Who votes for whom and why?

Stanford CS229 Project

Juliette Coly
Department of Economics
Stanford University
jcoly@stanford.edu

1 Introduction

Who vote for whom and why? This question is at the core of political science (and the electoral game). Several schools of thoughts can be opposed when it comes to answering this question. The sociological explanation states that socio-demographic variables (income, religion, race, education, ...) are all what matters: "a person thinks, politically, as he is, socially. Social char- acteristics determine political preference" in the words of the seminal work in this area (Lazarsfeld et al., 1968).

I'll explore this mechanisms using novelly available data on French electoral outcomes since the 1789 French Revolution. My input variables are demographic and economic variables (income, age distribution, religion, ...) and my output variable is the electoral outcome of the 1981 presidential election at the municipality level. I'm running three regression algorithms (Ridge, Adaboost, multi-layer perceptron) to predict the share of votes that the leading candidate, F. Mitterrand, obtained in the first-round of the election.

2 Related Work

Paul Lazarsfeld was pioneer in electoral sociology. Finding that relatively few voting intentions switched during an election campaign, he fell back upon previously established demographic patterns of voting to explain their findings such as social status, religion, and place of residence (Lazarsfeld et al., 1968). This work is based on the descriptive analysis of 600 respondents in the US in the 1940s. This work is fundamental in terms of the explanation that it proposes for voting behavior but ir lacks a rigorous data analysis.

As documented by Cranmer (2019), machine learning algorithms have recently started to be used in political sciences but few papers have been interested in predicting electoral outcomes from socio-demographic features. The closest article is Sinha et al. (2020). Sinha et al. (2020) predicts the incumbent vote shares in the US 2020 presidential elections using national-level economic and non-economic factors of previous elections. My approach differs in which they exploit the difference between elections with national-level data (time series) while I am exploiting the difference between municipalities for a given election (the 1981 Presidential election). Another difference is that they focus on the United States while I am looking at France.

Some works have used machine learning to study the impact of social media on political campaigns. Brito and Adeodato (2023) uses sentiment analysis with data from Twitter, Facebook, and Instagram to predict election outcomes in Latin America. Ali et al. (2022) discusses and estimates the stability of social media approaches to forecast election results in Pakistan. These works are more interested in the impact of the political campaign through social media than the impact of socio-demographic variables.

3 Dataset and Features

Data source The data comes from Cagé and Piketty (2023). The authors numerized electoral and socio-economic data for more than 36,000 French municipalities between 1789 (the first post-French Revolution election) to 2022 (last legislative and presidential elections).

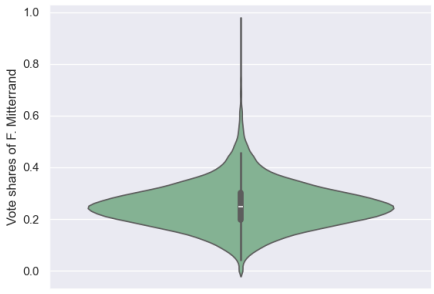
Output variable I am trying to predict the share of votes that F. Mitterrand (who ended up becoming President) obtained in the 1981 presidential election in each municipality. I focus on the share of votes instead of the winning candidate. The French system relies on the popular vote and not on an electoral college as the U.S one. Therefore, the share a candidate gets in a municipality is more informative than whether they arrived first.

Note that this is the result of the first round of a two-round electoral system. It is normal that F. Mitterrand mostly gets shares between 20% and 40 % of the votes. Figure 1a shows the distribution of the outcome variable, which has a median of about 23 %.

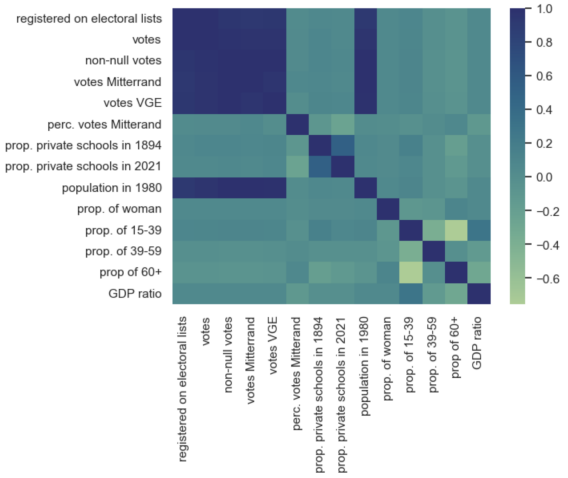
Input variables The data contains the following socio-economic information for each municipality [1]:

- Revenues: the revenues of the municipality as the fraction of France's revenues.
- Age: the percentage of 15-39; 39-59; and 60+ people.
- Sex: the percentage of man/woman.
- Percentage of private schools in 1894 and 2021. National statistics on religious belief being forbidden in France, the authors came up with proxies. The percentage of private school in the county is very correlated with how Catholic the county was.

Figure 1b displays the correlation between the variables of the dataset. There's not variable that is clearly correlated with the outcome variable (perc. votes Mitterrand).



(a) Distribution of the share of votes for the winning candidates, F. Mitterrand



(b) Correlation map between features.

Features and transformation The data contains 12,116 observations once the missing values are removed (I checked that there's no systematic biases in the observations that contain missing values). I kept 80% of the observations for training, about 9,600 of them, and allowed the 20% remaining to testing.

[1] More variables were available. However, including all of them was decreasing the sample size too much (only 400 observations were remaining). I thus kept a relative small number of variables to have enough observations.

I removed the following variables: population in 1980 and total number of votes as it is correlated with the number of registered voters[2]; the proportion of private schools in 1894 as it is correlated with the proportion of private schools in 2021.

