Iowa’s first congressional district

Project report

Dec 04, 2024

Jen-Kai Wang A20576222

Introduction

This project aimed to predict key electoral outcomes for Iowa’s First Congressional District by leveraging demographic and historical election data. Our primary objectives were to model three distinct targets: (1) voter turnout percentage, (2) the likely presidential party winner, and (3) the likely representative party winner. Focusing on party-level outcomes rather than specific candidates provided a more stable and generalized framework, since party affiliations and their associated ideologies often endure beyond individual election cycles.

Our team, composed of Naser Alkuhili (AI major), Jen-Kai Wang (CS major), Ping-Chun Shih (CS major), and Sakher Yaish (AI major), undertook a multifaceted approach. We integrated demographic data sourced from the Census API, historical election returns from official state election boards, and applied multiple modeling techniques including Random Forest Regressor, neural networks, and spatial comparisons. This combined approach aimed to identify patterns in electoral data and understand how demographic features influence voting behavior.

Project Approach

Our methodology began with a broad survey of data sources from multiple states (Iowa, Missouri, Illinois) to assemble a dataset with sufficient variance in demographic and electoral characteristics. From this, we performed exploratory data analysis (EDA) to understand how demographics, such as race, age distribution, educational attainment, and median income, correlate with voting behavior and turnout.

Once we had a clearer picture, we employed spatial analysis to identify and compare districts demographically similar to Iowa's First District. By measuring demographic similarity via Euclidean distance in feature space, we identified several comparable districts. This similarity-driven selection helped ensure that the training data captured patterns likely to be present in our target district. We then refined our models iteratively, testing various machine learning and neural network architectures to find the most reliable predictors of election outcomes.

Data Sources Explored

Data Sources:

- **Iowa**: Secretary of State election results, providing comprehensive district-level outcomes.

1- https://sos.iowa.gov/elections/results/index.html#2

- **Missouri**: Official state election archives with county-level and district-level returns.

1- https://www.sos.mo.gov/elections/resultsandstats/previousElections

2- https://uselectionatlas.org/RESULTS/state.php?fips=29&year=2020&f=0&off=0&elect=0

- **Illinois**: State Board of Elections results, allowing comparison across a neighboring state with demographic overlap.

1- https://www.elections.il.gov/electionoperations/votetotalsearch.aspx

We augmented the election data with demographic data from the United States Census via the Census API. This included features like total population, racial composition, age distributions (18–34, 35–64, 65+), gender ratios, median household income, educational attainment, and unemployment rates. Combining these datasets allowed us to explore not just past voting patterns, but also the socio-economic contexts that might influence those outcomes.

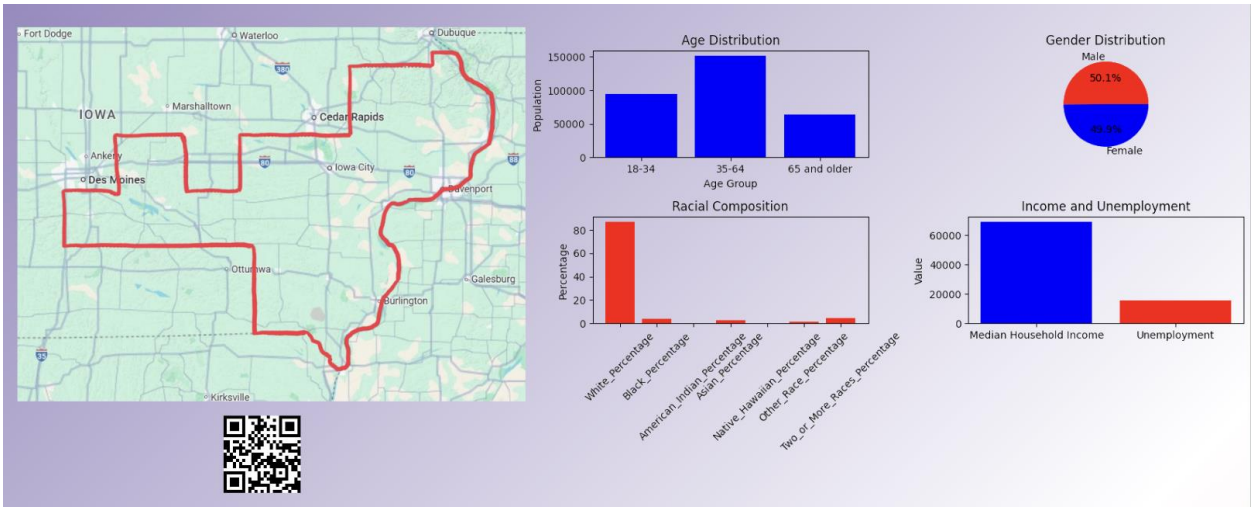
Exploratory Data Analysis:

