Rachel Orey
Prof. Amir
DATS6103: Introduction to Data Mining
28 April 2020

Final Project Report

Informing Election Officials' Resource Allocation to Reduce Wait Times at Polling Places

Introduction

The first to conduct a major primary election amid the COVID-19 pandemic, in early April Wisconsin attracted global attention for line lengths exceeding 3 hours at some polling places.[i] Interestingly, the cause of Wisconsin's long lines was not itself due to the virus, but to a lack of poll workers and associated polling place closures. While polling place closures and poll worker absenteeism *was* due to the pandemic, long lines are not unique to a world in crisis.

In 2016, 4.8% of election precincts saw wait times over 30 minutes and 1.5% saw wait times of over one hour.[ii] While seemingly meager proportions, given that long wait times inhibit voter turnout, all precincts should aim to keep lines short when possible. As shown by the case in Wisconsin, many times long lines are a result of poor resource allocation. Too often, it is the underinformed distribution of resources (specifically polling places and poll workers) that causes undue wait times for voters.

In an effort to inform election officials' resource allocation, this project does two things at the county level: (1) predicts polling place line length based on current demographic and policy factors, and (2) provides the optimal number of polling places and poll workers necessary to keep maximum wait time below twenty minutes.

This report will discuss the model's datasets, experimental setup (including any data cleaning, preprocessing, and mining tools used), results, and high-level conclusions.

Datasets

The field of election administration has long been plagued by a dearth of reliable data. With data collection left to state and local governments, the analysis of nationwide trends often involves the amalgamation of many different—often incomplete—data sources; the same is true here.

The primary dataset that underlies this project is from Harvard's 2016 Survey of the Performance of American Elections (SPAE).[iii] This survey includes responses from 200

individuals in every U.S. state, yielding 10,200 total responses. While SPAE included an array of different questions, my analysis focused on survey responses in two areas: line length and the time voters arrive at the polls.

Unfortunately, data on voting equipment used by precincts is only available at the state level. I obtained equipment data by state from Ballotpedia.[iv] Voter turnout data is also only available at the state level. I obtained this data from the U.S. Elections Project, as published on the World Population Review website.[v]

In addition to election-specific data, my model relied heavily on demographic data obtained through the U.S. census. I obtained population density data from the University of Minnesota.[vi] This dataset is based on U.S. census data I was unable to locate the original source of. Furthermore, I obtained demographic data—namely, the proportion of non-white individuals in an area—directly from the U.S. census.[vii] Luckily, both of these sources provided data at the county level.

Experimental Setup

Data Cleaning and Preprocessing

The bulk of my preprocessing involved joining together data from the five above listed sources. Every dataset I used that included county-level data also included the county's Federal Information Processing Standards (FIPS) code. This helped introduce a layer of standardization and ensure that when merging datasets from different sources I was not inadvertently joining data that does not correlate.

While I entertained the idea of filling missing data with means or medians, I ultimately decided to drop missing data entirely. Because much of the data is categorical (voting equipment type, survey responses), I did not want to take the mean or median and in doing so treat categorical as numerical data. The original survey data obtained from the SPAE included 10,200 responses, yet only 7,394 (72.4%) remained after dropping empty or "I don't know" responses.

I matched the county code of the survey data with its respective demographic data, dropping all demographic data for counties that were not included in the survey. All counties included in the survey had demographic data present in the imported datasets. Of the 3143 counties included in demographic data obtained from the U.S. census, only 3 counties were not present in the population density dataset. These three counties were not apart of the survey data and thus would have been dropped anyways.

Finally, two of my datasets only had data available at the state level (voter turnout and voting equipment type). For each county present in the survey, I matched it with its state's voter turnout data (a percentage). I applied this same logic of extending state-level data down to each state's individual counties with the voting equipment data. However, voting equipment was categorical and not numerical so I ultimately created one-hot values for each voting equipment type.

PyQT Interface

Both the Decision Tree and Queue predictive model are housed in a PyQt user interface. This interface allows users to input their county's unique features and predict line length, as well as see how many poll workers and polling places they need to keep wait time below twenty minutes. See Figure 1 for an image of the user interface.

