CSSM 502 Homework 3

Data

At the beginning of the homework, I determined features that I want to use in voting prediction. Here are my selected features:

- Gender: *categoric variable*
- Marital Status: *categoric variable*
- Age: *continuous variable*
- Household Income: *categoric variable*
- Number of Person in Household: *continuous variable*

Data Preparation

- I remove the “refused”, “don’t know”, and “missing” values according to CSES documentation.
- One-hot encode categorical variables – I tried to convert categorical variables which are gender, education level, marital status, and income status into a column vector. Then I also changed column names accordingly.
- Since there are large number of features, I decided to dimensionality reduction. After that, I focus on two main technique. One of them is Principal Component Analysis and the other one is Linear Discriminant Analysis. First, I tried to use PCA however, when I compere results between PCA and LDA, I found that LDA’ s performance is better than PCA in that specific data set. Since we try to solve binary classification problem, number of features used in LDA is 1. So, its highly preferable in terms of complexity.
- For training, I split those with the care of stationary, and test data size is 0.3 of training data.

Model Development & Evaluation

After that, I try 3 main algorithms to come up result.

- Grafiend Boosting (boosting)
- Random Forest (bagging)
- Multi-Layer Perceptron

I train my test data with all the trained algorithms. Then I prefer to choose gradient boosting since it’s had highest AUC score