Progress Report

Team 12

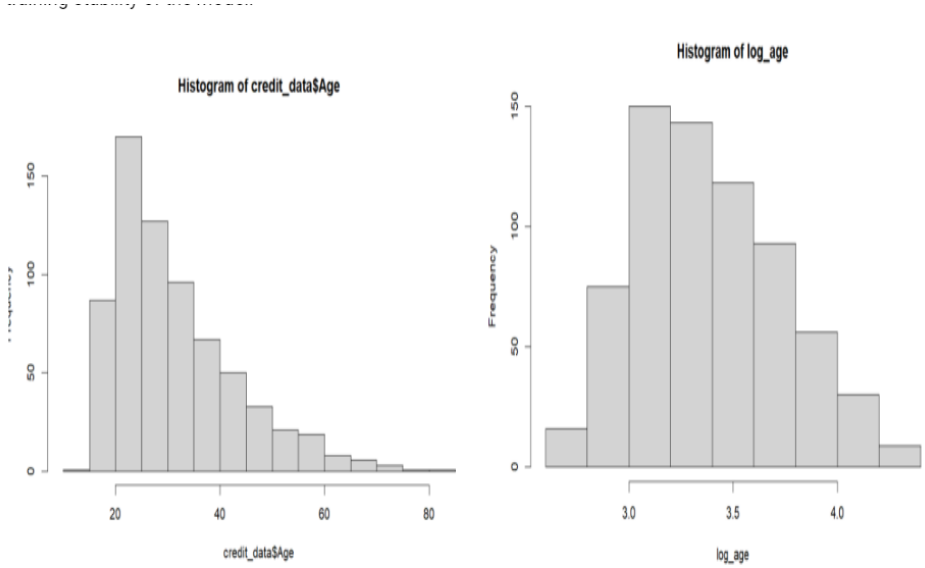
Topic: Credit Approval Analysis

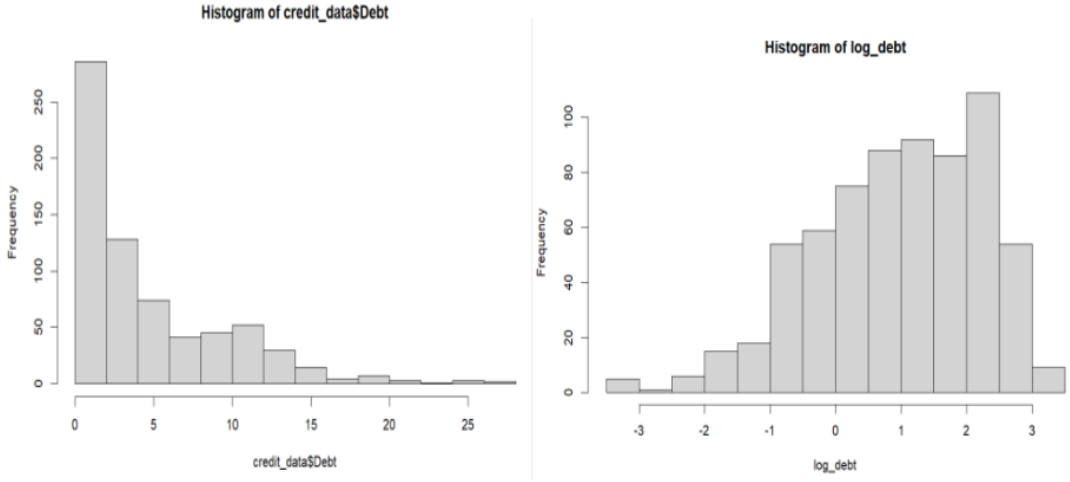
Credit is an important factor in our economy. Business needs credit to facilitate business activity, spend on capital expenditures, create jobs and thus indirectly stimulate the economy. Individuals need credit to fund major expenses, such as: education, real estates, autos, etc. Therefore, it is very important to facilitate lending in the economy without incurring excessive risks.

This is an interesting yet challenging model that requires frequent tuning and updates to ensure the proper risk management strategy in place. During the evaluation process, lenders look at socio-economic and demographic statistics of the applicant to make decision rules and minimize risk.