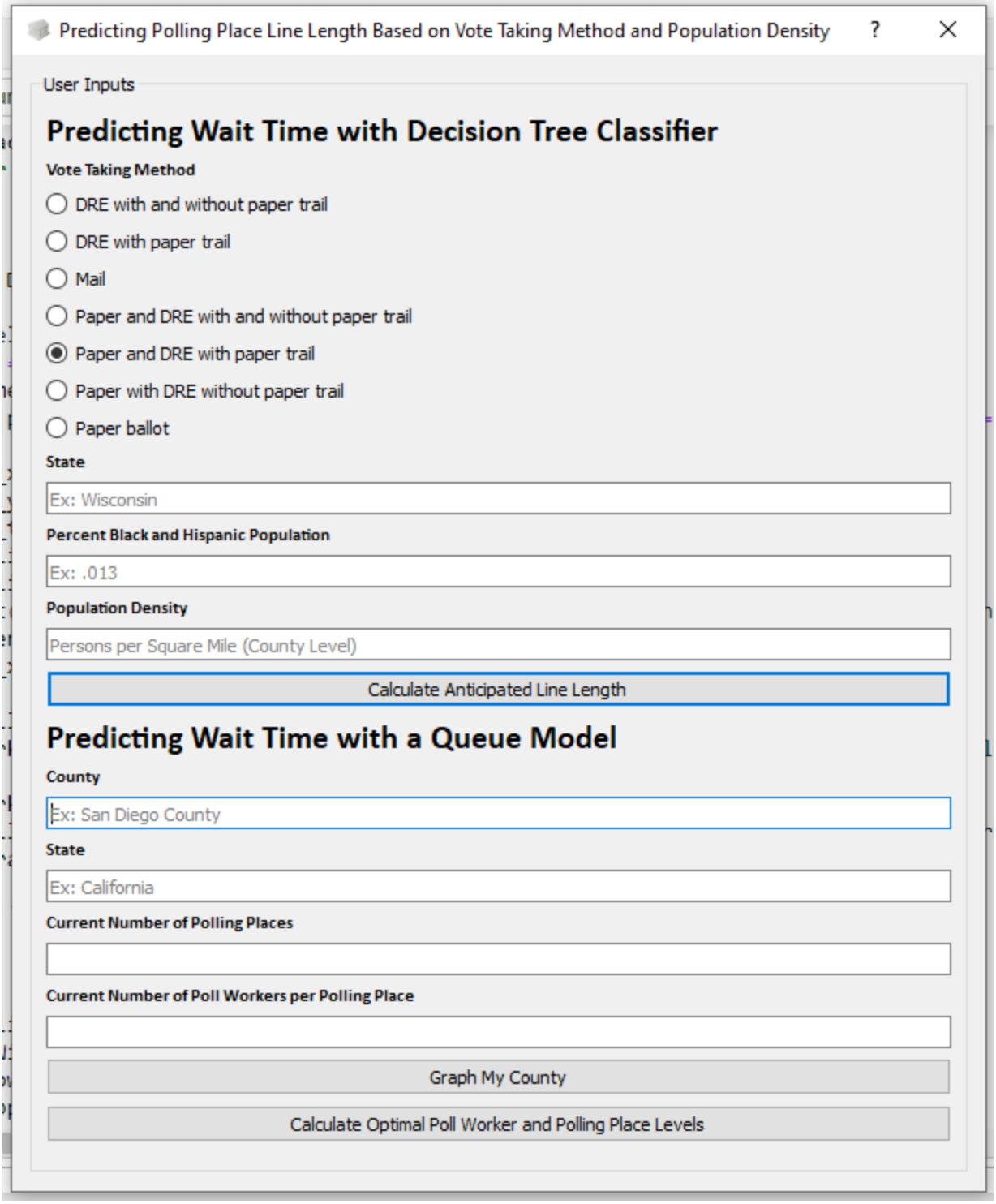
Predicting Polling Place Line Length Based on Vote Taking Method and Population Density
User Inputs
Predicting Wait Time with Decision Tree Classifier
Vote Taking Method
DRE with and without paper trail
DRE with paper trail
Mail
Paper and DRE with and without paper trail
Paper and DRE with paper trail
Paper with DRE without paper trail
Paper ballot
State
Ex: Wisconsin
Percent Black and Hispanic Population
Ex: .013
Population Density
Persons per Square Mile (County Level)
Calculate Anticipated Line Length
Predicting Wait Time with a Queue Model
County
Ex: San Diego County
State
Ex: California
Current Number of Polling Places
Current Number of Poll Workers per Polling Place
Graph My County
Calculate Optimal Poll Worker and Polling Place Levels

Figure 1

Predictive Modeling: Decision Trees

The first part of my project involved predicting line length based on a county's current demographic and policy factors (population density, percent non-white population, and voting equipment type). I predicted line length in two ways, first by using a decision tree.

