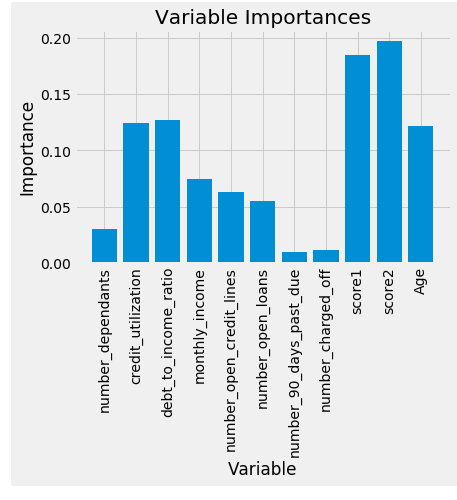
**Report of findings/insights from the given loan dataset:**

**Note:**

Please refer to the python notebook for the outputs, visualization plots and evidence

1. The training dataset has 20839 observations and 12 variables that includes target variable with 2 levels 1 and 0 which means default and non-default loans respectively.
2. Out of 12 variables, only one variable date\_of\_birth is of object type and all other variables are of numerical datatype.
3. Age variable is created using date of birth of the customer so as to make it as numerical.
4. Then, ID and date of birth variables were deleted leaving all the variables in the dataset as numeric in nature
5. A plot of histograms was plotted for all numerical variables to analyze the distribution of each of these variables.
6. The count of non-default loans is 18710 whereas number of default loans is 2129. This means around 10% of loans are defaulted.
7. The above distribution in the dataset is said to be imbalanced and hence it is required to under sample/oversample the data to make accurate/fair predictions on the new or test data.
8. I have chosen to under sample the non-default loans in such a way to make the ratio of default to non-default loans as 1:1
9. Then after, correlation heat map is plotted to determine the association of independent variables with target variable.
10. Outliers were identified for variables that are highly correlated with target variable and dropped those observations.
11. Five kinds of machine learning algorithms are implemented on the training data and used the cross validation approach to measure the model validation metrics.
12. Random forest classifier is used as the final model as it has reasonable auc and accuracy scores.
13. Feature importance of all the variables are plotted using the functions offered by the random forest classifier



1. Score1 and Score2 are the most important variables in predicting whether a loan is defaulting followed by credit\_utilization, debt\_to\_income ratio and age of the customer.
2. Further, score1, score2 and monthly\_income are negatively correlated with the target variable which means as score1, score2, monthly\_income increases the likelihood of defaulting of the loan decreases.
3. Finally, the random forest classifier is used to predict the loan default on test dataset as it has ability to overcome the problem of overfitting.
4. I would recommend to charge higher interest rates to the customers that have higher probabilities of defaulting the loan.
5. If a customer in the test dataset has high score1, score2 and high monthly income, then I would provide more flexibility to that customer in choosing the loan amount and loan term and I would offer lower interest rates.
6. If a customer has around 30-40 age, descent credit utilization and debt\_to\_income ratio, then I would offer him better interest rates, which are neither higher nor lower.
7. If scores of the customers in the test dataset are good and monthly\_income is higher than average, then I would prefer to offer them the refinancing option after few months of loan origination by providing lower interest rates.