MATH-475 Final Report: Predicting Voting Behavior in the 2020 U.S. Presidential Election

Levi Sweat

1. Introduction

This project analyzed voting patterns in the 2020 U.S. Presidential Election at the county level. The initial goal was to explore how the COVID-19 pandemic influenced voter behavior, specifically focusing on the Republican/Democrat vote share. However, due to data limitations (2024 results were unavailable), the analysis pivoted to predicting the 2020 Democratic vote percentage across counties using demographic, economic, and COVID-19 data. Various machine learning models, including linear regression, deep learning models, and binary classification models were developed and evaluated to predict voting percentages. The GitHub Repository containing the Jupyter Notebook, data, project proposal, and this report can be found here.

2. Data Exploration & Preprocessing

The dataset comprised two sources: county-level election data and COVID-19 case/death data. Key preprocessing steps included:

- Combining Datasets: Aggregated COVID-19 data by county/state and merged it with election data, yielding 2,620 matching counties.
- Feature Cleaning: Dropped irrelevant columns and standardized column names to snake_case.
- Type Conversion: Converted percentages to fractions and cleaned numeric columns with missing or invalid values.
- Feature Engineering: Combined education levels into a single weighted variable, reducing redundant columns in data.
- Outlier Identification: Used Isolation Forest to successfully identify and remove outliers that would poorly impact future models.
- Correlation Analysis: Identified weak correlations between most features and target variables (Democratic/Republican vote percentages), guiding feature selection.

3. Model Selection & Training

Many models were tested. We began with a deep learning model using ReLU activation, which provided reasonable results. We then experimented with a variety of different activation methods, hidden layers, epochs, target values, and prediction values. We found the most accurate model using Swish activation only predicting the Democratic vote percentage and removing outliers from the original dataset. We then created a binary classification model that attempted to predict whether the Democratic vote percentage would exceed 50%, achieving high accuracy but low F1 scores, indicating a trade-off between precision and recall. Each deep learning model used standardized features, early stopping to prevent overfitting, and dropout layers for regularization. We finished the modeling by creating a simple linear regression model with the features most correlated to the target. This offered poor results, which is unsurprising as the correlation values of even the best features were quite low.

4. Performance Evaluation

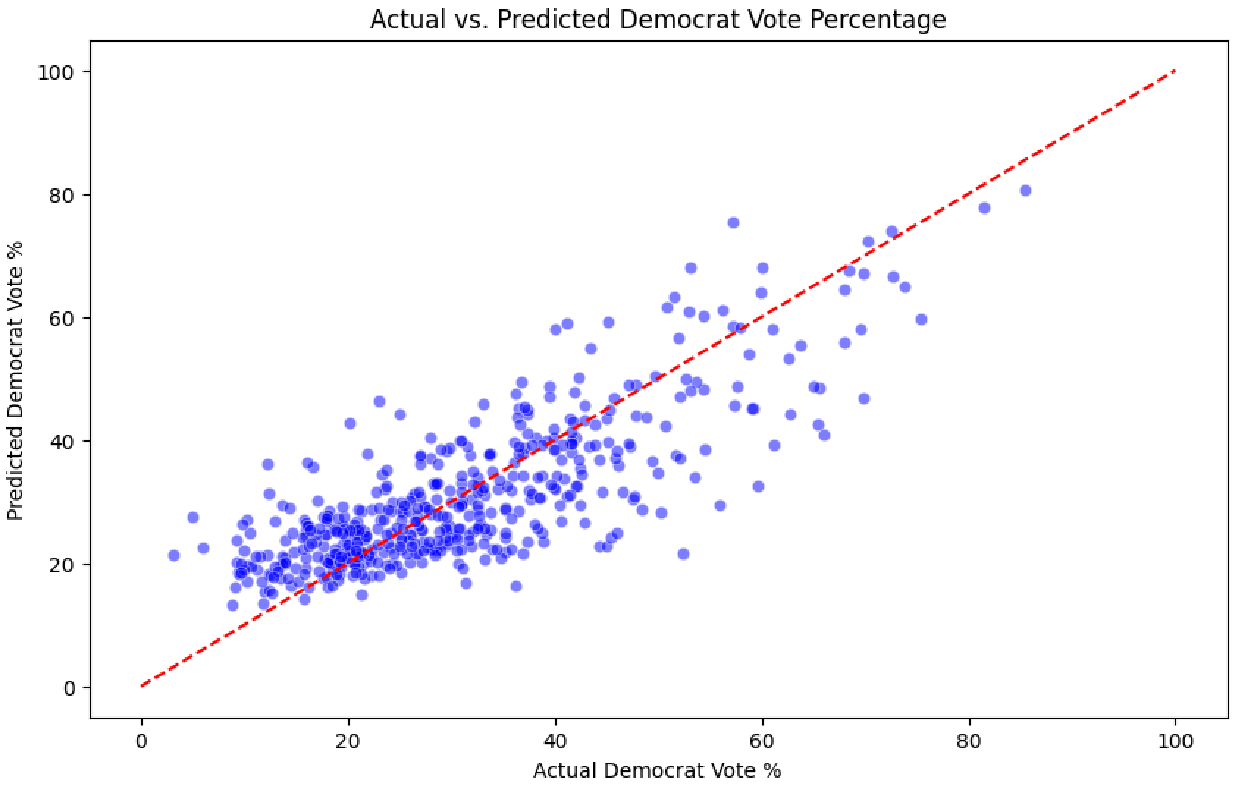
The Swish-based deep learning model predicting Democratic vote percentage and excluding outliers achieved the best results:

Mean Squared Error (MSE): 55

Mean Absolute Error (MAE): 5.8

R^2 Score: 0.72

This performance is notable given the weak correlations between features and the target variable. For binary classification, while the accuracy was high (>90%), the low F1 score highlighted issues with imbalanced recall and precision. Below is a graph comparing the actual Democratic vote percentage compared to the actual Democratic vote percentage.



5. Challenges & Future Work

Challenges:

- Data Limitations: Weak correlations between features and targets limited model accuracy. 2024 election data yet to be publicly accessible for all counties, resulting in a major data limitation early into the project.
- Outliers: Identifying and removing outliers significantly impacted performance.
- Feature Engineering: Combining and selecting meaningful features required iterative experimentation.

Future Work:

- Data Augmentation: Include additional demographic or behavioral data to enhance feature correlations.
- Time-Based Analysis: Considering how the COVID-19 dataset was originally formatted, could analyze how the Pandemic shifted over time and how it affected 2020 election results compared to 2024.

- 2024 Election Results: As 2024 election results are updated and become more publicly accessible, use models trained on 2020 election to predict 2024's election results (and see how they perform given actual results).
- Binary Classifications: Further analysis could be completed to create better/more accurate binary prediction models.

Conclusion

This project demonstrated the potential of machine learning models in predicting voting behavior. Despite data challenges, the deep learning models, particularly with Swish activation and outlier removal, showed promising results. Future improvements could enhance the robustness and interpretability of predictions.