Framing the question as a categorization problem led to the following features and targets. The original features that informed my model were population density, non-white population, and voting equipment type. An additional iteration of this decision tree approach involved including state as a feature. The target was line length: the categorical data provided from SPAE responses.

Decision trees operate by splitting data into categorical groups, each time hoping to increase the purity of the group. This project created two separate decision trees, one relying on the Gini index to determine the split and one relying on information entropy.

Gini and entropy are both metaethical algorithms which seek to determine a dataset's impurity. For instance, a group of data with 4 identical cases would have a total purity according to both Gini and Entropy; a group of data with 2 cases of one type and 2 cases of another type would have maximum impurity according to both Gini and entropy. However, a group of data with three cases of one type and one case of a separate type would have 75% impurity according to Gini and 81% impurity according to entropy. While Gini relies on more straightforward understandings of probability, entropy relies on logarithmic values – adding a level of sophistication to the analysis.

Using PyQt, I built a user interface in which users could select their vote taking method, their state, their county's Hispanic and Black population, and their county's population density. Upon clicking a button, these features were then fed into the predictive model and users were notified of their predicted line length according to both Gini and entropy (See Figures 2 and 3).

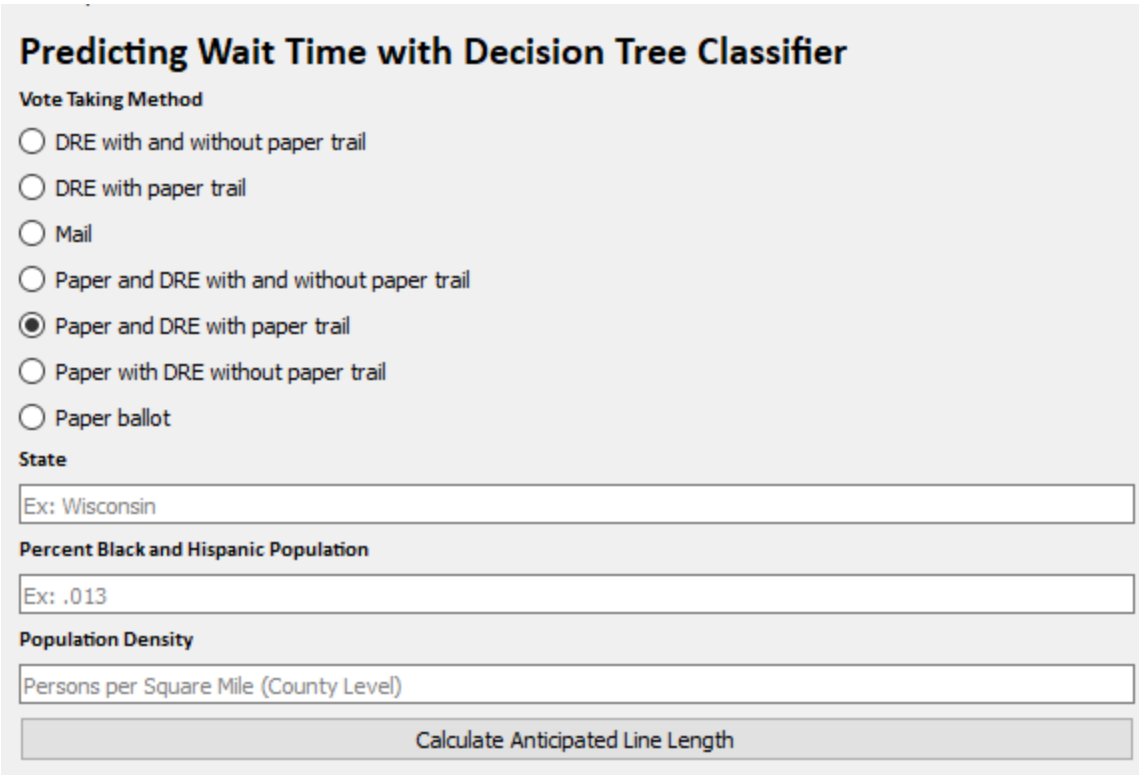

Figure 2

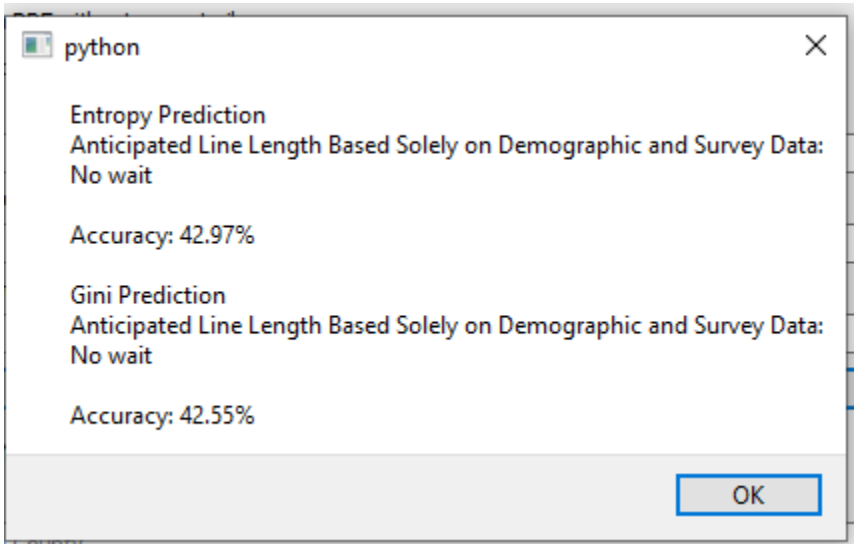

Figure 3

Predictive Modeling: Queueing Model

As will be discussed in the following results section, using decision trees to predict wait time performed poorly. As such, I built an additional queue model which predicts a county's maximum wait time based on county-level voter turnout, the number of polling places and poll workers, and the time voters typically arrive at the polls (this was taken from SPAE survey

responses and filtered by the county in question). This relied on a basic application of queue theory.

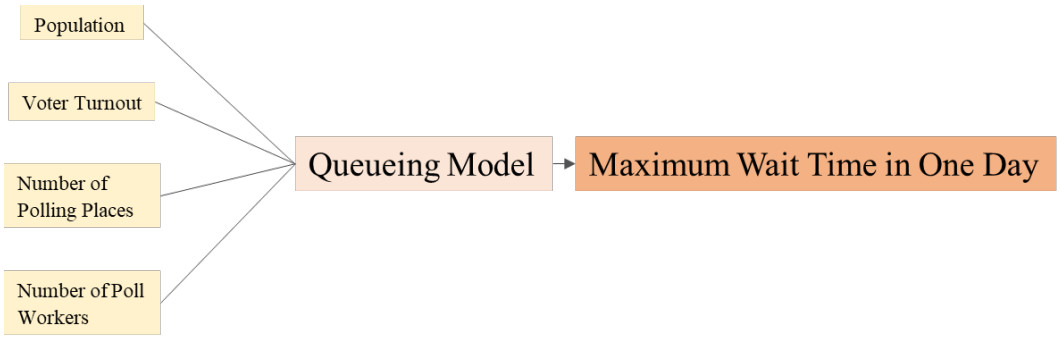

Figure 4

To order to calculate the number of voters arriving at polling places in a given day, I multiplied the county's total population by the state's voter turnout (as a percentage of the total population, not as a percentage of eligible voters) by the percentage of voters who vote in person (based on SPAE survey responses). I then multiplied an array with the percent of voters arriving at the polls each hour (also based on SPAE responses) by the estimated number of voters who will vote in person on Election Day to get an hourly break down of polling place arrivals (see Figure 5).

