# ISyE 6740 – Spring 2025
# Final Report

**Team Member Names:** Matthew Cheung

**Project Title:** Forecasting Voter Turnout and Presidential Election Outcomes in Clinton Township: A Machine Learning Approach

## Contents

## 1 Problem Statement

In an increasingly complex and politically charged global environment, the act of voting is one of the most critical mechanisms for shaping democratic societies. Within the United States, election outcomes have significant influence over economic policy, domestic and international relations, and the overall direction of public debate. At the center of this process lies the Presidential General Election, held every four years in November, which serves as both a unifying and divisive moment in American society.

These types of elections generate the most media coverage at the national level but hold an equally important role at the local level. Communities like Clinton Township—where I live and work—play a crucial part in shaping the broader political landscape through local participation. Understanding voter behavior in these local contexts can offer key insights into larger trends.

This project aims to leverage local election data from Clinton Township spanning 2012 to 2024 and apply a range of machine learning techniques to forecast two key outcomes for the 2028 Presidential General Election:

- Voter turnout in Clinton Township
- The projected winning party based on predicted vote shares

This project combines real historical data with thoughtfully synthesized variables to build predictive models and seeks to contribute a data-driven perspective on voter participation and political outcomes within my local community.

## 2 Data Exploration

### *2.1 Data Source*

The main data source for the project is the Clinton Township Elections Page [1] which hosted election results for previous elections dating back to 2020. This data detailed voter turnout numbers, total votes cast, and vote share per party. Contacting the Clerk's Department directly, additional data covering the age and gender breakdowns of the voters was supplied and subsequently added to the data set as additional features.

The election data from 2020 to 2024 was somewhat limited in terms of being able to train and develop machine learning models. Therefore, the range of the election years was extended back to the year 2012. By referencing the Macomb County Election archive [2], some of the data pertaining to the Clinton Township was able to be spliced into the data set. The remaining missing data for the 2012-2018 elections was imputed using the following logical assumptions based off the historical data available:

$$\text{Turnout}_{\text{Clinton Township}} = \text{Turnout}_{\text{Macomb County}} + 2.0\%$$

$$\text{Rep\_Share}_{\text{Clinton Township}} = \text{Rep\_Share}_{\text{Macomb County}} + 2.0\%$$

$$\text{Dem\_Share}_{\text{Clinton Township}} = 100\% - \text{Rep\_Share}_{\text{Clinton Township}}$$

$$\text{Voted} = \text{Registered Voters} \times \left(\frac{\text{Turnout}}{100}\right)$$

Historical data suggested that Clinton Township typically experienced voter turnout approximately two percentage points higher than the countywide average. Similarly, Clinton Township has historically leaned about 2% more Republican in its voting patterns compared to overall Macomb County results, and this adjustment was applied when imputing party vote shares.

### *2.2 Data Processing*

After sourcing the data from the various election sites and imputing missing data, several pre-processing steps remained to prepare the final datasets. The election records lacked important features to aid in prediction. To address this, synthetic variables were used to fill these features including flags denoting whether the election was considered a major race, whether COVID impacted voter behavior, and weather severity on Election Day.

Also included were voter demographic features such as age group distributions, median income, and education levels sourced from the U.S. Census Bureau [3] and Presidential betting odds for the 2028 election [4] to declare a potential winning political party.

Implied probabilities were calculated from betting odds to generate normalized predictor variables for party likelihood. For positive American odds (e.g., +120) and negative odds (e.g., -150), implied probabilities were computed using the following formulas:

$$\text{Implied Probability (Positive Odds)} = \frac{100}{\text{Odds} + 100}$$

$$\text{Implied Probability (Negative Odds)} = \frac{|\text{Odds}|}{|\text{Odds}| + 100}$$

These implied probabilities for Republican and Democratic candidates were included as additional features in the final vote share prediction model.

Ultimately, the data was split into two separate datasets. In the Appendix, Table 9.1 highlights the features of the dataset that was used to model and forecast overall voter turnout. Similarly, Table 9.2 incorporated the predictions from the Voter Turnout Dataset to predict the vote share between the Republican and Democratic Party. Together, both datasets help to project the total votes per party and identify the predicted winning political party for the 2028 Presidential General Election in Clinton Township.

## 3 Methodology

There were two main goals within this project: predicting voter turnout percentage and estimating party vote shares for the 2028 Presidential General Election in Clinton Township. By predicting this data, we would be able to estimate the vote totals and declare a winning party for the next presidency.

The modeling pipeline began with building and refining the dataset. Feature engineering included one-hot encoding categorical election types, standardizing numeric inputs, and imputing missing values. Synthetic variables were added to the data set, such as weather factors, to strengthen the dataset and provide better predictions for the future election.

