PENNSYLVANIA ELECTION ANALYSIS

ELECTION 2024

PRESENTED BY:

Makonnen Ramsey

Kristin Bell

Adrian Bayemi

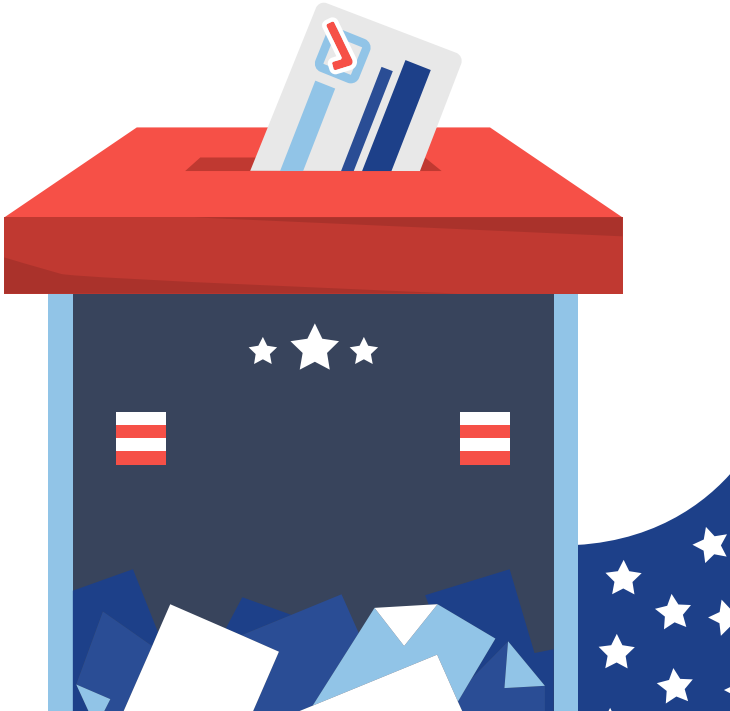


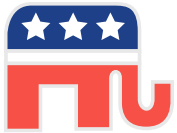
TABLE OF CONTENTS

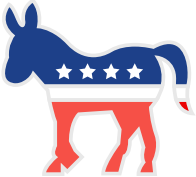
INTRODUCTION

It is abundantly clear that this country is divided in their political views, with most polls showing a near 50/50 split between Democrats and Republicans. Even so, there are plenty of undecided voters ‘up for grabs.’ With the upcoming presidential election in November, each party is eager to convince these undecided voters to cast their vote for their respective candidate.

You only have to turn the television or radio on for a short time to prove this point.

WHAT SETS US APART?





Although components that contribute to an individual's political preference are numerous and extremely complex, it's clear that factors affecting individuals' day-to-day lives, such as economic conditions, educational background, and social interactions can play a significant role in shaping their political affiliations. We were interested to see what, if any, of these traits of an individual have an impact on the party they support.

By taking a close look at socioeconomic conditions and demographic elements, could it be possible to train a deep learning model capable of determining a specific geography/population's election outcome?

PENNSYLVANIA: KEY BATTLEGROUND STATE

Pennsylvania has historically been a ‘swing state’ in presidential elections, and 2024 is no different. Not only has the candidate that won Pennsylvania been the same candidate that won the election as a whole in 47 out of the 59 elections held, it holds a large number of electoral votes (19) making it a pivotal state in the path to victory.

For these reasons, we decided to use Pennsylvania as a ‘test case’ to see if a voter profile can be determined using metrics obtained from the U.S. Census Bureau. There is a plethora of data to be sourced, but we found the most consistent data grouped by county. There are 67 counties in Pennsylvania, but the U.S. Census Bureau only had complete statistics (for the demographics we wanted to measure) for 40 of the 67 most populous counties.

METRICS COLLECTED

"DEMOGRAPHIC AND HOUSING ESTIMATES" (DP05)

- Total male and female percentage
- Male and female percentage over 18 years old (voting age)
- Race percentage: White, Black or African American, American Indian and Alaska Native, Asian

"EDUCATIONAL ATTAINMENT" (S1501)

- 18 to 24 years old
  - Percent High school graduate (including equivalency)
  - Percent Bachelor's degree or higher
- 25 to 34 years old
  - Percent high school graduate or higher
  - Percent Bachelor's degree or higher
- 35 to 44 years old
  - Percent high school graduate or higher
  - Percent Bachelor's degree or higher
- 45 to 64 years old
  - Percent high school graduate or higher
  - Percent Bachelor's degree or higher
- 65 & Older years old
  - Percent high school graduate or higher
  - Percent Bachelor's degree or higher