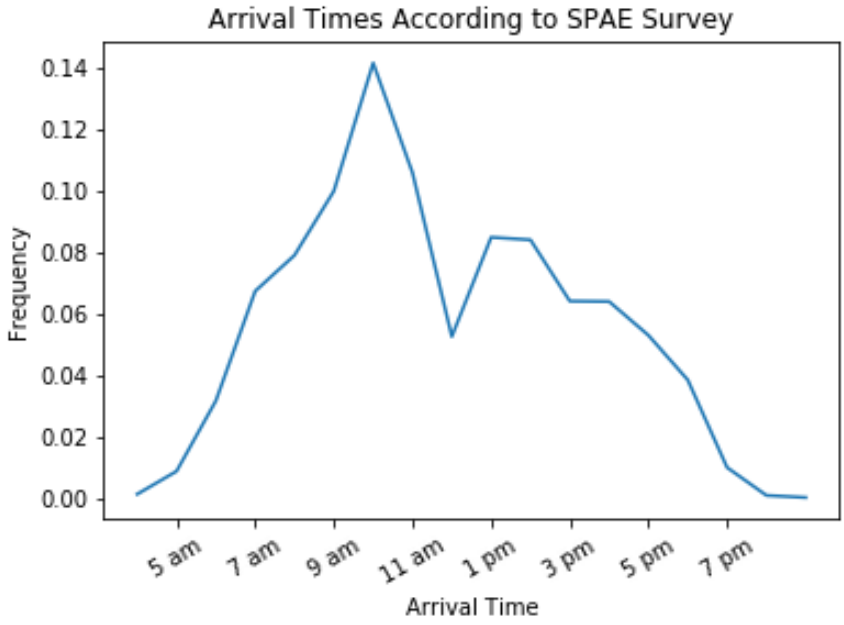

Figure 5

I then built a simple queue model which simulated wait time over the course of a single day. The maximum wait time was then returned as an output.

While I initially filtered arrival time survey responses by county, I ended up ruling out this filter as not all potential counties are included in the dataset. Additionally, filtering it shortened the sample size from over 7000 to at best 200 (each state had 200 survey respondents), increasing the risk of outliers affecting the data.

As with the decision tree approach, I used PyQt to build an interactive interface in which users can input their county, state, number of polling places, and number of poll workers to run a county-specific queue model. The results of the model are then displayed in a matplotlib graphic (Figure 6) and users can opt to see the minimum number of poll workers and polling places necessary to keep wait time below twenty minutes (Figure 7).