In our initial proposal, we laid out the 2 major questions that our analysis will focus on: (1) Identifying the significant factors that influence lender's credit decisions and (2) evaluating whether decisions made based on these factors models a reasonable risk management (by building a credit decision predictive model)

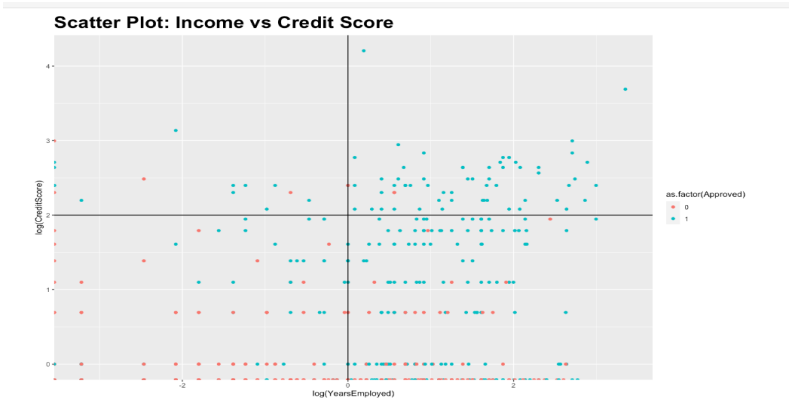
We started addressing the first question by exploring the dataset UCI credit approval. The dataset is clean and we did not need any imputation. We started assessing the characteristics of each independent variable and its relationship to the dependent variable. One thing we noticed is that all 5 continuous variables have right skewed distributions. This issue might cause the model to be unstable and inaccurate. We tried using log transformations and it worked well for some factors, such as Age and Income, but not for Debt and Credit Score (causing the distribution of underlying variable changes from right skewed to left skewed, coefficient of skewness changes from 1.485 to -.4)

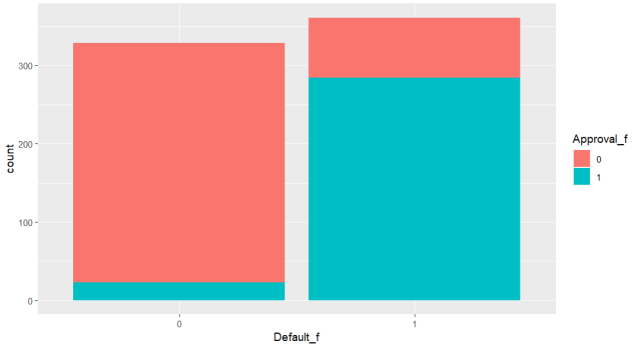




We are still exploring other options, such as: box-cox transform, inverse, etc.

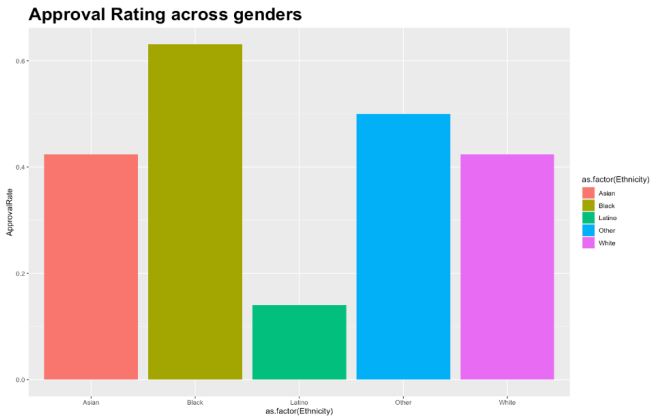
In our initial hypothesis, we proposed that prior default, employment status, credit score are the strongest predictors of whether an applicant will get approved. We also see high approval rates in the upper right corner of the below Scatter Plot: Income vs Credit Score. Those are instances in which applicants with both a high credit score and income are the most likely to get approved. We also recognize that applicants w/ prior default history are rejected more than 90% of the time versus 20% for those with no prior default.