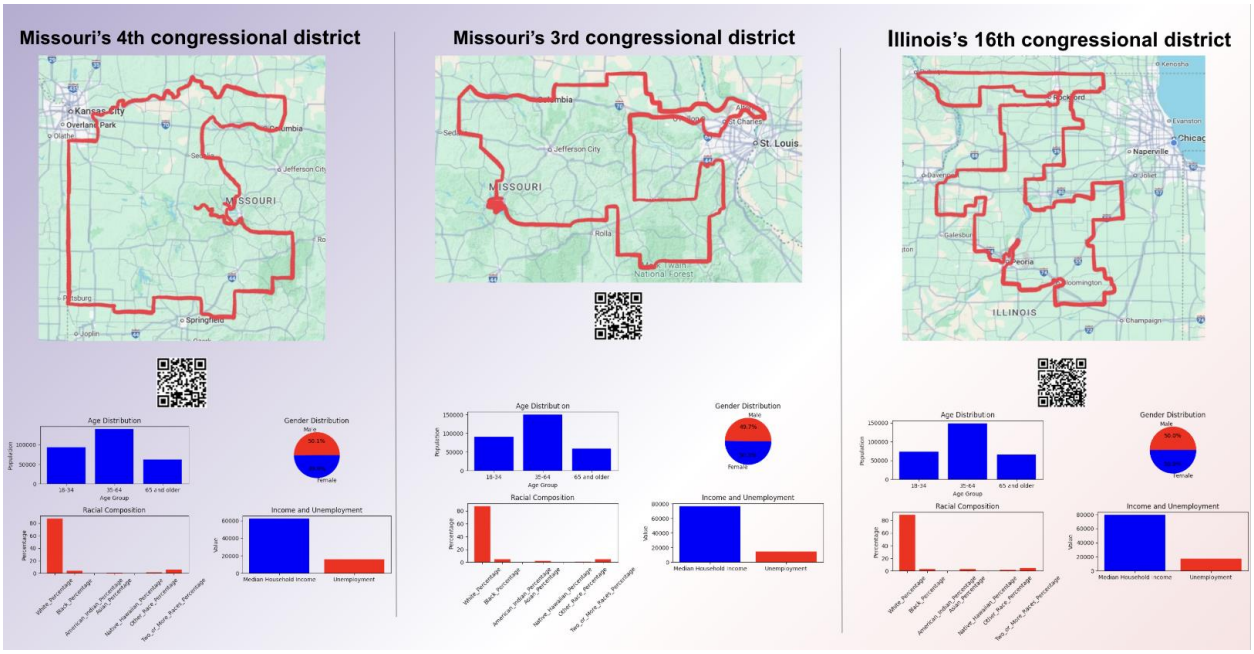
EDA involved examining correlations and distributions:

- **Demographic Trends**: Certain age brackets (e.g., 18–34) and income levels showed patterns related to turnout. Higher median household income and certain educational attainment levels often correlated with slightly higher voter turnout rates.

- **Racial Composition**: Districts with higher diversity sometimes displayed different voting preferences, potentially reflecting national patterns at a local scale.

- **Spatial Patterns**: Visualizing election results on maps revealed clear urban-rural divides. More urbanized counties within the district tended to lean Democratic, while rural counties were more consistently Republican.


Fort Dodge
Waterloo
Dubuque
IOWA
Marshalltown
Cedar Rapids
Ankeny
Des Moines
Iowa City
Davenport
Ottumwa
Galesburg
Burlington
Kirksville
Age Distribution
Population
150000
100000
50000
0
18-34
35-64
65 and older
Age Group
Gender Distribution
Male
50.1%
49.9%
Female
Racial Composition
Percentage
80
60
40
20
0
White_Percentage
Black_Percentage
American_Indian_Percentage
Asian_Percentage
Native_Hawaiian_Percentage
Other_Race_Percentage
Two_or_More_Races_Percentage
Income and Unemployment
Value
60000
40000
20000
0
Median Household Income
Unemployment


Missouri's 4th congressional district
Kansas City
Overland Park
Columbia
Jefferson City
MISSOURI
Springfield
Missouri's 3rd congressional district
St. Louis
Jefferson City
MISSOURI
Rolla
Illinois's 16th congressional district
Rockford
Chicago
Naperville
Peoria
ILLINOIS
Champaign

This thorough EDA guided feature selection and model development, ensuring we included the most relevant predictors and understood their potential influence.