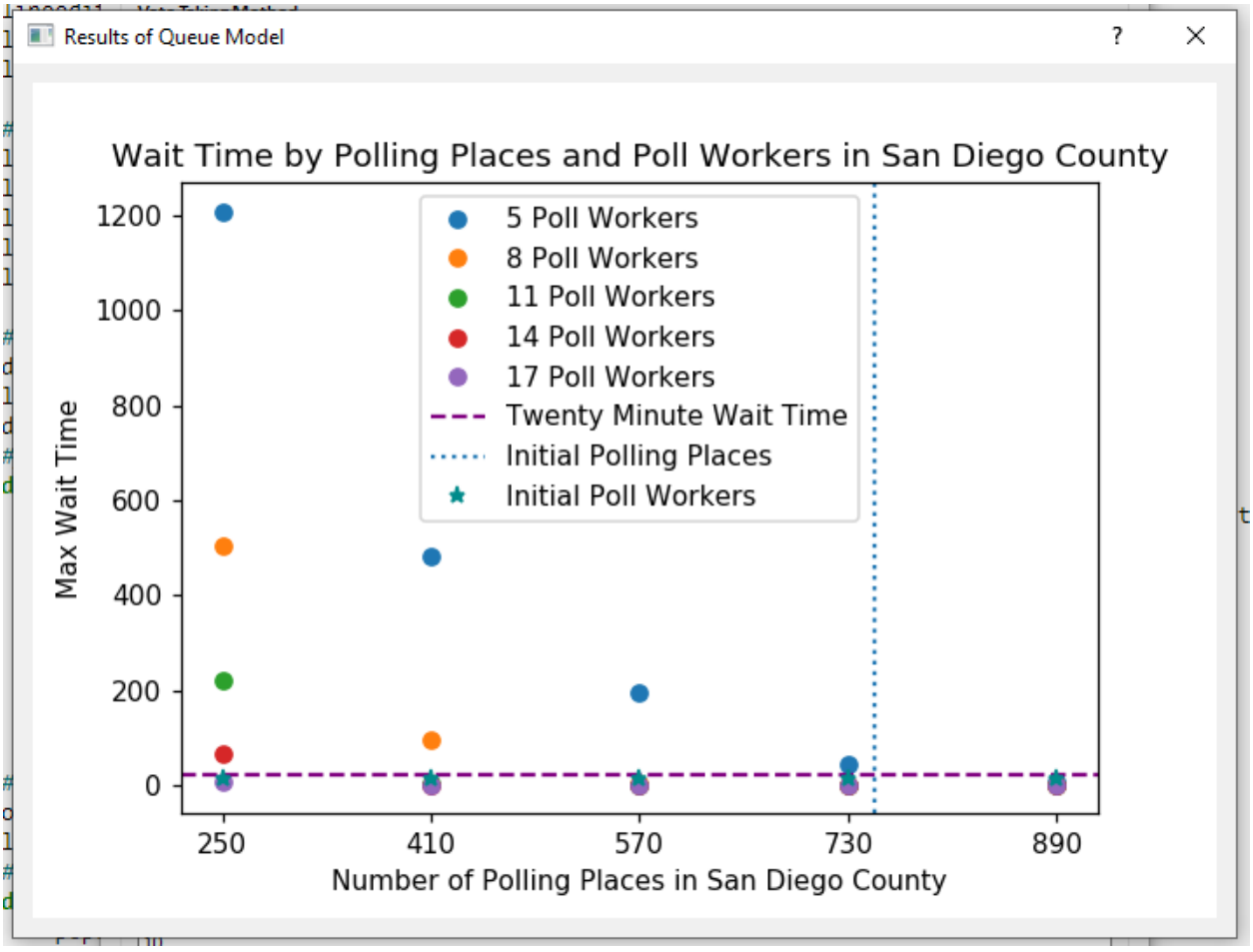

Figure 6

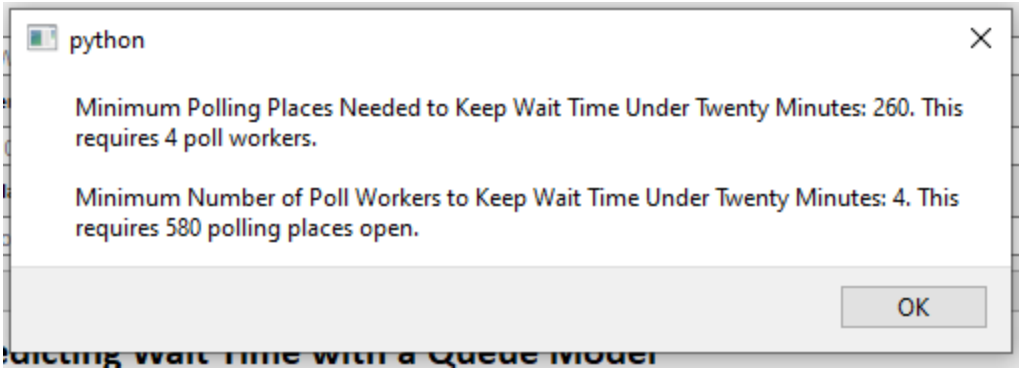

Figure 7

Results and Discussion

Predicting Wait Time with Sklearn Decision Tree Classifier

Each feature that I included in my analysis was based on former research which indicated that areas with high population densities and above-average proportions of Black and Hispanic populations typically face longer wait times at polling places.

However, basic regression analysis of both population densities and Hispanic and Black populations did not support these conclusions. Figure 8 shows a scatter plot of population density and wait time and Figure 9 shows a scatter plot of the proportion of county populations that is Hispanic and Black and wait time. Both graphs include a linear regression line which shows only a modest relationship between the features and line length.