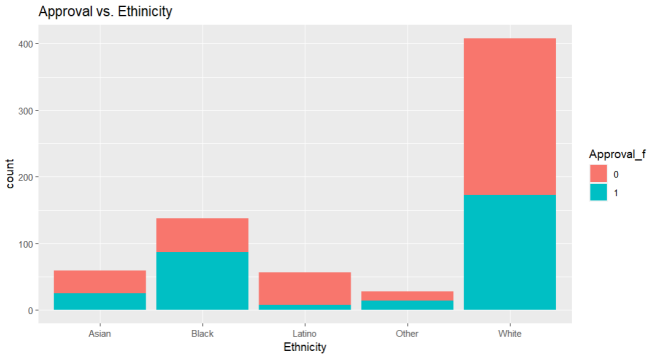




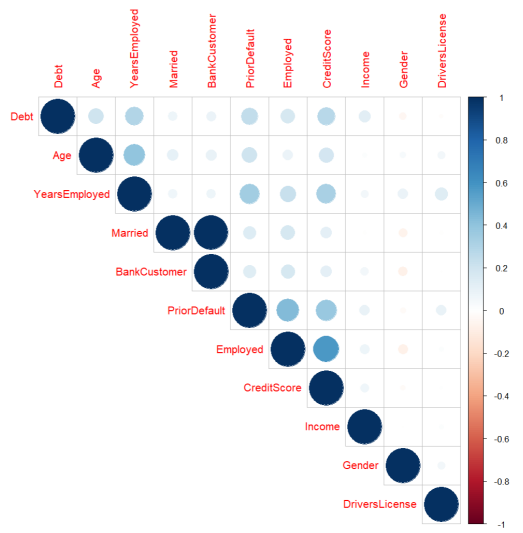
We plan to prove this hypothesis by either the actual coefficient and P-value of the factors in the logistic regression model or the weight of the input in the neural network model.

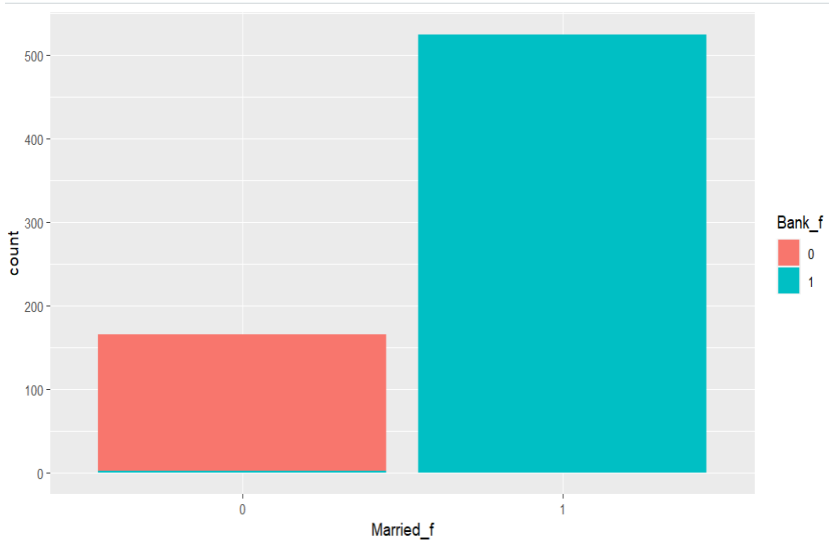
We also notice that ethnicity might be an interesting factor to look at. The below graph shows a low approval rate for Latino among other ethnicities but if we look at the entire dataset, the number of applications for Latino also is the lowest in the dataset. We are wondering if the low approval rate truly represents the industry norm or due to the small size of the dataset.





Since we are still working on transforming the data, we also started looking into the correlation between variables. As you can see, “ Married” and “ Bank Customer” are highly correlated. Almost 100% of applicants who are married will be the bank customers so we are considering eliminating one factor. By building a simple logistic regression using all provided attributes, we also see VIF scores for 2 factors: Bank Customer and Married are extremely high.





Since we are still working on transforming the variables, we haven't been able to do the PCA knowing that multicollinearity exists within the dataset and there are too many variables. We have 6 continuous variables, 6 binary variables can be splitted into 12 variables, 4 category variables that we are still researching on solution (Zipcode is masked data- will need to omit, Industry has 14 values- highly inaccurate due to its data collection method, and Ethnicity can be split into 3 variables).

Once we finalize on the transformation of the dataset, we will perform PCA and build the logistic regression as well as the neural network. We will then compare the confusion matrix between the two models to measure the accuracy. However, since our 2 major questions are two types of analytics (descriptive vs predictive) we will need to evaluate and explain deeply which model we are going to choose to optimize. Logistic regression is high interpretability but low accuracy while neural network is higher accuracy but low interpretability.

Citation

Khaneja, Deepesh. (2017). Credit Approval Analysis using R.

Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." *Neural networks* 61 (2015): 85-117.