Predicting Credit Card Approval of Customers using Machine Learning Algorithms

CIND 820 – Big Data Analytics Project, Module#3

Supervisor: Dr. Ashok Bhowmick
Date of Submission: March 07, 2022

Prabhkiran Kang
Student Number: 501068149

Ryerson University

Initial Results and the Code

Data Loading and Data Preparation

From the different programming languages available for data analysis, language of my choice is Python as it has number of packages that can be used to accomplish big data analysis project.

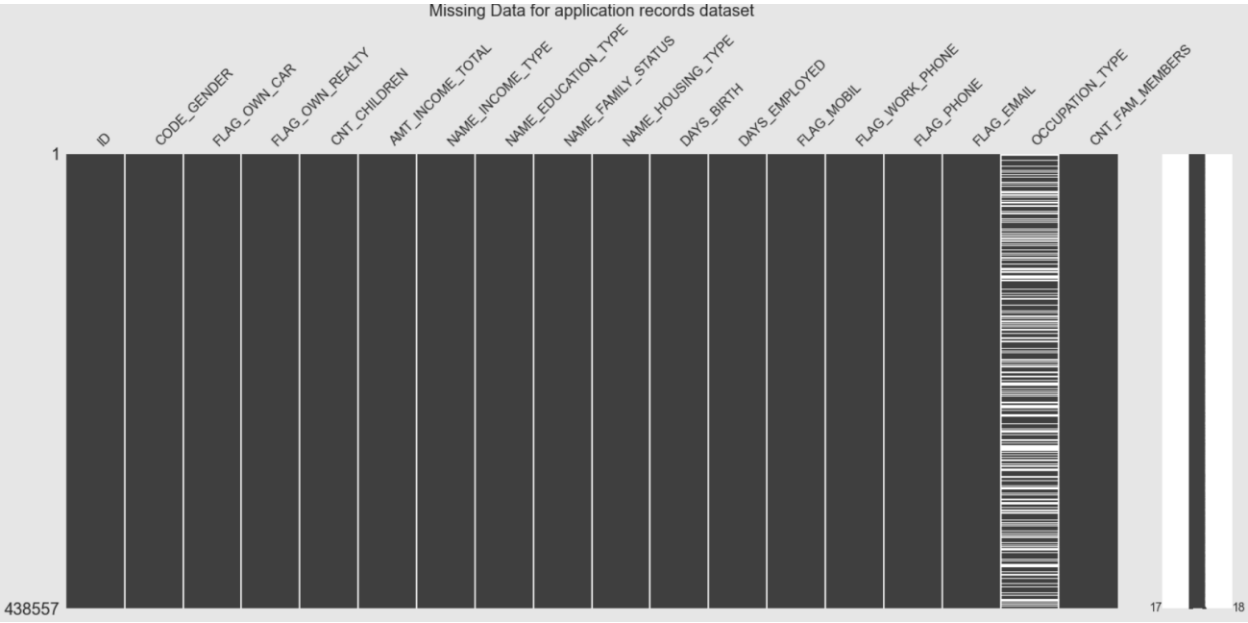
First step to complete this module is to import the number of useful packages, including *Numpy, Pandas, Sklearn, Matplotlib, Seaborn* to prepare, clean, build and plot the dataset.

In the next step, I loaded the dataset that includes 2 csv files using *Pandas* package in python. I named the application record as app_data and credit record as credit_record.

Data Exploration and Visualization

I started the data exploration and visualization process by exploring each csv file individually and then merging the 2 files.

I. While exploring the application record csv file, I found that number of unique clients and number of rows are not equal, which means there are duplicates. While checking the missing values, I found Occupation_Type feature has lot of missing values.



II. While exploring the credit record csv file, I found that number of unique clients are less than # rows. While checking the missing values, I found no missing data.