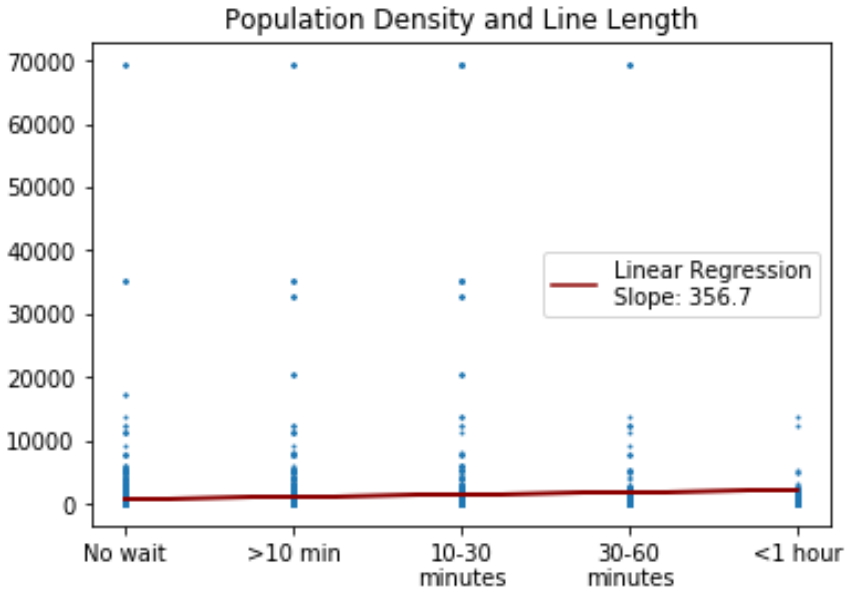

Figure 8

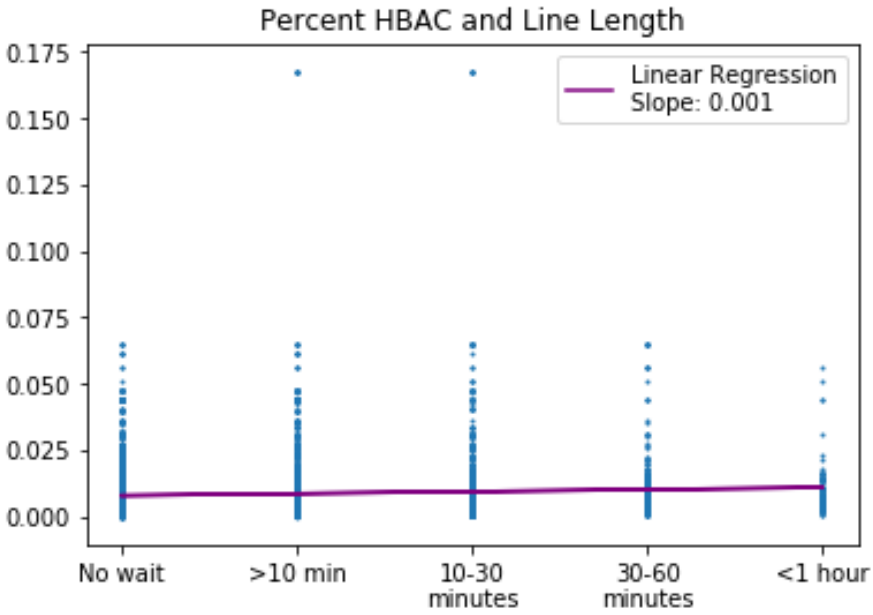

Figure 9

Given the poorly correlated relationship between line length and two of the prominent features—other features included voting equipment type (see Figure 11) and state—it is not surprising that the accuracy of the decision tree was low. For each iteration of the decision tree (Gini, entropy, with and without state as a feature), accuracy remained stable at roughly 40% (see Figure 10).

	Accuracy with Gini	**Accuracy with Entropy**
State Included as Feature	42.55%	42.97%
State Not Included as Feature	41.23%	42.13%

Figure 10

Taking a step back to see the model in aerial view helps to diagnose why accuracy is so low. The issue lies primarily in imbalances in the scale of feature and target data. The feature data (population density, minority populations, voting equipment type) was aggregated at the county, if not state, level. However, the target data was comprised of *individual* survey responses. It is entirely sensible that county and state level features would be a poor predictor for individual experiences. Instead, a better predictive model would need target data to include maximum wait time at the county or state—rather than individual—scale.

Predicting Wait Time with a Queuing Model

A queuing model is far more straightforward than a decision tree classifier. Additionally, the queue model is more hypothetical in nature; while accuracy can be clearly tested with the decision tree approach, there is no clear way to test the results of a hypothetical model. Instead, because the queue model relies on widely accepted information about wait time's relationship with the number of voters, service time, etc., the results of the queue model can be trusted as accurate insofar as their inputs reflect reality. The true measure of the queueing model's accuracy would require a live simulation of the inputs that are given to the model. Notwithstanding that live simulation, we can trust the model's results to be an accurate account of hypothetical inputs.

While the inputs are hypothetical, election officials could use the model to determine how many polling places and poll workers to utilize in order to keep wait time short.

The major area for improvement in the queue model lies in service time. Currently, the model assumes a mean service time of two minutes per voter (this is informed by personal work experience in the field of elections). However, as shown in Figure 11, line length differs greatly according to the voting equipment utilized by precincts. A more advanced queuing model could associate service time with voting equipment in use to more accurately tailor results to specific jurisdictions.

