Final Project: Credit Card Approval Predicter

Xiangbo “David” Wen

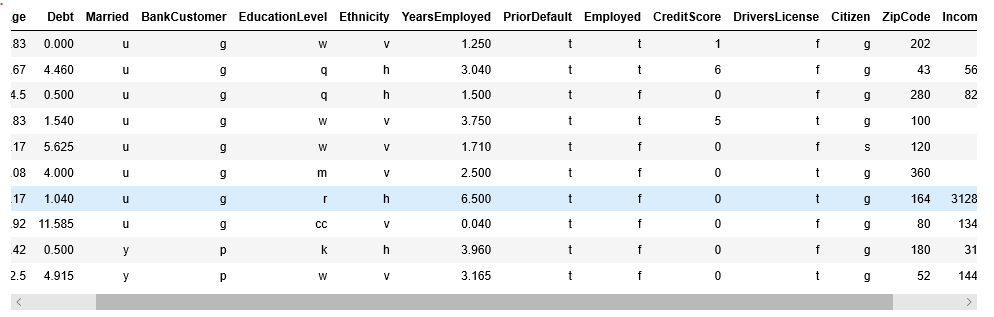
**Introduction**

The direct personal motivation for project is build a predicter to determine whether my credit card application will be approval or not. One of shocking facts of me is that I have never apply for any credit card in my 28 years of life. Based the extreme conservative economic education I inherited from my traditional Asian parents, any format of borrowing and spending money from future is cheating and shameful. Recently, I finally decide to apply for the very first credit card from Chase bank, and I am still waiting for my application result.

For other credit card appliers, I want this predicter can help more people understand what the critical factors are affecting our credit card approval rate. For these commercial banks that received large number of credit card applications each day, this predicter could reduce manual analyzing with power of machine learning, which has adopted by every commercial bank.

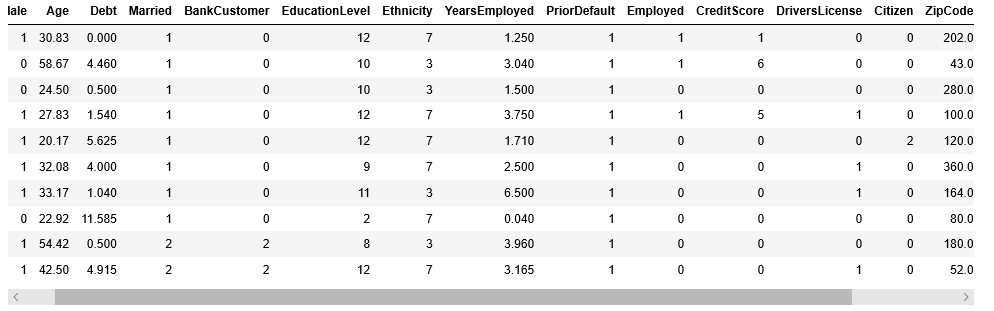
The data used in this project is retrieved from UCI machine learning repository, named *Credit Approval Data Set.* This dataset contains 690 instances and 15 attributes, with data types of categorical, integer and float. The 15th attributes show the result of application, where “+” means approved and “-” means denied. The first 14 attributes described personal information of appliers, classification labels include: gender, age, debt, marriage, bank, education level, ethnicity, employed years, prior credit card default, employed or not, credit score, drivers license, citizenship, zip code, income.

The original view of this dataset (partial view):



The challenges for data cleaning and preparation focus on replacing missing value and convert categorical value into numeric. To protect diversity and integrity of data, numerical missed values are replaced with mean value in its column, and categorical missed values are replaced with the most frequent value in its column. With machine learning Label Encoder function from Sklearn, all categorical/object type data transformed into labeled integers.

Here is the dataset after cleaning:

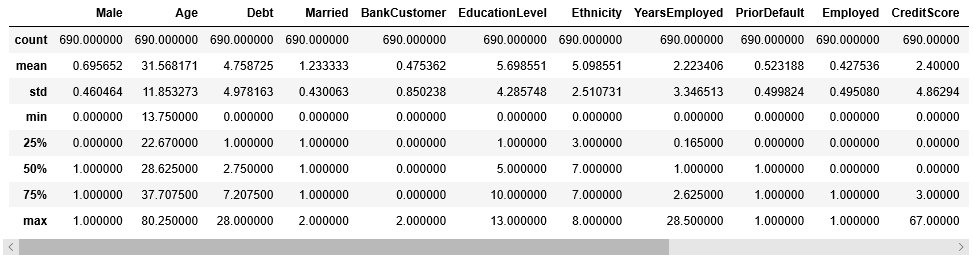


One downside of this dataset is that all values have been converted to meaningless symbols to protect confidentiality of the data.

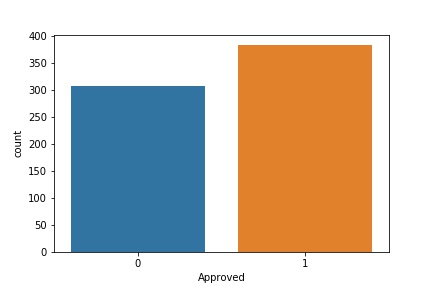
**Data analysis**

Let’s have some basic understanding of this dataset before we dive deep into data mining.