Data Selection and Cleaning

Data Cleaning Steps:

- **Consistency Checks**: Ensured uniform naming conventions for demographic variables across states.

- **Missing Values**: Imputed or removed incomplete records to maintain a high-quality dataset.

- **Outlier Handling**: Examined outliers in turnout or demographic factors and assessed whether to retain them as part of the natural variation.

- **Feature Engineering**: Derived meaningful variables, such as calculating demographic age ranges, percentages of various racial groups, and converting absolute counts into proportions.

Data Organization:

After cleaning, the data was organized into a single cohesive dataset linking election results (presidential and representative party winners, turnout rates) to demographic and socio-economic features. Standard scaling was then applied to ensure all features had comparable scales, benefiting models, especially neural networks, that are sensitive to input magnitudes.

Relation of Selected Data:

All selected features related to broad categories that could influence voting behavior:

- **Socio-Economic Status (income, unemployment)**: May affect political priorities (e.g., economic policies, job growth).

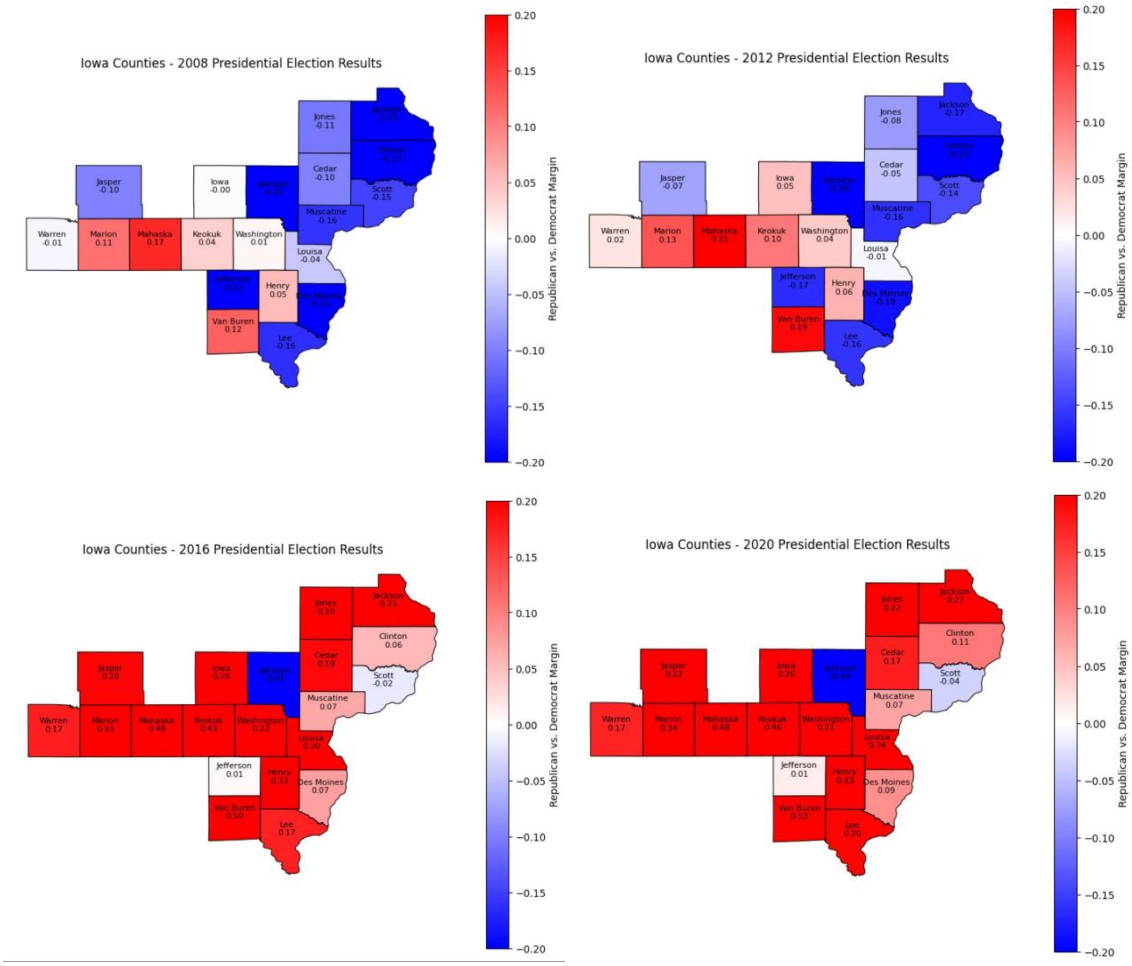
- **Demographic Structure (age distributions, gender ratios)**: Different age groups and gender balances can favor certain policies or candidates.