"INCOME IN THE PAST 12 MONTHS (SURVEY YEAR INFLATION-ADJUSTED DOLLARS" (S1901)

- Households, Families, Married-Couple Families, Nonfamily Households
  - Total count for each group
  - Median income (dollars) for each group
  - Mean income (dollars) for each group
- Percentage of each group that falls into the income brackets:
  - Less Than $10,000, $10,000-$14,999
  - $15,000-$24,999, $25,000-$34,999
  - $35,000-$49,999, $50,000-$74,999
  - $75,000-$99,999, $100,000-$149,999
  - $150,000-$199,999 & $200,000 or More

"VETERAN STATUS" (S2101)

- Total Population, Veteran Population and Non-veteran Population
  - For each of the above groups:
    - → Percentage in age brackets "18 to 34", "35 to 54", "55 to 64", "65 to 74" and "75 and Over"
    - → Labor force
    - → Unemployment
    - → Percentage with income below poverty level in the past 12 months
    - → Percentage with any disability

"FERTILITY" (S1301)

- Women with births in the past 12 months
- Birth rate per 1,000 Women in the following age brackets:
  - "15-19 Years Old", " 20-34 Years Old", 35-50 Years Old"
- Received public assistance income rate per 1,000 Women

CLEANING THE DATA

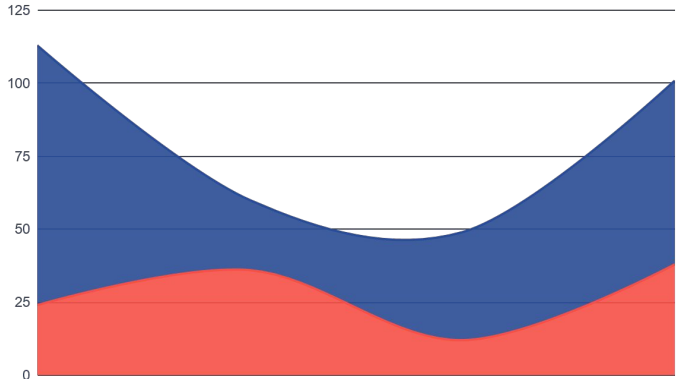
PRESIDENTIAL RESULTS BY COUNTY (2000-2020)

We dropped unnecessary columns, parties other than Democrat or Republican, added a ‘winning party’ column which was amended to a binary classifier and filtered to PA counties only for years 2012, 2016 and 2020

U.S. CENSUS TABLES BY COUNTY (2012, 2016, 2020 AND 2023)

Attributes for the chosen metrics were loaded into dataframes and cleaned. We utilized imputation and dropping of ‘NaN’/null values to ensure the machine learning techniques could be implemented on the data