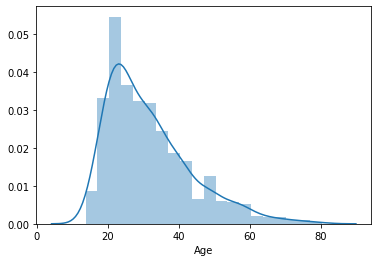
The basic statistics for this dataset are a little bit hard to observe, due to the large number of attributes and many of those attributes are binary data.



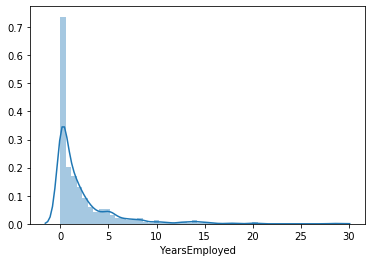
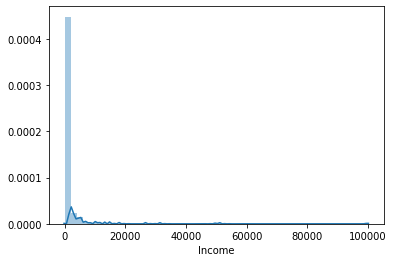
For better observe, a few key attributes are picked and made into plots:

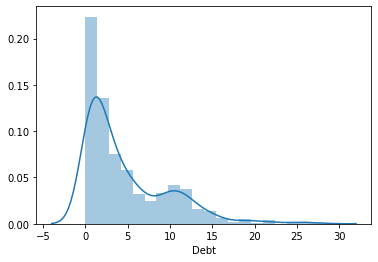
This figure shows the ratio of approved and denied, where 0 represent approved. This ratio is relatively balanced, which provide a balanced data for analyzing.

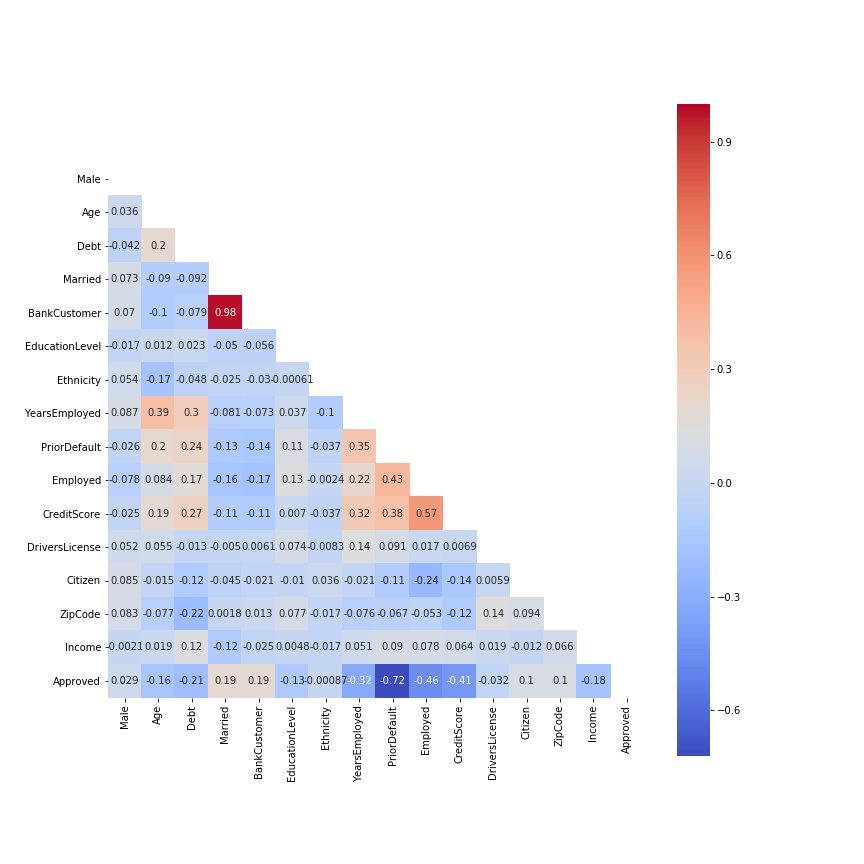
The distribution of Age shows this dataset is primarily focused on people from 20 to 40 years old, which corresponds with the research done by the world bank that people in the age group of 20s-30 leads the number of credit card applied and owned in worldwide.



Due the age of this dataset, it is easier to understand that the young age group won’t have long years being employed and won’t have much income with their early years of career.

