Predicting elections has always been extremely difficult due to the large number of factors at play, including demographic, political, technological, economic, and historical. Due to the recent proliferation of machine learning, a technique used to create predictive models from large sets of data, it is possible to isolate a number of these factors and create a model to predict elections based upon past data, specifically economic factors. Some of these factors are median income, poverty rates, unemployment rates, education, and GDP (Hummel & Rothschild, 2014). These factors are connected to elections because of the links between how poorer neighborhoods vote as opposed to rich ones, especially when in a specifically prosperous or unprosperous time, or how economic performance can affect people’s perception of the president or party responsible. Oftentimes, location and geography can have an impact on economic factors and election results as well (Braha, 2017).

Machine learning techniques rely on the idea of using large amounts of data to slowly optimize a model through multiple iterations. This process is called “training”. Machine learning models can take large amounts of data that are difficult for humans to interpret and discern patterns. There are also multiple types of models, e.g. linear regressions, logistic regressions, decision trees and random forests, and support vector regressions. Additionally, there are multiple types of neural networks that are used to find nonlinear relationships. Because each model is extremely different in the internal mechanics, different models can work for certain types of data or data sets, but fail miserably for others. Because of this, and due to the complex nature of elections and economic data, the proper model to be used is still not certain, and multiple models have been attempted (Zolgader, Niaki, & Niaki, 2017).

Beyond this, the models themselves can be easily edited and tuned for hyperparameters that affect how the model learns and trains. For example, neural networks allow for multiple levels of data abstraction by creating neurons that represent a function that are stacked in layers, allowing it to slowly build levels of complexity on a model. However, because of the abstract nature of neural networks, the actual architecture of the network, the number of neurons per layer, how the layers are connected, hyperparameters such as the learning rate, the activation functions used, the loss functions used, and how many layers are present can be difficult to optimize and can have profound effects on the performance of the model. Furthermore, because of this high degree of customizability, there is a significant danger of overfitting the model, resulting in a model that achieves high accuracy within the dataset, but without any predictive power, due to memorizing the data rather than actually learning how to model it. Because of the difficulty in both optimizing and picking a model to be used, many times multiple models are compared to determine the best one for the specified problem.

Additionally, most of the more traditional models also have hyperparameters that can be changed to optimize the model. For example, support vector machines can use multiple different kernel functions that essentially map the data to a higher dimensional space in different ways in order to find an optimal hyperplane that separates it with minimal error, allowing for the exploitation of unintuitive but effective nonlinear relationships in the data. Some of these kernel functions include linear, hyperbolic tangent, and radial basis. These different functions also have parameters that can optimized, such as cost, or how much the model is punished for incorrectly classifying a point (or for having an inaccurate regression), as well as function specific parameters. Therefore, multiple models can be created, tested, and altered to suit the data structure and optimize results (Zolgader, Niaki, & Niaki, 2017).

Current and past research has either used traditional linear regression techniques to create models, used non-economic variables, or focused on national or possibly statewide elections. Many of them also ignore the demographics of the candidate and focus on the party alone or ignore the previous performance of the local economy and its composition. For example, Zolgader conducted a study in which they used three models, a linear regression, a support vector regression (SVR) and an artificial neural network (ANN), each trained with national presidential approval rating data from 1952-2012, with 2004, 2008, and 2012 being set aside for training. While they found that SVR was the most accurate model, with the ANN just slightly behind, they both improved on linear models by 50%. They recommended using more localized statistics (state-wide) or possibly “combining” SVR and ANN to create a superior model for future research (Zolgader et al., 2017). This shows that both SVR and ANN, as machine learning models, were applicable to the problem as opposed to the more traditional statistical learning or linear regression based models. Another study by Hummel and Rothschild used a linear regression and state-wide data on approval ratings, incumbency, GDP, past results, state ideology, and the change in income over time, as well as biographical information about the candidate, such as home state advantage, or how being from a certain state may improve your chances of winning that state. Additionally, they added multiple controls and variables to influence how certain variables are treated by the model in importance, such as the aforementioned home state advantage. They also included multiple variables on the incumbency, such as the ideology of the last person in office. Overall, their model correctly predicted the state-wide outcome 90.0% of the time within the sample, and 85.8% correct outside (Hummel & Rothschild, 2014). These different categories of data and examples of feature engineering show how the data can be manipulated to extract more information from it. This implied that while more complex models can be more accurate, what may be more important is the accuracy of the data used and what impacts are accounted for, especially as linear regressions are much simpler than support vector machines or neural networks.

Additionally, Abrams found that state economic factors can have an effect on voting results, and another study found that certain regions of the US are more prone to certain economic and political leanings (Abrams, 1980). For example, the study says that the Midwest generally attains a lower education, is more white, conservative, and less economically innovative, while the middle Atlantic is the opposite, which points to a relationship between political beliefs, socioeconomic status, and location (Braha, 2017). Additionally, a study of the 1992 presidential election showed that negative real income growth and an increase in unemployment from the last presidency in each state were extremely significant in the defeat of Bill Clinton, showing that income and unemployment have a significant effect on elections. They also found that the introduction of a relatively popular third party candidate did not have a significant effect on the overall outcome(Abrams, 1993). Furthermore, research into predicting elections for a parliamentary system has been completed, using constituency level data. Munzert shows how the combining of models into one, also known as creating an ensemble model, can be useful if the models are significantly different in the data they take and how they predict. This is because the combination of models allows for the exploitation of more data while also being able to reduce the biases inherent to each model by balancing it with others. Through combining a model based upon historical data and a model based upon current polling data, Munzert was able to predict 93.3% of the seats correctly, with 75% of those incorrectly predicted being within 4 percentage points of winning (Munzert, 2017).