A variety of machine learning techniques were implemented to analyze the data including linear regression, ridge and lasso regularized models, random forests, support vector regression (SVR), and k-nearest neighbors. Due to the relatively small size of the training dataset, we used Leave-One-Out Cross-Validation (LOOCV) to fairly evaluate model performance without discarding valuable data. The top performing models were used to predict voter turnout which helped to estimate the vote share across parties and predict a winner.

## 4 Results and Evaluation

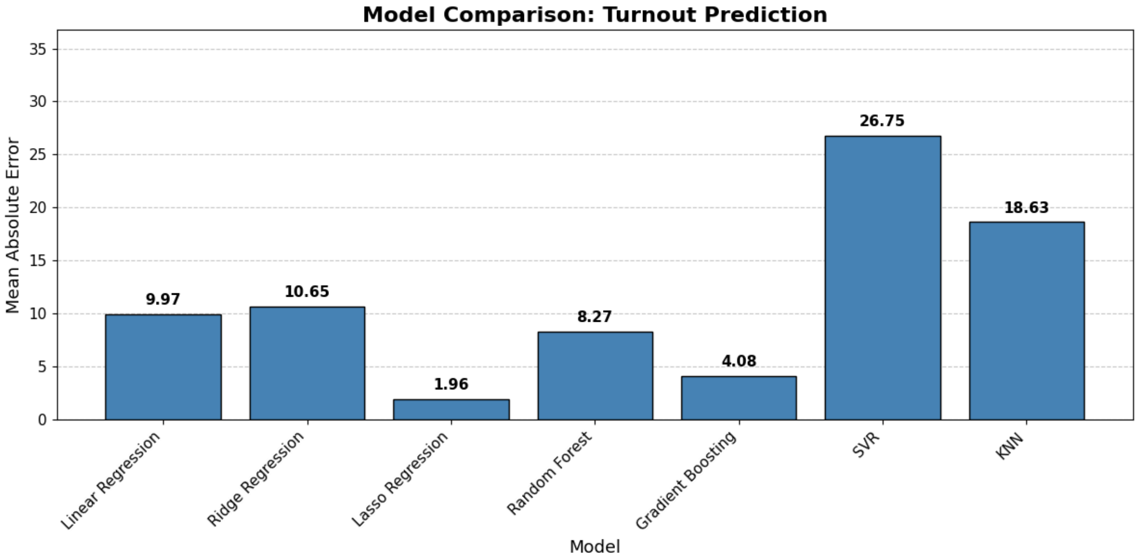
### *4.1 Voter Turnout Prediction*

The first dataset, the Voter Turnout Dataset, was trained against a variety of Machine Learning models including Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regression (SVR), and K-Nearest Neighbors (KNN). Due to the small nature of the dataset, it was best practice to use Leave-One-Out Cross-Validation (LOOCV) to fairly evaluate model performance because a k-fold Cross Validation would leave the dataset with unreliable folds. Again, due to the limited nature of the dataset, using $R^2$ as a performance metric was not reliable, so instead we opted for Mean Absolute Error (MAE). Mean absolute error can be calculated as the following where $n$ = the number of data points, $y_i$ = the actual observed value, and $\hat{y}_i$ = the predicted value:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Using LOOCV and MAE allowed us to train the model on $n - 1$ samples, predict on the 1 held-out sample, and compute the Mean Absolute Error for the prediction error. The process was repeated for $n$ samples. The comparative MAE results for voter turnout prediction models are visualized in Figure 4.1.

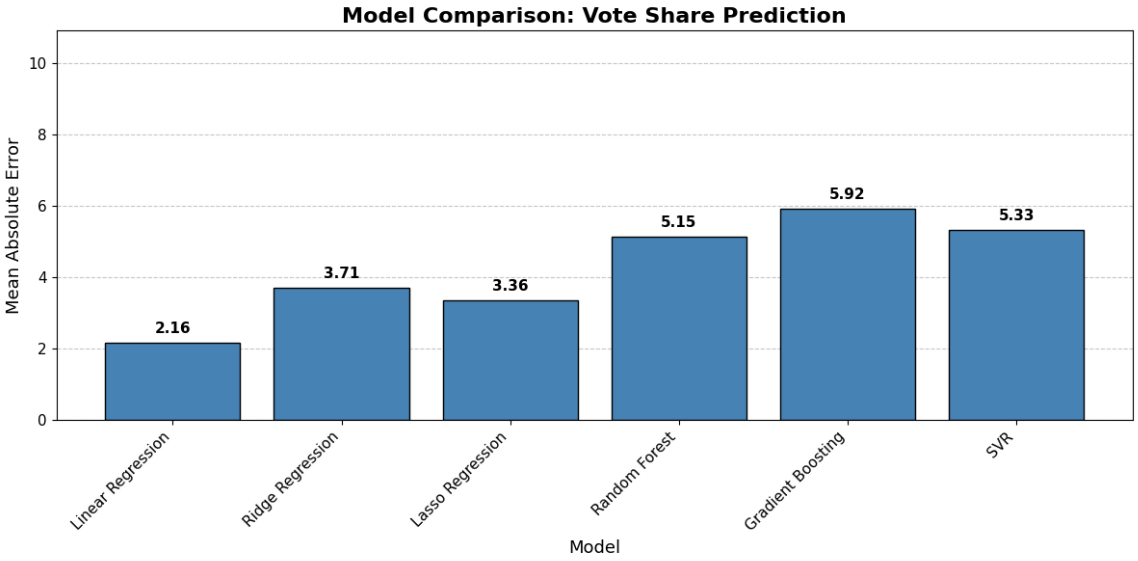
Figure 4.1 – Model Comparison for Turnout Prediction



