Background Research Paper                                                                      Ronan Tegerdine

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 advent and proliferation of machine learning, a technique used to create 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, or GDP (Zolgader, S. A. Niaki, & S. T. Niaka, 2017).  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 from the data. There are also multiple types of models, from linear regressors and classifiers to other types of convolutional neural networks or support vector regressions. Each have their own methods of creating a model and are thus differently suited to different tasks, but 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. For example, neural networks work by creating neural networks that allow for multiple levels of data abstraction by having a “filter” for each level of “neurons” (which are essentially computations) allowing it to slowly build levels of complexity on a model.  These models can have extra layers or neurons added that affect the performance of the model based upon what is required from it. Therefore, multiple models can be created, tested, and altered to suit the data structure and optimize results. While these models can find nonlinear correlations and are very robust in their options to find patterns and correlations, they often require a large amount of data to train, which only increases as neurons and layers are added. They can also be difficult to optimize, specifically in finding the right parameters to use for each model. A relatively new platform called TensorFlow can be used to create these types of models, which has multiple systems and methods to create and edit models easily. TensorFlow has already been used by Google and in multiple biological research projects and has native support for neural networks and other types of models that may be used, with a way to edit and change the models as to allow easy modification and testing (Rampasek, Ladislav & Goldenberg, 2016).  
 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, S. A. Niaki, and S. T. Niaki 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 data (including GDP, approval rating, how many terms the candidate has been in office, personal income, recent votes and the unemployment rate). While they found that SVR was the most accurate model, improving on existing 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). Another study by Hummel and Rothschild used a linear regression and state-wide data on approval ratings, GDP, unemployment, and income over time, as well as biographical information about the candidate, such as home state advantage or how an existing party aligned with them in the state affects the results. Additionally, they added multiple controls and “dummy” variables to influence how certain variables are treated by the model in importance, such as the aforementioned “home state advantage”. Overall, their model correctly predicted the state-wide outcome 90.2% of the time within the sample, and 89.1% correct outside (Hummel & Rothschild, 2014). 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. Additionally, a study by 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 is less educated, 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).

This research will be different because not only will multiple models be tested and optimized, 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 socioeconomic conditions. Additionally, unlike many of the studies previously, these models will include the demographics of the candidate and other variables that could potentially affect the outcome, such as 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.

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

Rampasek, L., & Goldenberg, A. (2016). TensorFlow: Biology’s gateway to deep learning? *Cell Systems, 2,* 12-14. doi:10.1016/j.cels.2016.01.009