In this research, I will be using county wide demographic and historical data with multiple machine learning models, from support vector machines to artificial neural networks, in order to predict presidential election results. This will be different because not only will multiple models be tested and optimized, or possibly combined into an ensemble, but countywide data for both election results and demographics will be used to create a more accurate model. Previous research has shown that statewide data not only affects voting behavior but also can be used to create a better model than simply national data, so county data will hopefully be able to narrow down communities of interest and create a more balanced picture of how a group of people will vote based upon their demographics and socioeconomic conditions. Additionally, unlike many of the previous studies, these models will include recent election history and associated economic trends. Also, because there are far more counties than states in the US, there will be more data available to train a neural network with, which may make that model more accurate considering that neural networks require much data to be trained effectively. Other models in the past have either not used machine learning to predict elections, relied heavily on polling data, or have not accounted for demographic variables.

ED:

The final model will be a custom-built ensemble model from two separate models: A standard prediction model based upon localized polling data leading up the election, and a swing model based upon the socioeconomic and demographic factors compared to the previous cycle. Ideally, the degree of localization will be to the county level, but perhaps state level data will be the best option for being able to gather data. This is because, as Munzert showed, the ensembling of two models that use fundamentally different data and create predictions in different ways can drastically reduce error and increase prediction accuracy. They also used a prediction and swing model, and they used constituency level polling data in order to predict their results. However, data scarcity is a large issue when dealing with constituency level data, which is why state level data may be used at first (Munzert, 2017). Additionally, as Zolghadr showed, complex techniques using simple data can give accurate results, at least on the national level. In their conclusion they advocated for attempting to use state level data to predict the elections in future research, and they also used polling data on presidential approval ratings in order to predict the elections (Zolghadr, 2017). Furthermore, Hummel and Rothschild were able to show that even with simple linear model techniques, you can accurately predict multiple types of elections by using a combination of demographic, economic, and dummy variables that allow for the fine tuning of data through feature engineering (Hummel & Rothschild, 2014). As Abrams showed, economic factors significantly affect elections, or at least in the case of 1992, where there was a large economic recession. They showed how Bush’s results in the electoral college were directly related to the unemployment changes and changes in income (Abrams, 1993). This allows the socioeconomic and demographic data to play a large part in the swing model, which will determine whether or not a certain region will switch parties.

The data for polling data will be obtained from Zolghadr if possible, and, if not, from Gallup polling. Additionally, there are other polling sources that may be introduced if it is found that the Gallup data is inherently biased towards one type of victory or another, such as the Roper center. The data will be manipulated and tested through Weka, a program built for machine learning applications that allows for data visualization, engineering, manipulation, and testing. The data will be imputed for missing values, scaled, clustered, and will go through a stepwise regression to determine what components will have an effect on election results. The data will also be recorded at each step in the process to determine which preprocessing techniques will manipulate the data so that it will predict most optimally. Additionally, the preprocessing techniques themselves will be able to tuned, such as what bounds we use for scaling, or how we impute missing values. Once the data has been handled, it will go through multiple model and statistical types to determine which models operate the best, including PCA, random forests, logistic regressions, gradient boosting algorithms, SVMs, and ANNs based in Tensorflow. Each model will then be tuned using cross validation and grid searches for each dataset to determine what hyperparameters are most optimal for each dataset, and finally the combination of best dataset, model, and hyperparameters will be used for the final model. For the swing model, this process will be repeated using different data, and then the two will be ensembled and adjusted for weights so that they will improve overall accuracy of the model together. Most of the data visualization and model creation will be done in python Jupyter notebooks using a combination of Pandas, Sklearn, Seaborn, and Tensorflow with Keras.

References:

Zolghadr, M., Niaki, S. A. A., & Niaki, S. T. A. (2017). Modeling and forecasting US presidential election using learning algorithms. *Journal of Industrial Engineering International, 2017*, 1-10. doi:10.1007/s40092-017-0238-2

Hummel, P., & Rothschild, D. (2014). Fundamental models for forecasting elections at the state sevel. *Electoral Studies, 35*, doi:10.1016/j.electstud.2014.05.002

Braha, D., & de Aguiar, M. A. M. (2017). Voting contagion: Modeling and analysis of a century of U.S. presidential elections. *PLOS ONE, 12(5),* e0177970. doi:10.1371/journal.pone.0177970

Abrams, B. A. (1980). The influence of state-level economic conditions on presidential elections. *Public Choice, 35(623)*. doi:10.1007/BF00140090

Munzert, S. (2017). Forecasting elections at the constituency level: a correction-combination procedure. *International Journal of Forecasting, 33(2), pp.467-481*, doi:10.1016/j.ijforecast.2016.12.001

Abrams, B.A., Butkiewicz, J. (1993). The influence of state-level economic conditions on the 1992 U.S. Presidential election. *Public Choice, 85(1-2), pp.1-10.* doi:10.1007BF01047898