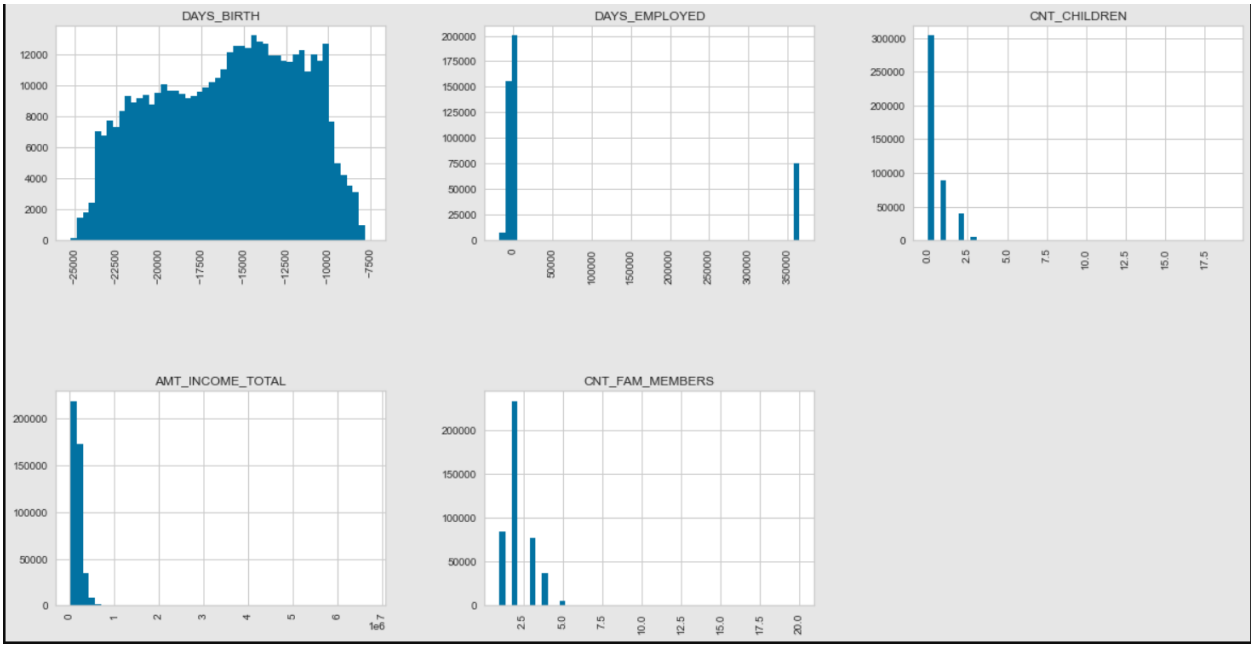
III. After exploring each csv file, I explored the categorical features of the application record csv file.

While exploring, I found out:

