Election Prediction Using Random Forests in R

Introduction

This project aims to demonstrate the power of machine learning, specifically using Random Forests, in predicting the U.S. election winners by county. The model uses a dataset from the 2016 U.S. elections and applies Random Forests to predict the winners.

Dataset Overview

Source: The data has been compiled from Ben Hamner's 2016 election dataset available on Kaggle.

Response Variable

mostVotes: Candidate who obtained the majority of votes in a given county (either "Clinton" or "Trump").

Predictors

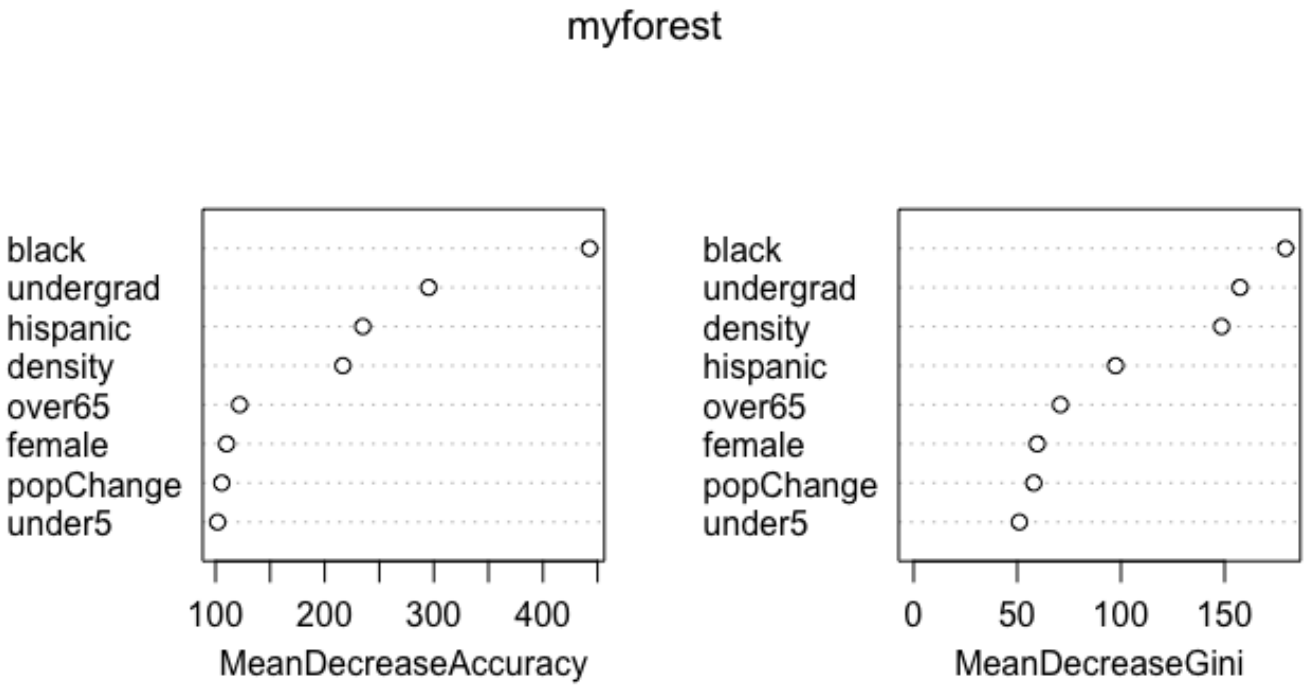
- state: 2-letter state abbreviation.
- county: Name of the county.
- fips: Federal Information Processing System (FIPS) code for the county.
- popChange: Population change in percentage terms between 2010 and 2015.
- under5: Percentage of the county population under 5 years of age.
- over65: Percentage of the county population over 65 years of age.
- female: Percentage of the county population that is female.
- black: Percentage of the county population that is Black.
- hispanic: Percentage of the county population that is Hispanic.
- density: Population density (inhabitants per square mile).
- undergrad: Percentage of the population (aged over 25) holding a bachelor's degree.