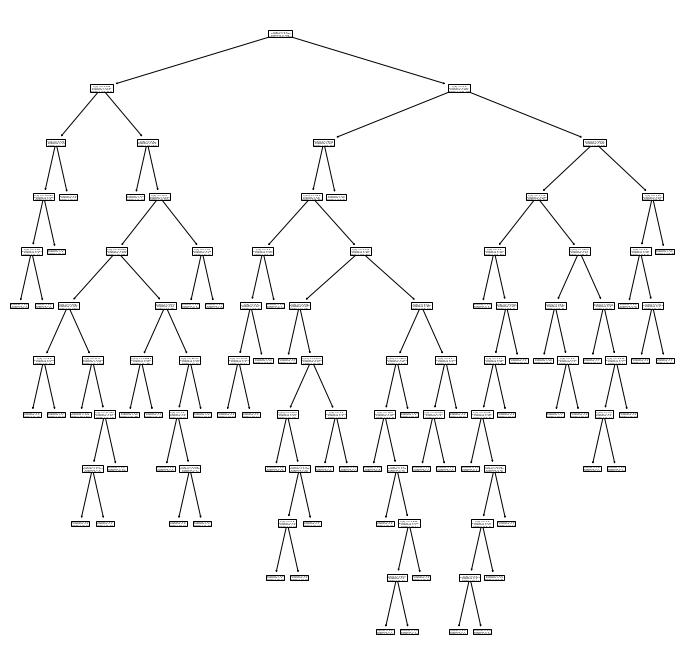


When people are young and have unlimited energy and desire, due to lower income in life cycle, young people’s debt in the relative high compare to the rest of their life. Research shows young generation in America and worldwide has more debt than any other generations in the past. In America, high college tuition fee and student loan contribute a large part of this debt. 

When put all these attributions together, we can clearly observe correlation among them. But correlation does not mean causations, where we need deeper analysis. 

**Methods**

In the project, 3 models are built for finding the best classifier with best accuracy in prediction. The first model is logistic regression, which is designed for finding the probability of a binary dependent outcome. In this project, one application of credit card can either be approved or denied, which is perfect suit for logistic regression.

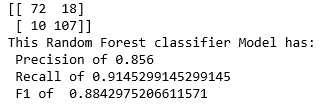
The second model is formed by decision tree classification that also suitable for binary outcomes. Decision tree work by splitting the dataset at series of nodes that eventually segregates the data into the target variable. Eventually, it will form a tree-like decision structure with layers of nodes determines the path of decision. The decision tree of this project is complicated but could help us with better understanding than logistic regression. 

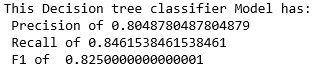
The third model is random forest, also called random decision forest. Random forest is an ensemble learning method for classification with constructing a multitude of decision trees and average results. Therefore, with random forest, the risk of bias is lower and accuracy with large dataset is higher compare to decision tree classification.

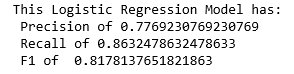
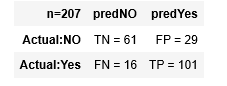
All models are fitted with 70% of data as training data, and then tested by the rest 30% data in examine accuracy.

**Results**

In comparison among these three models, random forest brings the highest accuracy with 86.47%, where logistic regression has accuracy of 78.26%, and decision tree has accuracy of 79.71%. I also compared their confusion matrix, precision, recall, and F1:

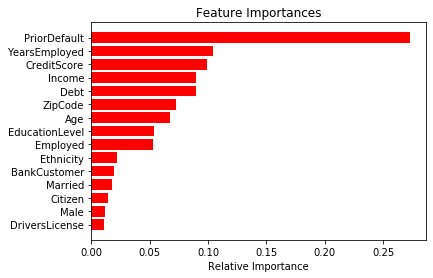


From all comparisons, random forest bring the most accrurate prediction model for this data.

In addtion, from the analysis in random forest, the importance of features are differanted:



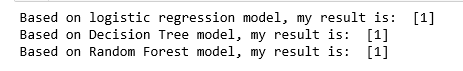
**Discussion**

From the results in modeling, clearly, credit default history is the most significant factor impacts credit card application. Followed by employment length, credit score, income and debt amount. In my understanding, the ability of repay credit debt or the risk of credit default is the key factor when bankers issuing credit card applications. People with history of credit default, low income, new in career life, own large amount of debt and low credit score will not have enough ability to repay their credit card in the short future. Therefore, they are unlikely get new credit card approved.

In the process of this project, I learned decision tree and random forest classification. These two are very useful in machine learning and data mining. I also practiced more functions in pandas and numpy.

**Conclusion**

In the end, I apply my own data with all three models I build in this project to find out the probability of approval if I apply for credit card. Unfortunately, all three models return result of deny.



But as declared in previously, all data in this dataset has been converted to meaningless symbols to protect confidentiality of the data. I can only guess the meaning behind the data for some part of features. For example, in gender section, there are two results “a” and “b” in the original dataset, which is impossible to determine which one means male or female. With this in mind, the results of this project are significant in determine the importance of each feature in impacting application result. This means, in case of gender, I can only tell the importance of gender is lower than education, but I can’t tell the difference in probability of male applicants against female applicants.