to predict consumers who default on their credit card payments based on variables on data gathered.

Data consists of 13,444 observations and 14 variables.

Data Source: <https://www.kaggle.com/surekharamireddy/credit-data>

Load Data and Inspection

```
▶ path = '/content/credit_data.csv'  
df = pd.read_csv(path)  
df.head()
```

	CARDHLDR	DEFAULT	AGE	ACADMOS	ADEPCNT	MAJORDRG	MINORDRG	OWNRENT	INCOME	SELFEMPL	INCPER	EXP_INC	SPENDING	LOGSPEND
0	0	0	27.250000	4	0	0	0	0	1200.000000	0	18000.0	0.000667		
1	0	0	40.833332	111	3	0	0	1	4000.000000	0	13500.0	0.000222		
2	1	0	37.666668	54	3	0	0	1	3666.666667	0	11300.0	0.033270	121.9896773	4.8039364
3	1	0	42.500000	60	3	0	0	1	2000.000000	0	17250.0	0.048427	96.8536213	4.5732008
4	1	0	21.333334	8	0	0	0	0	2916.666667	0	35000.0	0.016523	48.1916700	3.8751862

Source : <https://www.kaggle.com/surekharamireddy/credit-data>



Description of Data

CARDHLDR	1 if application for credit card accepted, 0 if not
DEFAULT	1 if defaulted 0 if not (observed when CARDHLDR=1, 10,499 observations)
AGE	Age in years plus twelfths of a year
ACADMOS	months living at current address
ADEPCNT	number of dependents
MAJORDRG	Number of major derogatory reports
MINORDRG	Number of minor derogatory reports
OWNRENT	1 if owns their home, 0 if rent



Description of Data...continuation

INCOME Monthly Income (divided by 10,000)

SELFEMPL 1 if self employed, 0 if not

INCPER Income divided by number of dependents

EXP_INC Ratio of monthly credit card expenditure to yearly income

SPENDING Average monthly credit card expenditure (for CARDHLDR = 1)

LOGSPEND Log of spending



Cleaning Data

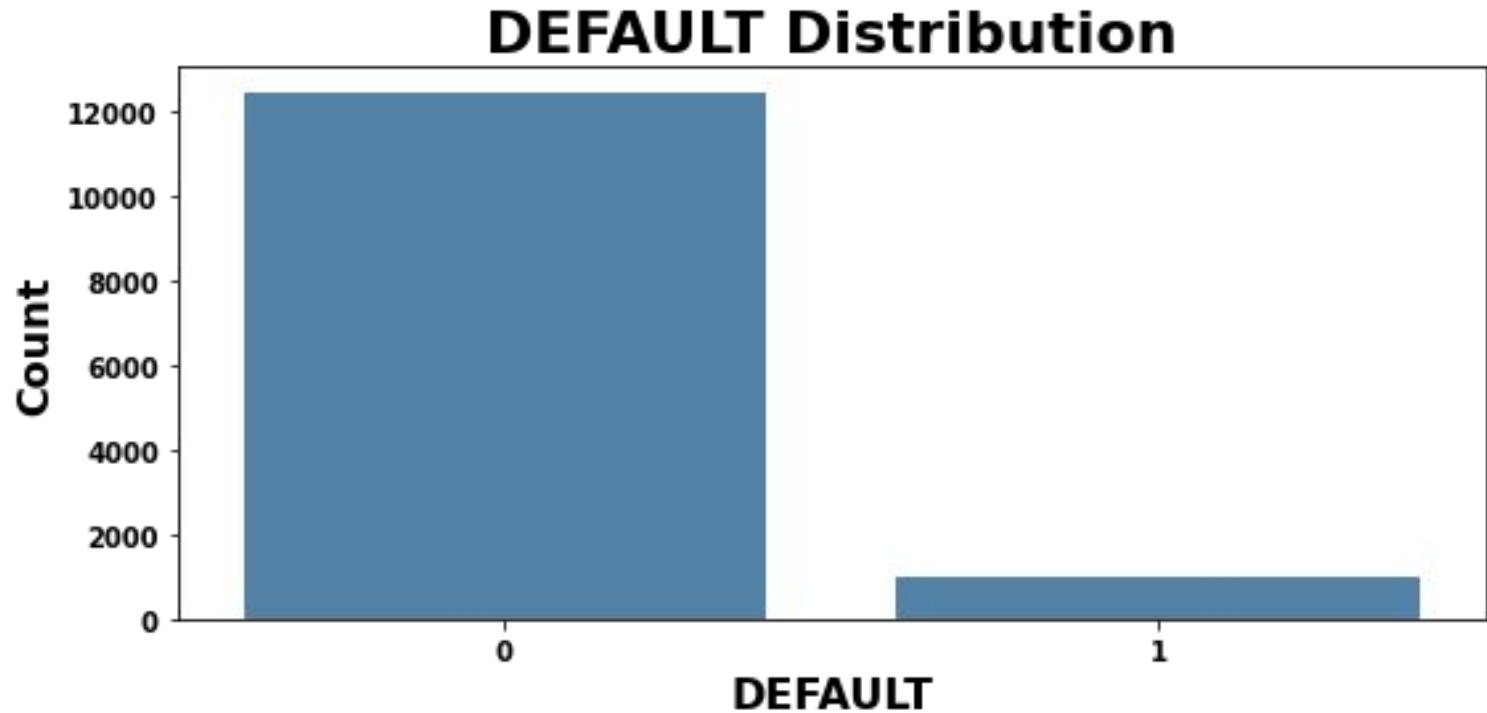
There are no zero (0) null values on the dataset