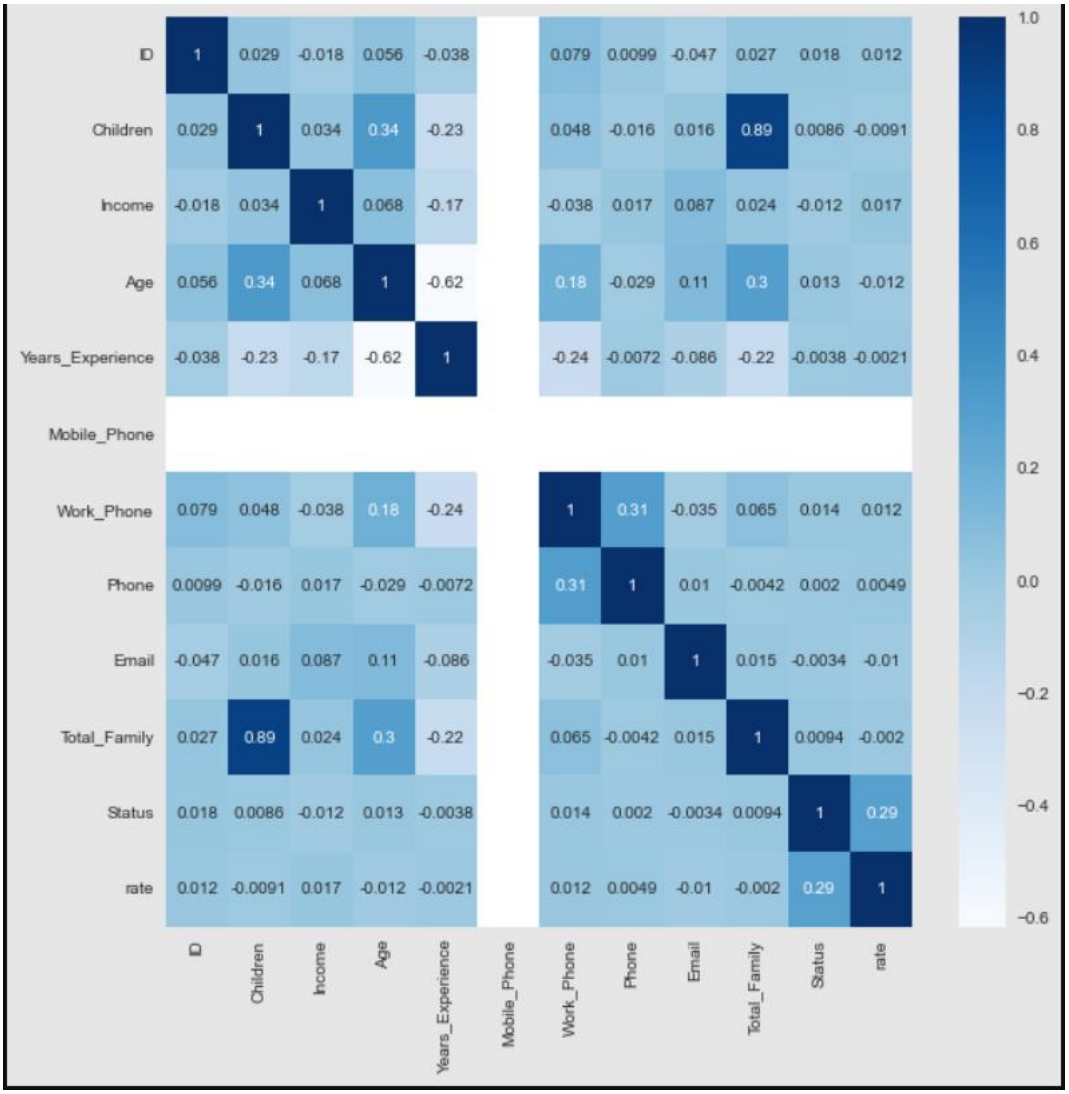
- The feature FLAG_MOBILE has only one unique value.
- The feature NAME_HOUSING_TYPE is very dominated by a single category.
- In the feature NAME_INCOME_TYPE, the category 'Student' has very few observations.
- In the feature NAME_EDUCATION_TYPE, the category ‘Academic Degree’ has very few observations.

IV. Next, I explored the numerical features of the application records csv file. My findings were:

- The numerical features should be scaled.
- To avoid the effect of the outliers, we can use a machine learning model that works well against the outliers (i.e. Random Forest).
- The outliers in DAYS_EMPLOYED are not actual days, but an indication that the client is unemployed



V. In the next step, I merge the 2 csv files into one and named it final_data. I renamed the columns of the final_data dataframe. I performed correlation of the features of the final dataset. I found out that Children feature has high correlation with Total_Family, thus I would remove it. I would also remove Mobile_Phone feature.



VI. In the Next step, after dropping Children and Mobile_Phone features. I performed another correlation. I observed that Income and Age plays a big factor.

Data Preprocessing

In this part, I preprocessed data on the basis of observations made during data exploration and visualization. I handled the missing values, numerical values, and categorical values.

After which, I got a final data. I split this final data into test and train datasets.

I performed Over-Sampling using Adaptive Synthetic (ADASYN) Algorithm and scaling to the datasets.

Thus, my dataset is ready for machine learning algorithms.

Machine Learning Algorithms

In this module, I only applied 1 machine learning algorithm, i.e. Logistic Regression Model.

I got accuracy score of 0.87.

I believe once I apply another iteration to this model and apply other machine learning algorithms in the next module. I would get a better result and a clearer picture.

GitHub Link: https://github.com/prabhkang1/CIND820-2022