- **Racial and Ethnic Composition**: Historical data indicates some racial groups may align more consistently with one party.

By including these features, we provided the models with a comprehensive view of the electorate, increasing the likelihood of capturing underlying patterns in voting behavior.

Spatial Analysis

Our spatial analysis centered on Iowa's First Congressional District, examining the distribution of electoral outcomes and demographic factors at the county level. We created a series of choropleth maps to visualize the partisan margins (Republican vs. Democratic) across multiple election cycles (e.g., 2008, 2012, 2016, and 2020). Each county was shaded according to its voting margin, enabling us to quickly identify patterns and shifts over time.



By studying these maps, we observed how certain counties consistently leaned toward one party, while others fluctuated between election years. This allowed us to discern potential underlying demographic or socio-economic trends influencing electoral behavior. For example, some previously competitive counties gradually became more partisan, while others showed more subtle swings. Through these visualizations, we gained a clearer understanding of the district's internal electoral dynamics, patterns in voter turnout, and evolving political landscapes.

This localized spatial analysis, focusing solely on our target district rather than comparing it to others, reinforced the insights drawn from our modeling process. It helped ensure that the data-driven methods we employed were grounded in a concrete understanding of how political preferences vary not just by demographic attributes, but also by geographic location within the district.

Modeling Approach

Models Considered:

1. Voter Turnout Prediction

- **Initial Attempts**: We initially considered a neural network (Multi-Layer Perceptron Regressor) to predict voter turnout as a continuous variable. While this method could model complex non-linear relationships, we also wanted to explore other robust techniques

- **Neural Network Selection**: To refine our approach, we introduced a Random Forest Regressor and compared its performance against the neural network. The Random Forest model predicted a turnout of approximately 69.52%, which aligned more closely with the actual turnout of 72.77% than the neural network's prediction.

- **Rationale**: Random Forests, as ensemble methods, often provide strong predictive performance and can handle complex feature interactions without intensive hyperparameter tuning. In this case, the Random Forest Regressor offered a better balance of accuracy and complexity, making it the more suitable choice for our turnout prediction task.

2. **Presidential Winner Prediction**

- **Model**: A neural network classifier was built using Keras/TensorFlow. This approach allowed us to model the problem as a binary classification: predicting whether the Democratic or Republican party would prevail in the presidential race within the district.

- **Design**: The model included multiple dense layers with dropout to reduce overfitting, aiming to improve generalization.

- **Reasoning**: Party-level classification is well-suited to neural networks, as they can handle subtle interactions between demographic factors and likely political leanings.

3. **Representative Winner Prediction**

- **Model**: The approach mirrored the presidential prediction model, using a neural network classifier with similar architecture and training procedures.

- **Features & Training**: Emphasized demographic and economic indicators, attempting to discern how these features might tilt the legislative outcome.

Training and Testing:

- Data was split and scaled. We performed cross-validation on the turnout prediction model to ensure robust performance estimates.

- For classification tasks (presidential and representative predictions), we utilized label encoding and appropriate loss functions (binary cross-entropy) along with accuracy metrics.

- Our final chosen models for each task reflected the best compromise between complexity and predictive accuracy, with neural networks chosen over simpler methods due to their stronger performance and lower error.

Results

Turnout Prediction:

- **Chosen Model**: Random Forest Regressor.

- **Sample Result**: Actual turnout ~72.77%, predicted ~69.52%.

- Although slightly lower than the actual figure, this prediction comes much closer to the real turnout than our previous neural network attempt.

- The remaining discrepancy may still reflect factors not represented in the demographic and socio-economic data, such as local campaign activities or last-minute events on election day.

Presidential Winner Prediction:

- **Chosen Model**: Neural network classifier.

- **Results**: Using the current data from Iowa's First Congressional District, our model predicted a Republican party victory for the presidential outcome. The actual outcome was also Republican.

Representative Winner Prediction:

- **Chosen Model**: Neural network classifier.

- **Results**: Similarly, using the same current data for the representative race, the model predicted a Republican winner. The actual result for the representative race was also Republican.

Assessment and Discussion:

Our models performed well, especially in predicting the direction of party winners. The turnout model also showed promise, though it was slightly off from the actual numbers. Notably, we did not include polling data, which can be a strong predictor of turnout and outcomes. The complexity of human voting behavior, shaped by local issues, campaign efforts, weather conditions on Election Day, and last-minute events, means some variance is expected.

