**CSCI 381/780 – Applied Data Science**

**Final Report**

# Project Abstract

The purpose of this study is to analyze how the turnover rate from past election may determine upcoming election results. Determining turnover rates is not a simple science, however by using machine learning techniques we want to create models that when presented with new data can predict with accuracy the turnover rates. In this project we use the General Elections dataset from 2018 By MIT Election Lab to conduct this study. This repository contains county-level demographics and past election data on the federal elections from 2012-2018. In this study we hypothesize that election turnover rates will not trend too much differently when compared to previous elections. Limitations that we have encountered is that the data itself provides demographic information based on the year of 2018 however provides us information from elections in 2012 and 2016. We approached this by first creating columns for the turnover rates from 2012 and 2016 using the Current Voting Age Population (CVAP) from 2018. This is assuming that the current age voting population has stayed the same from 2012-2016 for each county. We believe that this is a safe assumption to make as population does not drastically change to where it will affect the overall demographic in a few years. Only major external factors can be the catalyst for such major changes in numbers.

# Accomplishments

### What are the major goals of your project?

The long-term goal for this project would be to eventually with a dataset with turnover rates of many previous elections predict accurately the turnover rates on new data. This will also include the other demographic features so we can gain greater insight on to why turnover rates were at that given election.

### What did you accomplish?

Specific Objectives

We accomplished in creating models that perform relatively accurately with the given data. We also understood which of our features has more of an impact when creating our models. This led us to have more insight to see if our model is generalizing with out given data.

Major Activities

Since the Turnover rate we are calculating is done by using the numbers from 2018 we had to clean up the data by removing any extraneous calculations. We also interpreted if values were to be imputed or dropped.

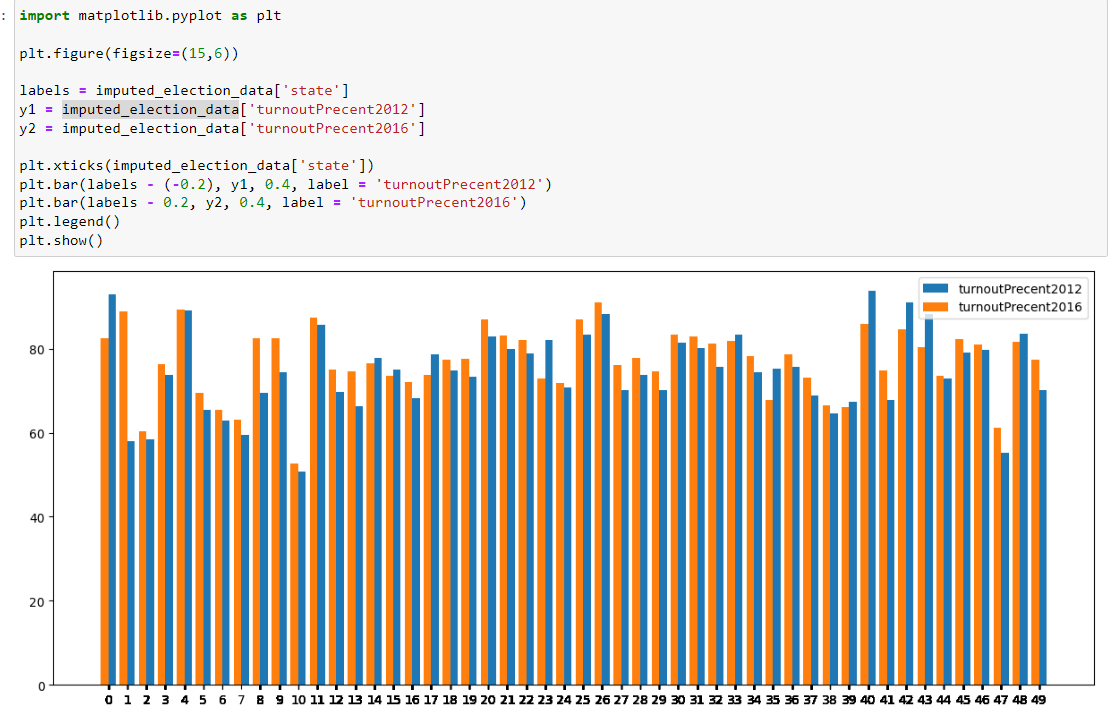
After data was cleaned up and eligible to be used for our models, we first displayed the trend with turnover rates from 2016 and 2012 using a bar graph to establish relationships before we began to create our models.

Afterwards we then created two models, a linear regression model and a random forest regressor model. Finally, to display our results we provided analysis using SHAP to see how the features we have affect our models.

