Task 3 Course 2 Final Report

David Splawn

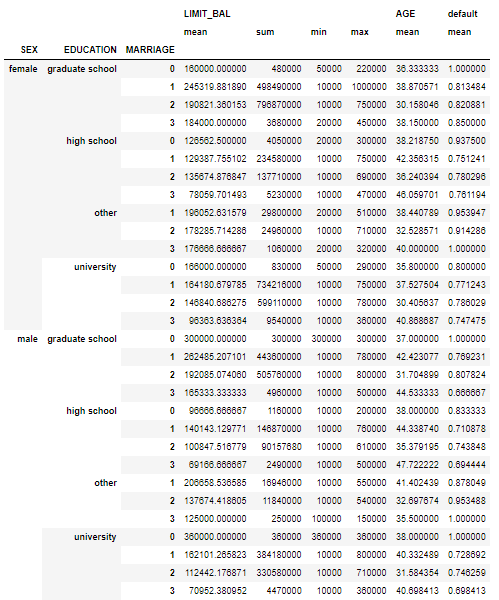
1. Cleaning Preprocessing

- delete duplicates - fix headers to meaningful info - understand what label encoding data means

2. EDA

In my analysis of the variables, I learned that this is a classification problem and not a regression problem. The fundamental question of will someone default or not results in a binary function.

The Demographic data yielding import insights and can be summarized by the table below:



3. Modeling Problem and Approach

The underlying problem of solving whether a customer will default or not is still somewhat cryptic because of lack of income and spending patterns. The variables and model building did provide an uplift in the model accuracy and will yield value for the credit provider. The problem at hand is a binary classification problem.

Machine learning algorithms such as linear regression, logistic regression, support vector machines, Random Forest, Naïve Bayes, K-NN, and ANN(separate jupyter notebook) of tensorflow in a different kernel were used to solve the problem.

The “PAY\_N” features were simplified into one feature seeing whether someone is paying or not.

pay\_features = ['PAY\_0','PAY\_2','PAY\_3','PAY\_4','PAY\_5','PAY\_6']

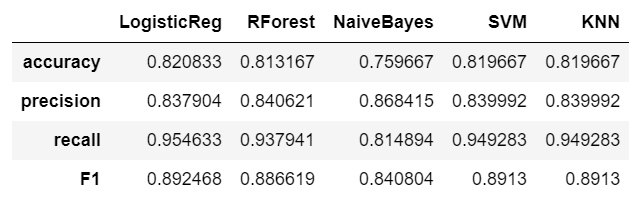
for p in pay\_features:

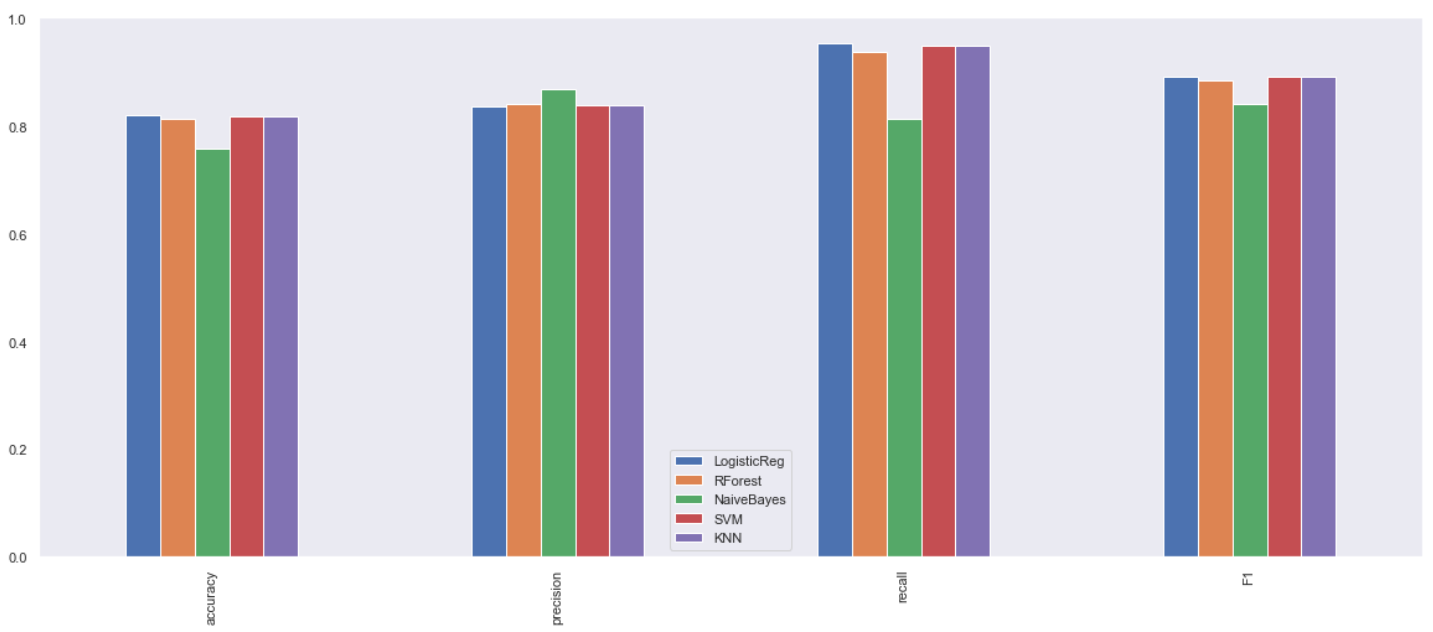
df.loc[df[p]<=0, p] = 0  
  
4. Model Building

The model was divided into 80% training and 20% testing and added a random state and stratification to eliminate order bias. After splitting the model, a standard scaler was added to eliminate scale biases, and it was introduced after the split to eliminate the biased from the test set since the test set should be 100% blind. Finally, a confusion matrix of the model results was created to understand the accuracy, precision, recall, and F1 of the models.

3. Result of Models

Using Linear Regression yielded poor results. The rmse for the prediction came out to .3762 and the Rquared value was .1783 suggesting this is not a linear regression problem.





5. Discussion

Linear Regression performed poorly for the model and a different approach had to be taken that classifies binary predictions better.

All the models did really well except for the NaïveBayes. The best model seems to be the Logistic Regression with an uplift of 82% accuracy. An Expect Monetary Value analysis can be done depending on time/current discount rates and using a Monte Carlo simulator to quantify the actual uplift in value.