Nick Hemauer
PLSC 597
5th Oct. 2023

Homework 2

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZABHCA

1. The R file included has the results replicated. Below is the table that matches the table in the paper. Voter turnout is the target variable of interest.

```
Call:
glm(formula = vote18 ~ depression + female + age + educ + income +
    attend + married + unemployed + black + hispanic, family = "binomial",
    data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.161189   0.510613  -4.233 2.31e-05 ***
depression  -0.225730   0.127456  -1.771 0.076555 .
female      -0.286849   0.151325  -1.896 0.058016 .
age          0.044733   0.005358   8.348  < 2e-16 ***
educ         0.194480   0.054077   3.596 0.000323 ***
income       0.049025   0.025533   1.920 0.054845 .
attend       0.077902   0.054909   1.419 0.155973
married      0.106738   0.166158   0.642 0.520622
unemployed  -0.111092   0.194513  -0.571 0.567914
black        0.103711   0.236296   0.439 0.660733
hispanic    -0.263527   0.194898  -1.352 0.176335
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1315.6  on 1013  degrees of freedom
Residual deviance: 1110.5  on 1003  degrees of freedom
AIC: 1132.5

Number of Fisher Scoring iterations: 4
```

2. Training Accuracy

```
Logistic Regression Training Accuracy: 0.7108603667136812
SVM Training Accuracy: 0.8251057827926658
Random Forest Accuracy: 0.9943582510578279
Regularized Logistic Regression Training Accuracy: 0.3554301833568406
```

3. Test Model Fit. In this case, logistic regression performs the best on the test data as it has the lowest mean squared error and R2 value. But none of them are particularly good matches.

```
Logistic Regression:
Mean Squared Error: 0.28
R-squared: -0.23

SVM Regression:
Mean Squared Error: 0.29
R-squared: -0.26

Random Forest:
Mean Squared Error: 0.31
R-squared: -0.39

Regularized Logistic Regression:
Mean Squared Error: 0.29
R-squared: -0.28
```

4. The permutations vary significantly between models. The logistic regression permutations show that the most important features in the model are Education and Age. Theoretically, this makes sense with the target variable being voter turnout. Furthermore, those are the two features that showed significance at the 3* level in the original paper.