No duplicates were found

Dropped three (3) columns:

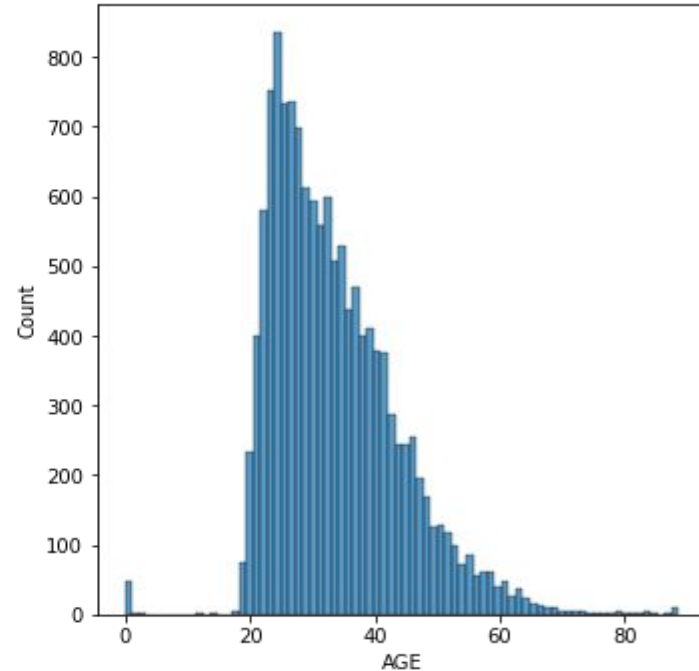
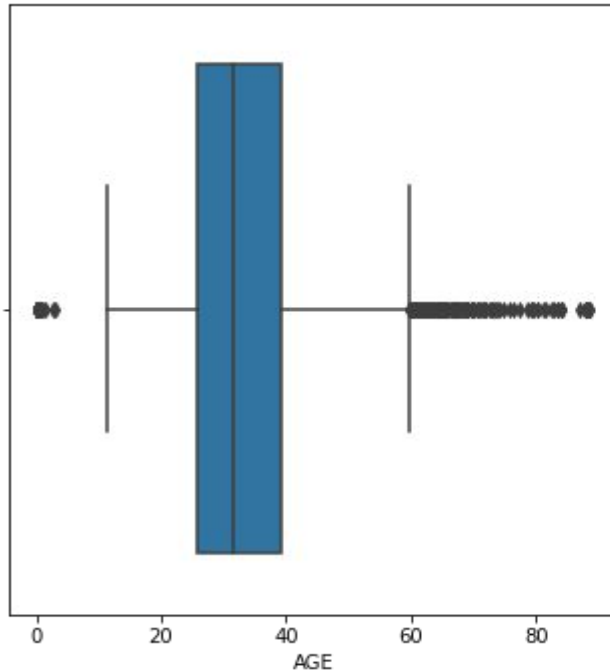
- CARDHLDR : As description states, this 'title' is given to person if credit card application is accepted. To be able to use this data, they should already be cardholders.
- LOGSPEND : 'Log of Spending' - not much information at to what this data is for.
- SPENDING : Again, this data column has information based on CARDHLDR application, accepted or not.


Initial finding on features of the data shows higher number probability of cardholders not to default on their credit card payment.