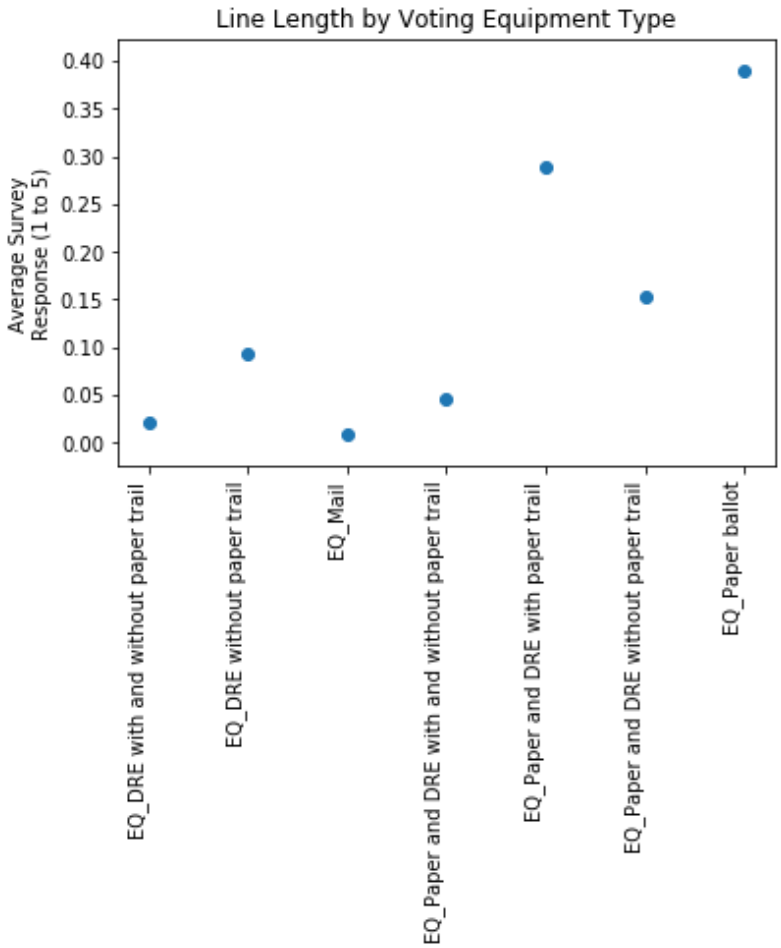

Figure 11

Summary and Conclusions

Given the imbalance in scale between feature (county) and target (individual) data, a 40% accuracy rate is not as bad it may seem. While the decision tree part of this project may be a poor predictor of county-level maximum wait time, it shows the promise that such an approach could provide with the proper inputs.

Furthermore, the queuing model aspect of this project is a useful tool that could help election officials across the county determine how many poll workers and polling places they need to keep wait times low. One aspect of the queuing model that could be improved is service time. Service time is an integral aspect of any queue, and the next iteration of this model should include a more tailored estimate of service time that is based on the type of voting and registration equipment in use.

In short, this project highlights the challenges caused by a lack of commensurable elections data on a national scale. Leaving data collection and reporting to state and local jurisdictions makes any sort of aggregate modeling or analysis difficult.

References

[i] Maayan Silver. "Long Lines Reported As Wisconsin Election Proceeds Despite Coronavirus Threat." *NPR.org*. 4 April 2020. Available at: https://www.npr.org/2020/04/07/829091968/long-lines-reported-as-wisconsin-election-proceeds-despite-coronavirus-threat.

[ii] Matthew Weil. "The 2018 Voting Experience: Polling Place Lines." 4 November 2019. Available at: https://bipartisanpolicy.org/report/the-2018-voting-experience/.

[iii] Stewart, Charles, 2017, "2016 Survey of the Performance of American Elections", https://doi.org/10.7910/DVN/Y38VIQ, Harvard Dataverse, V1, UNF:6:/Mol52fZ59fx6OsPWIRsWw== [fileUNF]

[iv] "Voting methods and equipment by state." *Ballotpedia.org.* Available at: https://ballotpedia.org/Voting_methods_and_equipment_by_state .

[v] "Voter Turnout by State." *World Population Review.* Available at: https://worldpopulationreview.com/states/voter-turnout-by-state/

[vi] Schroeder, Jonathan P. (2016). Historical Population Estimates for 2010 U.S. States, Counties and Metro/Micro Areas, 1790-2010. Retrieved from the Data Repository for the University of Minnesota, http://doi.org/10.13020/D6XW2H. Available at: https://conservancy.umn.edu/handle/11299/181605 .

[vii] "Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2018 (CC-EST2018-ALLDATA)." *County Population by Characteristics: 2010-2018*. U.S. Census Bureau. Available at: https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html#par_textimage_1383669527