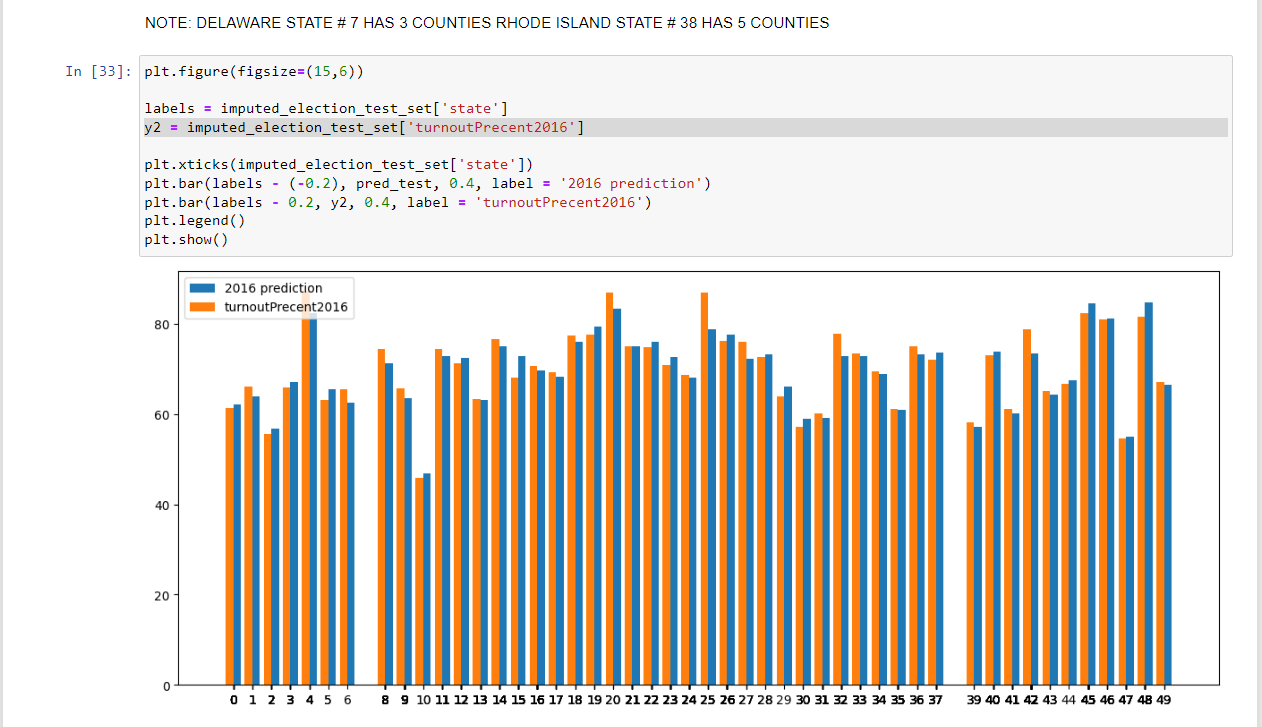
Significant Findings

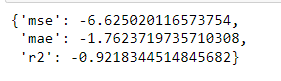
After creating the random forest regressor model the metrics that we got were a lot better than from when we used the standard regression. Our MSE was about 12 and MAE was 2 which is a lot better than the negative values we got from doing linear regression. Our r^2 value was 0.8 which shows a high correlation in our data.

