Credit Approval Estimator Using Personal Data

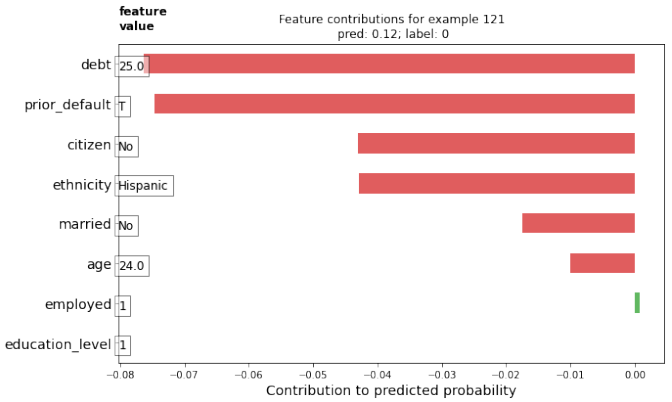
Matthew Cheng

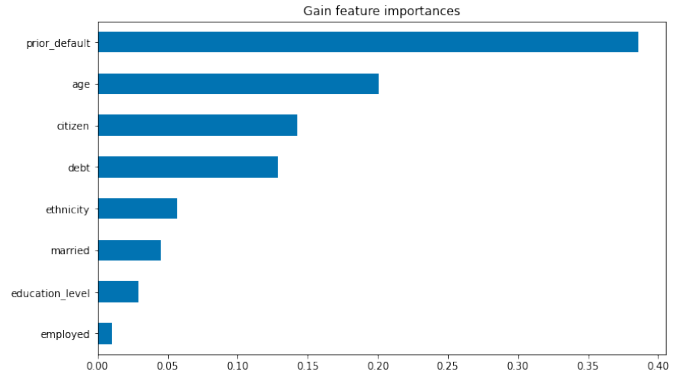
Problem Statement:

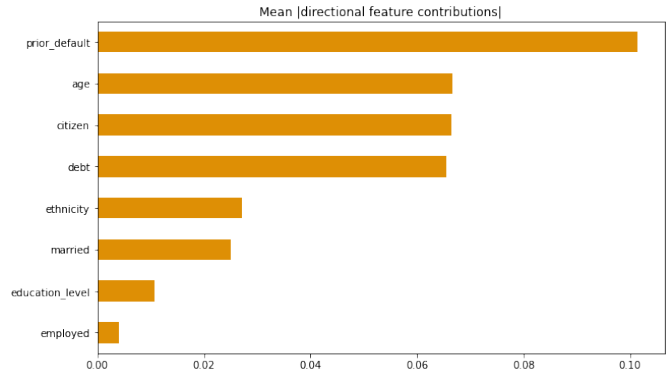
The purpose of this project is to determine what groups should be provided with better education about credit scores and their importance. Credit scores are important because they can determine what kinds of credit you can get and how much of it you're allowed. As we learned in our philosophy class, there can be many biases present in various algorithms, like a FICO score. These scores impact everybody in the country, so if biases are present, that is an important factor to keep in mind. Despite credit scores only being impacted by 5 factors, I want to look at other factors, like age, gender, ethnicity, and education level, in order to determine the aforementioned biases. Who suffers the most from the current credit score structure? Are different groups of people being impacted more than others? Hopefully my results will be able to provide insight so that we have a better idea of who to help educate about the issue.

Data:

Although I could not find a source of data that actually provided me with both personal data and credit scores, I was able to find a dataset that had credit approval as the target variable. This means that the target is categorical instead of discrete like I had planned. This is a relatively small dataset with only 690 observations, but it was the most detailed dataset I could find. There are 15 features including gender, age, ethnicity, and employment. In total that is 9 categorical and 6 discrete variables. I got this data from the University of California Irvine Machine Learning Repository and unfortunately, their method of collecting the data is confidential.







Methods:

In this project, I used a Boosted Trees Model, very similar to the one we used with the Titanic dataset, in order to categorize each individual as approved or not approved for various forms of credit. I did have to do some preprocessing and transform the categorical data with one-hot encoding. For my boosted trees model, I used 60 trees with a max depth of 5, as I found this gave me the best results.

Conclusion:

I found that after training, my model was able to produce an accuracy of about 83.1%. On the left a three charts. The first shows the feature contributions for individual 121 from the validation set, as I felt like it did a good job highlighting the importance of the various features. As you can see, the two features that are the most detrimental to the result are debt and having prior defaults. This individual is also a young unmarried hispanic that is not a citizen, which all also hurt the outcome of being classified as no approval. What did help was being employed, but interestingly, it did not help very much. This could be because of both students and retirees counterbalancing those who are employed. I also thought it was interesting that education did not have much of an effect in this individual's case, being a high school graduate. Looking at the importance of the features as a whole, we can see some similar trends. Prior default, age, citizenship, and debt all make up the most significant features. Features like zip code did not seem to have much of an effect in this case, and was not shown in the chart.