Factors impacting the results positively included the robust demographic features, which provided a strong, static backbone of electoral tendencies. Negatively, the absence of granular polling data, volunteer canvassing efforts, fundraising totals, and local events reduced the model's ability to perfectly mirror real-world variability. County-level granularity, rather than precinct-level, also limited the fine-resolution insights into within-district variations.

Conclusion and Discussion

Accomplishments:

- We successfully built predictive models for turnout, presidential, and representative races in Iowa's First Congressional District.

- The neural network approach captured complex demographic and socio-economic relationships, producing results that closely aligned with known electoral outcomes.

What Worked Well:

- The chosen neural network architectures were flexible and performed better than simpler linear methods.

- Party-level prediction proved robust, demonstrating that demographic and historical data alone can be strong indicators of electoral leaning.

Challenges:

- **Data Collection**: Gathering consistent, high-quality data from multiple states and various official sources was time-consuming and sometimes required creative problem-solving.

- **Granularity and Completeness**: County-level data, while easier to aggregate, may not represent the full complexity of local voting behaviors.

What Could Be Done Differently:

- In future iterations, using precinct-level data instead of county-level data could substantially increase the model's accuracy by capturing more localized demographic variations.

- Incorporating additional data sources, such as polling, fundraising, or candidate campaign stop information, might refine the turnout and winner predictions.

Future Work

To further improve and extend this project:

- **Refine Granularity**: Incorporate precinct-level rather than county-level data.

- **Broaden Feature Set**: Add real-time polling data, historical weather conditions on Election Day, and socio-political indicators (e.g., public sentiment analysis from social media).

- **Generalization**: Apply the same modeling approach to additional districts and states to verify the broader applicability and adaptability of the methods.

- **Integrate Temporal Dynamics**: Consider modeling changes over multiple election cycles to predict long-term trends rather than single-year outcomes.

In summary, this project laid a solid foundation for data-driven electoral forecasting. By employing spatial comparisons, comprehensive demographic data, and advanced modeling techniques, we have shown that machine learning models, particularly neural networks, are capable of providing meaningful insights into electoral outcomes, even in the absence of direct polling data.

Team Performance

How did the team perform

What worked well:

We effectively divided the various tasks required for this project. Our approach involved identifying three similar districts to improve the accuracy of our predictions. To achieve this, we collected data from these districts and performed data cleaning for each of them. Similarly, the tasks related to prediction results and model training were also carefully divided among the team. In the end, we successfully completed the project.

What could have been improved:

The main area for improvement is time management. Since each team member was responsible for a different part of the project, balancing this work alongside other ongoing coursework proved challenging. As a result, there were times when we were unable to complete tasks on schedule.

Work breakdown:

Parts of the project each member do:

Naser Alkuhili: voter turnout prediction model, GIS, collect and cleaning Iowa 1st district data, midterm report, final poster

Sakher Yaish: model predicting the presidential race outcome, collect and cleaning Missouri 3rd district data, midterm report, final poster

Jen-Kai Wang: predicting the representative race outcome, collect and cleaning Missouri 4th district data, midterm report, final poster

Ping-Chun Shih: predicting the representative race outcome, collect and cleaning Illinois 16th district data, midterm report, final poster

Percentage of the work for the entire project each member do:

Naser Alkuhili: 25%

Sakher Yaish: 25%

Jen-Kai Wang: 25%

Ping-Chun Shih: 25%

Leader of the team:

Naser Alkuhili

Each team member grade:

Naser Alkuhili:

Communication: A

Technical quality: A

Follow-through: A

Sakher Yaish:

Communication: B

Technical quality: A

Follow-through: A

Ping-Chun Shih:

Communication: A

Technical quality: A

Follow-through: A

Code:

https://drive.google.com/drive/folders/19sSP5YlgkxWhptjGsur6bDney5Ud_ZGL?usp=sharing

References

1. Iowa Secretary of State. (n.d.). *Election Results.* Retrieved from https://sos.iowa.gov/elections/results/index.html#2

2. Missouri Secretary of State. (n.d.). *Previous Elections Results and Statistics.* Retrieved from https://www.sos.mo.gov/elections/resultsandstats/previousElections

3. U.S. Election Atlas. Retrieved from https://uselectionatlas.org/RESULTS/state.php?fips=29&year=2020&f=0&off=0&elect=0

4. Illinois State Board of Elections. (n.d.). Retrieved from https://www.elections.il.gov/electionoperations/votetotalsearch.aspx