Our initial graph to determine if there is a relationship before creating models

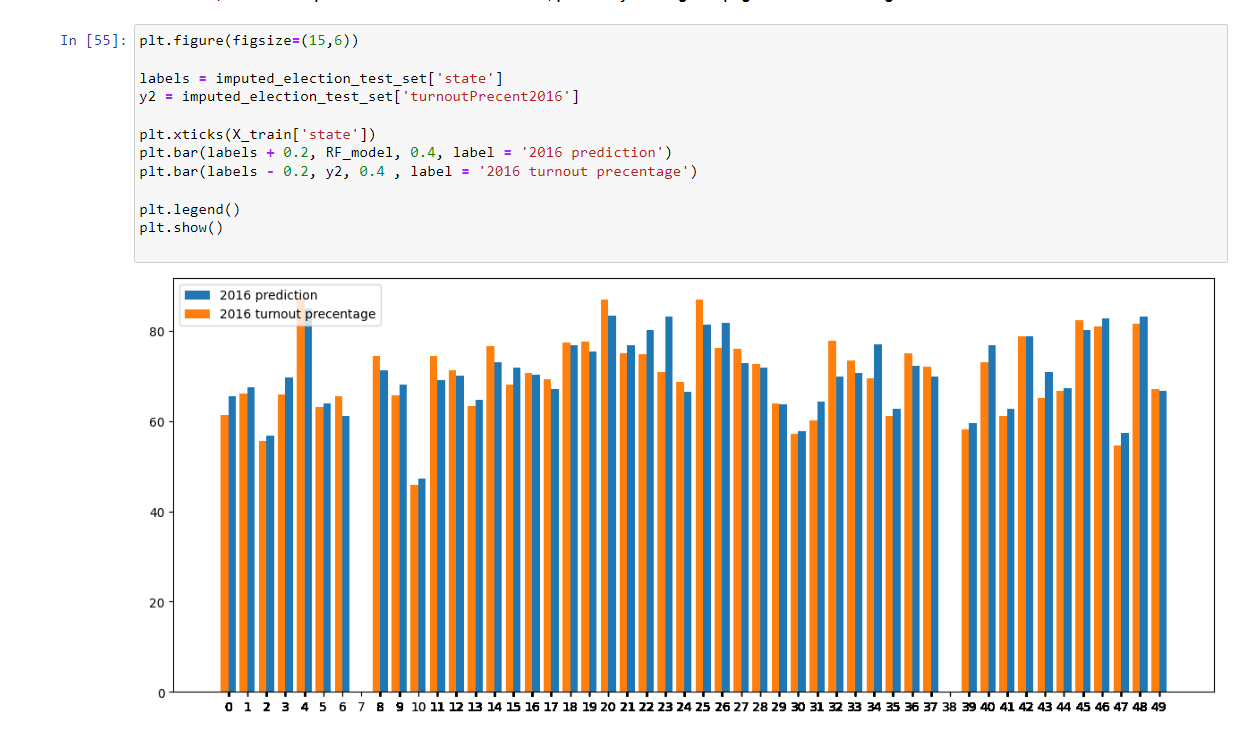


Our models predicted data vs Turn over rate in 2016



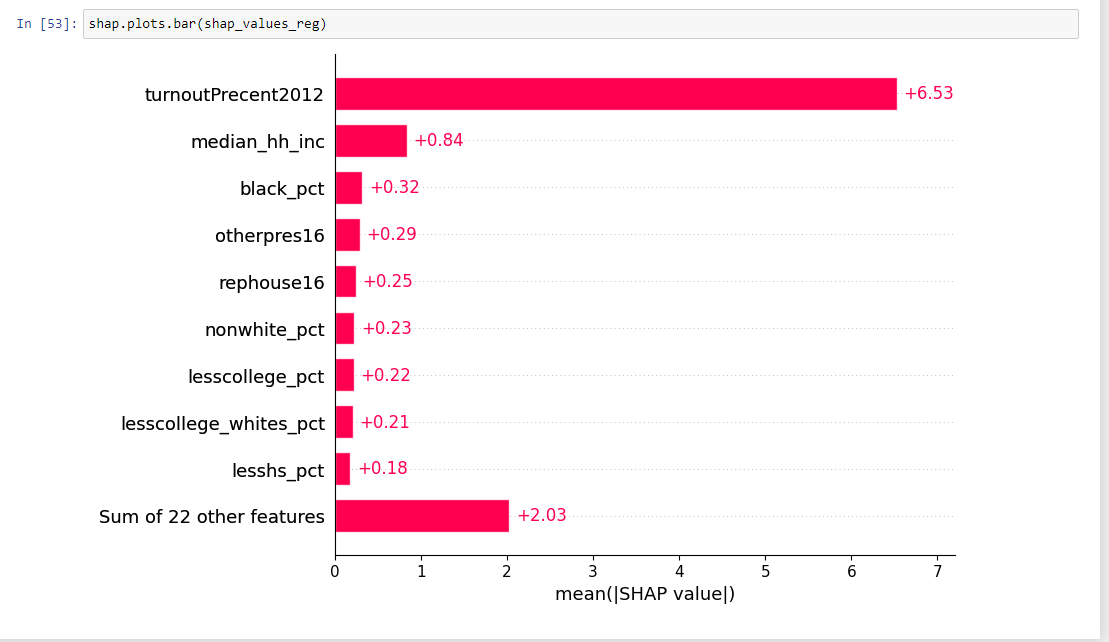


Our graph for our model using Random Forest Regressor

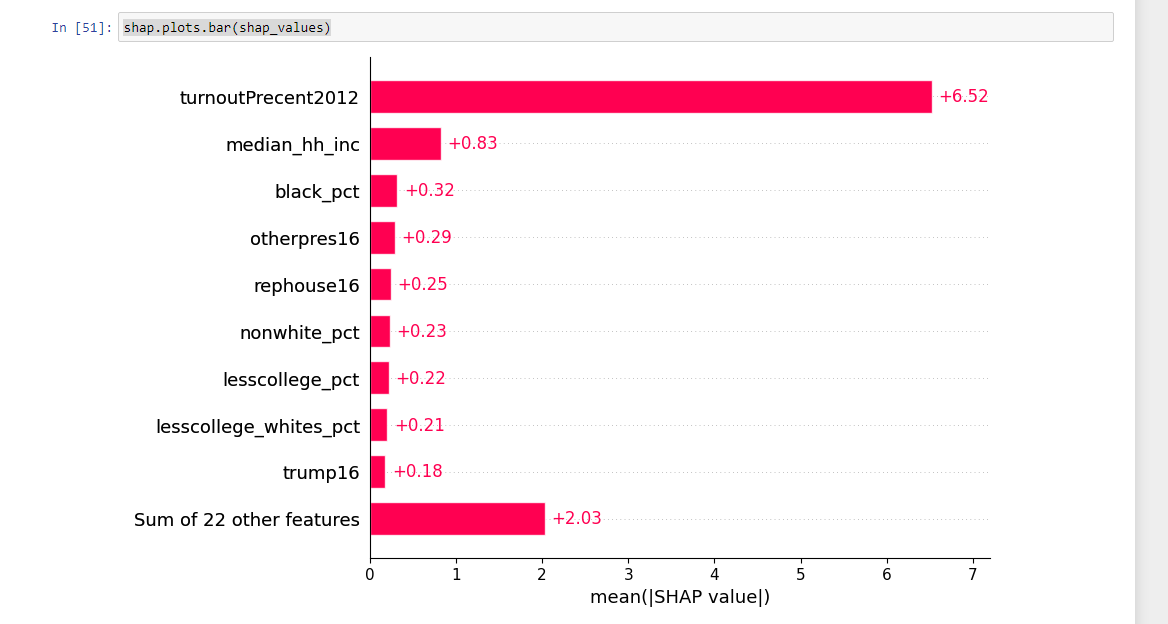




Finally our metrics using SHAP with our Regression Model



Metrics using SHAP for our Random Forest Regressor Model



Other Accomplishments

When creating my models, we ran into a problem in which I was not getting any outputs for state 7 and 38. Initially we thought it was an error in our models however we couldn’t understand why this was happening. Until after conducting further research we realized that Delaware and Rhode Island have very few counties. Delaware with only 3 and Rhode Island with only 5 meaning when creating out test split they were not included because they were left in training. This makes sense because comparing them to state such as Texas with 254 counties it would be impossible to include them in the test and validation sets.

# Changes/Problems

### Changes in Approach

There were many changes that we had to create when analyzing this dataset. Our initial plan for this dataset was to create a time series model on how the racial demographics will change overtime however after analyzing our dataset we realized this was not possible because we only had the demographics from 2018 so there was nothing to train our models to solve this problem.

Our next idea for this project was to find out of the total voters how many of these people are 29. This problem was also fundamentally flawed because the data provides a precent of the people of the TOTAL POPULATION under 29, meaning from age 0 – 29. This means that it also includes people who are illegible to vote. With the current data we have it would be impossible to do with without creating too many assumptions.

Finally, after conducting more research on our data, we settled on a problem that was more realistic with our given data. We wanted to see how analyzing past election turnover rates we can predict turnover rates in the future. This is possible because we have given data for our turnover rates for both the elections, the only assumption we must make is that the CVAP does not change since we are only given for 2018.