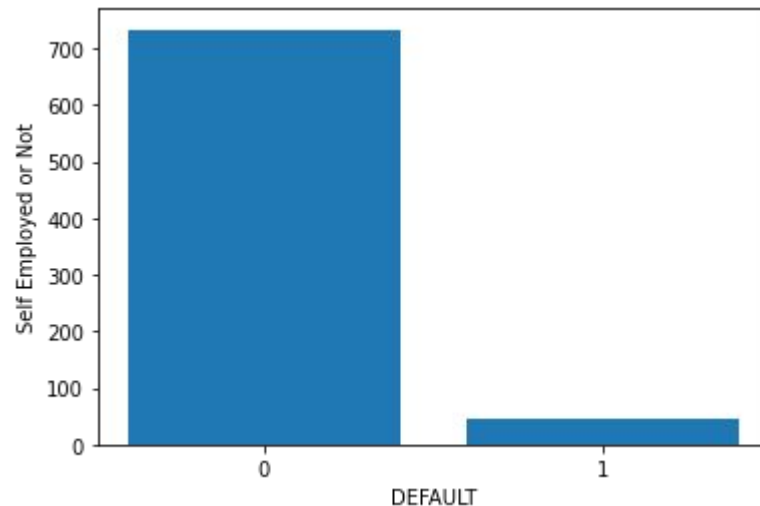
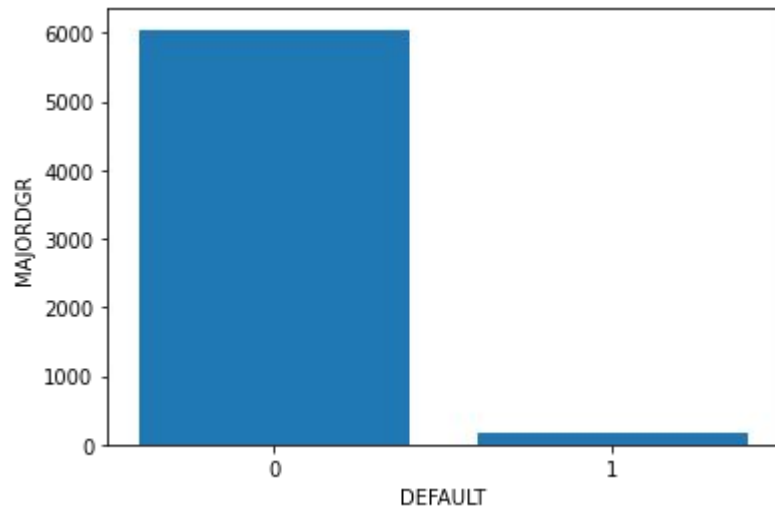
During initial analyzation of data, 'AGE' showed a number of outliers that had to be researched upon.

- Minimum age for card holders = 18 yrs old
- First credit card introduced 1950's
- Not much information from Kaggle as well.
- Decided not to drop

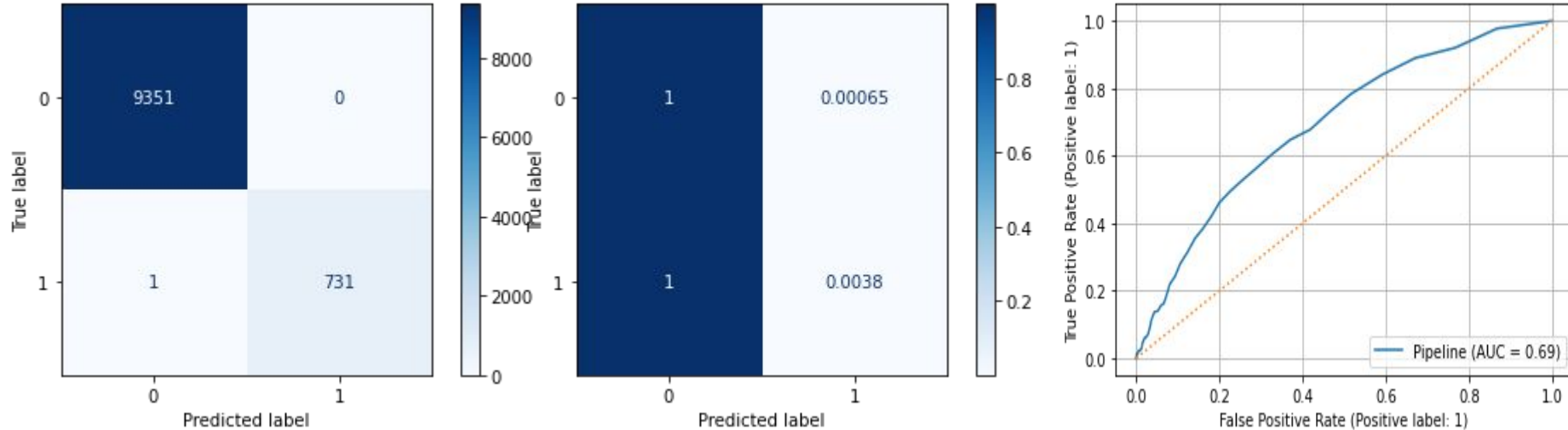




Most features of Dataset were imbalance which leaned most towards credit card holders not defaulting. These are just two samples from Major Derogatory Reports and Self- Employed or Not credit card holders.



Random Forest Model is recommended due to having the highest Train/Test Accuracy score. Though "Accuracy is not the best metric to use when evaluating imbalanced datasets as it can be misleading."





Conclusion :

The Dataset was not a good model for predicting default payment for credit card holders. Dataset had inaccuracies and uncertainty in the variables as stated in the description/content of the data gathered by the bank. The heatmap showed as well the weak correlation between features and the target.