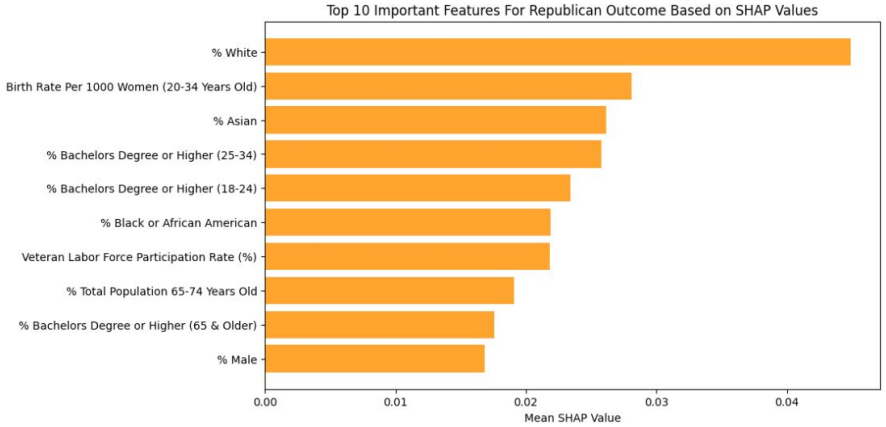
TRAINING THE MODEL

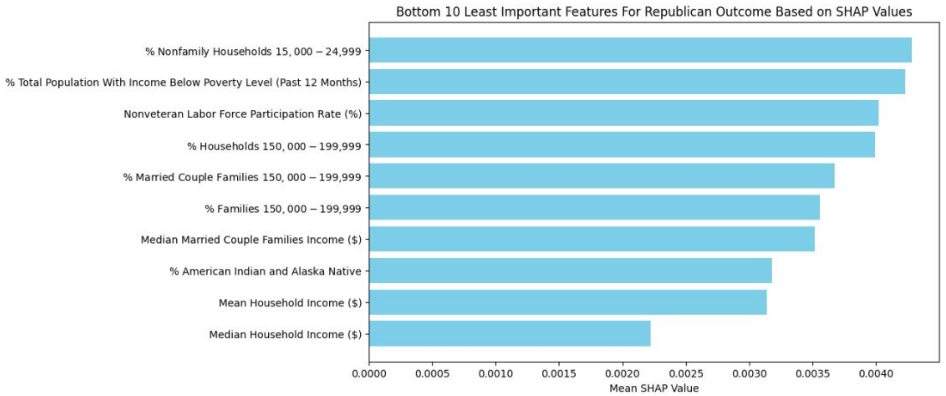


The cleaned data was then separated into a labeled/target outcome and an attributes array with 102 input dimensions. The dataset was subsequently split into training and testing sets, scaled and then utilized to compile, train and evaluate an initial neural network model. This yielded a **loss metric of 0.1034** and **accuracy metric of 0.9666**.

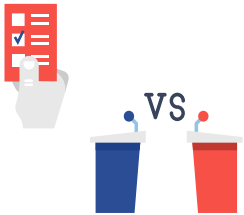
Model optimization was conducted via a Keras-Tuner approach to determine the best activation function, number of layers, number of neurons/nodes per layer, number of epochs, etc. The optimized model returned an **improved loss metric** and **accuracy metric of 0.0386 and 1.0 respectively.**

FEATURE IMPORTANCE





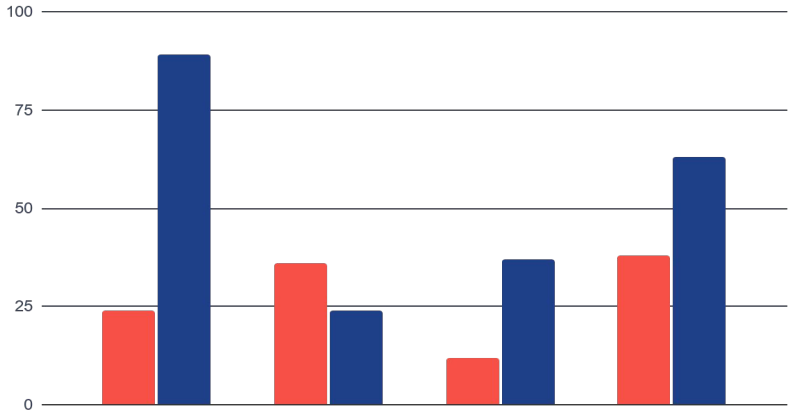
Feature importance for our model was also determined by the SHAP (SHapley Additive exPlanations) method, which was graphically represented as the top ten features with highest SHAP values meaning they are the most influential in driving the model's predictions toward a (1) or Republican outcome and the bottom 10 features with the lowest SHAP values describing the least contributing features to the model's predictions toward a (1) or Republican outcome.



MAKING PREDICTIONS

The optimized model and its associated scaler were saved/exported in order to generate predictions on the most recent available county attribute data from 2023. The algorithm's predictions for the winning political party for each of the 40 Pennsylvania counties was added back into the dataset, which was then combined with the 2012/2016/2019(2020) dataset to create a complete county dataset for additional visualizations.

Our model predicted 18 of the 40 counties will vote Democrat and 22 of the 40 counties will vote Republican.



CHALLENGES AND LIMITATIONS

Data only available for 40 of 67 counties (populations over 65,000)

2020 Census Data unavailable due to COVID-19 (2019 data used)

High dimensionality of political data makes predictions complex

Real-time model application may be difficult

CONCLUSIONS & EXTENSIONS

The model shows potential for predicting election outcomes based on demographic and socioeconomic factors. Such a model could aid in understanding voting behaviors and trends, facilitating informed decision-making for electoral engagement. Incorporating data from all counties could enhance accuracy. The model also identifies key areas for targeted campaigns, optimizing resource allocation and outreach based on influential features.

THANK YOU FOR YOUR ATTENTION.

THE FLOOR IS OPEN FOR QUESTIONS