### *4.2 Vote Share Prediction*

Similarly, the same models were trained to predict for the Vote Share between the Democratic and Republican parties. The same validation logic applied: LOOCV and MAE were used for evaluation due to the small sample size. In this case, K-Nearest Neighbors failed during LOOCV due to instability caused by the limited number of samples. Nevertheless, the other models produced valid MAE scores, which are displayed in Figure 4.2.

Figure 4.2 – Model Comparison for Vote Share Prediction



### *4.3 Model Selection Summary*

The MAE scores for each model, across both voter turnout and vote share prediction tasks, were compiled into Table 4.1 below:

Table 4.1 – Prediction Model Scores

| **Model** | **MAE (Turnout Prediction)** | **MAE (Vote Share Prediction)** |
|---|---|---|
| Linear Regression | 9.972 | 2.160 |
| Ridge Regression | 10.649 | 3.707 |
| Lasso Regression | 1.960 | 3.358 |
| Random Forest | 8.270 | 5.154 |
| Gradient Boosting | 4.078 | 5.919 |
| SVR | 26.747 | 5.334 |
| KNN | 18.629 | NaN |

Based on the MAE scores:

- **Lasso Regression** was selected as the final model for predicting voter turnout.
- **Linear Regression** was selected as the final model for predicting vote share.

These models were subsequently used to forecast turnout percentages, predict party vote shares, and estimate the winner of the 2028 Presidential Election in Clinton Township.

## 5 Prediction of 2028 Presidency

Using the previously selected best-performing models for each dataset, forecasts were made for the 2028 Presidential General Election in Clinton Township.

First, Lasso Regression, selected from LOOCV, was used to develop the profile of the election for 2028. The predicted number of registered voters was 84,005 with a 54.73% turnout. Next, the predicted turnout percentage was used as an input for the Linear Regression model of the Vote Share dataset. Including the implied betting odds, voting share between the two Political Parties was predicted for. From there, using the following formulas we were able to calculate our final predictions:

$$\text{Votes Cast} = \text{Registered Voters} \times \left(\frac{\text{Predicted Turnout}}{100}\right)$$

$$\text{Democratic Votes} = \text{Votes Cast} \times \left(\frac{\text{Democratic Share}}{100}\right)$$

$$\text{Republican Votes} = \text{Votes Cast} \times \left(\frac{\text{Republican Share}}{100}\right)$$

The resulting prediction metrics have been summarized in Table 5.1 below:

Table 5.1 – Prediction Summary

| **Metric** | **Value** |
|---|---|
| Predicted Registered Voters | 84,005 |
| Predicted Turnout % | 54.73% |
| Predicted Votes Cast | 45,976 |
| Predicted Democratic Share % | 54.29% |
| Predicted Republican Share % | 45.71% |
| Projected Democratic Votes | 24,960 |
| Projected Republican Votes | 21,016 |
| Predicted Winner | Democrat |

Overall, the predicted voter turnout for the 2028 Presidential General Election in Clinton Township was **54.73%**, and the projected winner was the **Democratic Party**, based on forecasted vote shares and total votes cast.

## 6 Analysis of Models

Through the various machine learning models explored for voter turnout and vote shares, Lasso regression stood out for the turnout prediction, while Linear regression was selected for vote share modeling. The Lasso regression offered strong performance on the limited dataset by applying $L1$ regularization.

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\}$$

The $L1$ penalty encourages a simpler model by shrinking some coefficients exactly to zero, effectively selecting a subset of predictors. This allows the model to perform feature selection while preventing overfitting. In this dataset, avoiding overfitting was key due to the limited number of samples. For vote share, Linear Regression was selected because it was able to provide clear relationships in the dataset without great model complexity.