Methodology

I used the randomForest package in R to build my predictive models. I experimented with multiple versions of Random Forests:

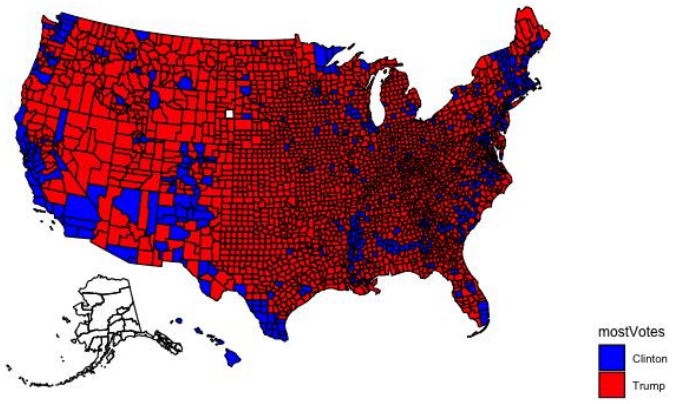
- **10,000 Trees**: This forest achieved an Out-of-Bag (OOB) error rate of around 6-7%, indicating a highly accurate model.
- **50 Trees**: This version also produced similar results to the 10,000-tree model with around 7-8% OOB error rate.
- **5 Trees**: This version shows some misses but still performed reasonably well at around 11% OOB error rate.

The following shows the relative importance of each variable in predicting which candidate won with the random forest with 10,000 trees:



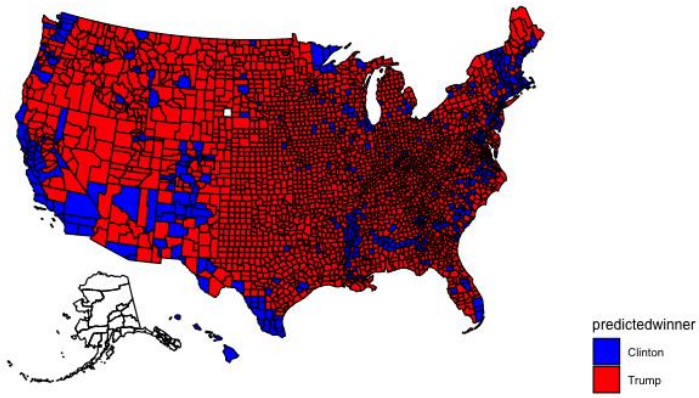
Visualizations

Actual Winner by County



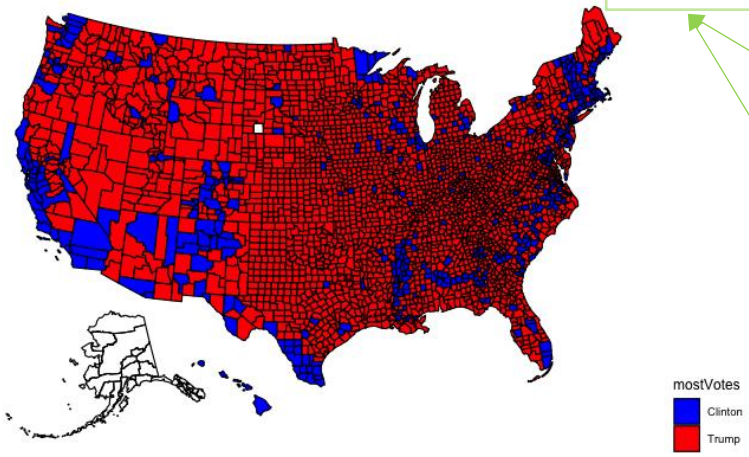
Predicted Winner by County

Predictions from a 10,000 tree random forest.



Predicted Winner by County

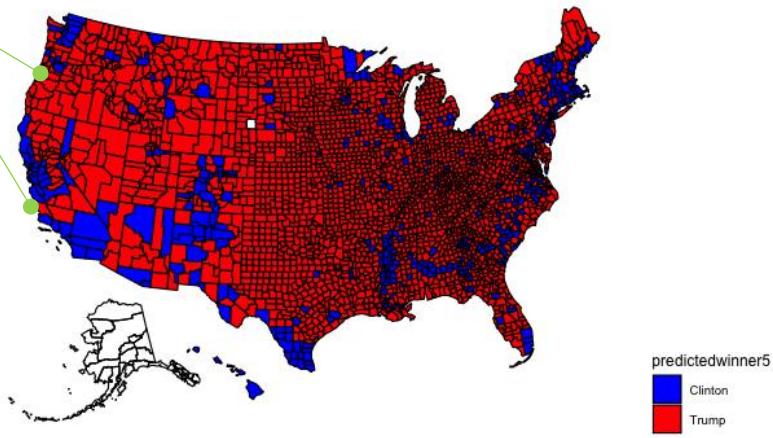
Predictions from a 50 tree random forest.



A few predictions went wrong here!

Predicted Winner by County

Predictions from a 5 tree random forest.



Conclusion

My models demonstrate the efficacy of using Random Forests for predicting election outcomes. Even a forest with only 50 trees can generate results nearly as accurate as a 10,000-tree forest, showcasing the robustness of this machine learning approach.