### Problems or Delays Experienced and Corrective Actions Taken

There were many problems when working with the data. Initially our biggest problem was trying to formulate a problem that is plausible to execute and solve with out given data. Secondly there were many issues when trying to manipulate the columns. When trying to iterate through columns and drop values that are N/A for our preprocessing it was difficult to work with since we did not encode our data at that point of the project. We found a solution by using .notna() which ended up doing what we wanted.

# Impact

### Impact on the Domain

After finalizing the results of our project, it is clear that there has high correlation in turnover rates from the past two elections 2016, and 2012. Using this data in the world of politics candidates can organized their campaign to heavily promote in areas that has high turnover rates because its clear that they already vote a lot. Or focus their attention into areas with less turnover rate to try and win over these voters who have not been voting at all to begin with.

### Impact on the Individual

This project initially seemed simple however after getting started and after many multiple failure and realizations it became evident that its not coding that goes into machine learning. Before even writing a single line of code there is a lot of thinking that must go into the planning. The question that we choose to answer must be answerable with the given data. This promote lateral thinking because its not as clear at first glace if something is answerable. It leads you to question your questions to find out if the questions that is being attempted to answer is plausible.

Aside from that we learned a lot about python and machine learning. Understanding why we do certain things and understanding the reasoning between minor things that may have been glanced over. For example, we were not exactly to sure on how regression worked which was the reason on why our first two attempts at the project ended up being failures. It is one thing to follow steps and do a project and something completely different to conduct a project on your own and understand why we did what we did.