The other tested models, such as Ridge Regression, Random Forest, Gradient Boosting, Support Vector Regression (SVR), and K-Nearest Neighbors (KNN), that were evaluated using Leave-One-Out Cross-Validation (LOOCV) provided valuable insight to the problem but were not selected for prediction. Some of the more complex models scored well in terms of Mean Absolute Error, but risked overfitting or instability given the limited and partly synthetic dataset. Simpler models ultimately balanced predictive performance and generalization ability more effectively for this project.

Through the analysis of the machine learning models, it was important to recognize the limitations of the datasets. The relatively small dataset size constrained model training and evaluation, and the inclusion of off-year and primary elections introduced noise that may have skewed the predictions. While the final selections were appropriate for the available data, future models with a longer historical dataset focused solely on Presidential elections could further improve the modeling and forecasting accuracy.

## 7 Conclusion

The goal of this project was to forecast the outcome of the 2028 Presidential General Election in Clinton Township using a mix of historical election data and synthetic features based on real-world factors. To do this, several machine learning techniques were deployed to predict the overall voter turnout and the winning political party.

The final predictions predicted a voter turnout of approximately **54.73%** or about **45,976 votes**. Based on the predicted vote shares between the Democratic and Republican party, the **Democrats** were projected **24,960** of these votes, edging out the Republican party with **21,016** projected votes.

The overall outcome of these forecasts are a bit surprising given real-life context and the slight Republican lean that was shown in the exploration of the historical data. However, there is important context to the interpretation of these results. There were several limitations to the modeling including:

- There was a small number of historical Presidential elections that had data available, which limited quality training data
- Due to the limited number of Presidential data available, there was the inclusion of off-cycle elections (primaries and special elections) which may have introduced noise unrepresentative of typical Presidential Election voter behavior
- Only the Democratic and Republican parties were considered, and other third-party candidates were not included
- The use of synthetic features, even with careful construction, introduces uncertainty

Despite the limitations, the project was able to successfully apply machine learning techniques to forecast potential local election results. Overall, the model evaluation and the considered limitations of the data sets set a solid baseline for future improvements. With expanded datasets focused exclusively on Presidential elections and further refinements in feature engineering, future iterations could yield even more accurate and reliable forecasts.

## 8 References

[1] Clinton Township Elections Department. *Past Election Results and Voter Turnout Data*. Retrieved from:
https://www.clintontownship.com/184/Elections

[2] Macomb County Clerk / Register of Deeds. *Macomb County Past Election Results Archive*. Retrieved from:
https://www.macombgov.org/departments/clerk-register-deeds/elections/past-election-results

[3] U.S. Census Bureau. *QuickFacts: Macomb County, Michigan*. Retrieved from:
https://www.census.gov/quickfacts/macombcountymichigan

[4] Sports Betting Dime. 2028 U.S. Presidential Election Odds Tracker. Retrieved from:
https://www.sportsbettingdime.com/politics/futures/us-presidential-election-odds/

## 9 Appendix

Table 9.1: Voter Turnout Dataset

| **Feature** | **Description** |
|---|---|
| Year | Year of the election |
| Turnout_Percent | Percent of registered voters who voted |
| Voted | Number of ballots cast |
| Major_Race_Flag | 1 if major race (presidential or midterm), 0 otherwise |
| COVID_Impact_Flag | 1 if COVID-19 affected election, 0 otherwise |
| Weather_Severity | 0 (normal), 1 (bad weather) |
| Percent_Age_18_29 | % of voters aged 18–29 |
| Percent_Age_30_44 | % of voters aged 30–44 |
| Percent_Age_45_64 | % of voters aged 45–64 |
| Percent_Age_65plus | % of voters aged 65 and older |
| Median_Income | Median household income |
| Percent_College_Grad | % of population with college degrees |
| Past_Turnout_1 | Turnout % from previous comparable election |
| Past_Turnout_2 | Turnout % from two elections prior |
| Election_Type_Presidential General | Indicator (1/0) for Presidential General elections |

Table 9.2: Vote Share Dataset

| **Feature** | **Description** |
|---|---|
| Turnout_Predicted | Predicted turnout percentage from turnout model |
| Major_Race_Flag | 1 if major race (presidential or midterm), 0 otherwise |
| COVID_Impact_Flag | 1 if COVID-19 affected election, 0 otherwise |
| Weather_Severity | 0 (normal), 1 (bad weather) |
| Dem_Implied_Odds | Implied probability of Democratic win from betting odds |
| Rep_Implied_Odds | Implied probability of Republican win from betting odds |
| Past_Dem_Share_1 | Democratic vote share from previous election |
| Past_Dem_Share_2 | Democratic vote share from two elections prior |