QMBU 450 – Spring 2021
Abdulhakim Özcan
Homework 3 Report

Predicting Voter Behavior

For this homework, I picked 6 variables that I was theoretically most interested in. These variables are gender, education, marital status, household income, race, and the residence of the survey participants. None of these variables are more than half of the time missing. However, all of them had some missing values with varying degrees. For the first step, the encoded missing values are replaced with NaNs for efficiency. Then, the new NaN values are replaced with the most frequent value in the corresponding column. There are multiple options to deal with missing variables. In this case, all the variables picked, have a relatively high number of non-missing values except “race” and “rural and urban residence” which still contain more non-missing values than missing values. Thus, it seemed valid to replace the missing values with the mode of each column. On the other hand, no data points (row) were eliminated not to waste any information. Out of the 6 variables picked, 3 features were encoded using one-hot encoding. The rest are considered to have natural and intuitive ordering to them and thus left untouched. In order to predict whether some turned out to vote, 4 different supervised machine learning models were utilized. These are Gaussian Naive Bayes, Random Forests, Support Vector Machines, and Logistic Regression models. Two of these models were designed using grid search to evaluate models for different combinations of parameters and find optimal parameters for a given model. Among these models, Logistic Regression seems to slightly outperform others with a cross-validation accuracy of just over 82%. Even though the accuracy scores found for each model seems good enough, they are not considerably high. Two steps can be taken to potentially improve the accuracy scores for models. The easiest step is perhaps to increase the number of features fed into the model. Although more data usually means a more accurate model, it may also have reverse impacts if there is too much noise in the data. In this case, also the second step, dimensionality reduction may prove to be useful because it can behave as a noise-filtering tool, keeping the signal and throwing out the noise.