I scaled the features variables by demeaning them and dividing by the standard deviations (note that all the features are numeric) as I am using a regularized regression and a neural network.

4 Methods

I try to predict the percentage of vote of the winning candidate (F. Mitterrand) in each municipality based on the socio-demographic features described earlier. I use a Ridge regressor, an Adaboost algorithm with decision trees, and a multi-layer perceptron regressor.[3]

Ridge regression Noting y_i the outcome variable and x_i the vector of features for municipality i, $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$, and $y = (y_1, \dots, y_n)^T$ the Ridge regression problem is written as

$$\min_{\beta} \frac{1}{2}||X\beta - y||_2^2 + \lambda||\beta||_2^2$$

where $||.||_2$ is the ℓ^2-norm and λ the regularization parameter. I am using a Ridge regression because the number of features and the number of observations are low in ML standards (respectively 6 and about 12,000), which make complex (and data-intensive) algorithms unsuitable.

Adaboost with decision trees Decision-tree regressors capture non-linear relationship between the outcome variable and the feature variables and are robust to outliers. However, a single decision tree tends to overfit the data. This is why I use the Adaboost ensemble method to aggregate the predictions from many weak learners. While its results are harder to interpret, Adaboost handles overfitting well and improves accuracy.

Multi-layer perceptron regressor I use a neural network with hyperparameter tuning to analyze the electoral data. This neural network, with its interconnected layers composed of neurons that apply the weights and biases, had its first layer described by

$$f\left(\sum_{i=1}^{n} w_i x_i + b\right),$$

where w_i represents the weights, x_i the input values, b the bias, and f the activation function.

5 Experiments / Results / Discussion

Since this is a regression problem, I use the mean-squared error (MSE) as metric, defined as:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ is the predicted value of the model.

Ridge regression I use cross validation to tune the hyperparameter λ. The parameter λ is in the range $\{0.1, 0.5, 1.0, 2.0, 5, 10\}$ and there are 5 cross-validation folds. The λ that minimizes the root mean square erros on the training data is $\lambda = 10$. The MSE onf the training data is $MSE_R^{TR} = 0.00824$. The MSE on the test data equals $MSE_R^T = 0.0086$.

[2]Note that the turnout was about 80%, so the number of registered voters and the number of actual votes are close

[3]I also tried a Lasso regression but I omitted this result due to limited space constraints in the report.

n. of voters	prop. private schools	prop. women	prop. 15-39	prop. 40-59	prop. 60+	GDP ratio
0.0001	-0.0071	0.0024	0.0042	-0.0006	-0.0013	-0.0043

Table 1: Coefficients of the Ridge regression

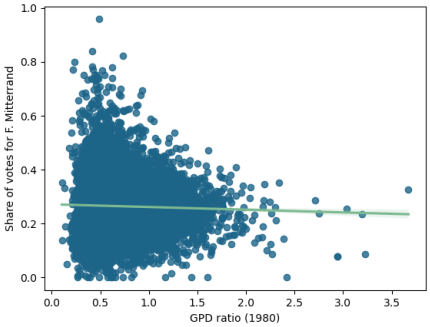


Figure 2: GDP and shares of votes for F. Mitterrand

Table 1 displays the coefficients of the Ridge regression associated to each variable. The variables the most positively correlated with voting for the winning candidate, F. Mitterrand, who was left-wing is the number of voters, the number of women, and the number of young (15-39 year-old) people. The richer (with a higher GDP ratio) and the more religious (with a higher proportion of private schools) the municipality, the fewer votes F. Mitterrand received.

However, looking at the relationship between a municipality relative wealth (GDP ratio) and the share of votes for F. Mitterrand, we see that the relationship cannot be really captured by a line (Figure 2). The triangular shape suggests that as a municipality gets richer, the number of votes decreases, but the high variance among poorer municipalities (between 0.5 and 1 GDP ratio) is not well-captured. This non-linear relationship suggests to use non-linear learning algorithms, such as Adaboost with regression trees.

Adaboost with regression trees I use cross-validation to tune the depth of the trees used as weak learners with 5 cross-validation folds. I used up to 100 trees.

The model that performs the best is made of stumps (depth 1 trees) with with a MSE on training equalling $MSE_A^{TR} = 0.085$ and MSE on the test data of $MSE_A^{T} = 0.0086$.

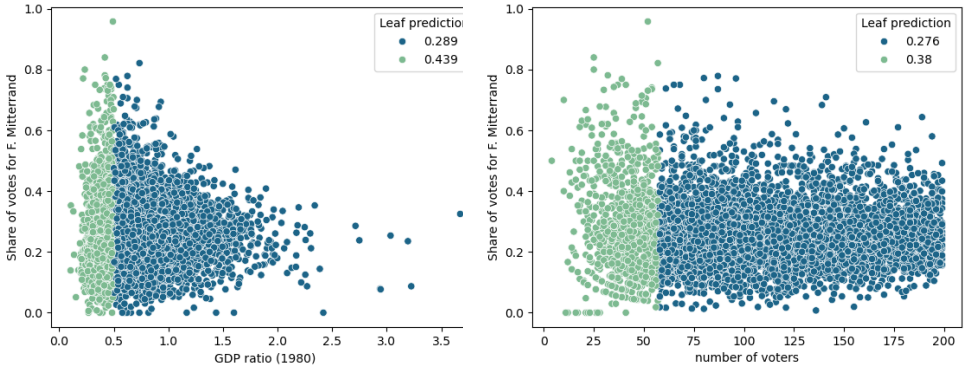
Table 2 contains the number of time each feature is used as the splitting variable [4] The number of voters and the wealth of the municipality are the most important splitting variables.

feature	count
number of voters	8
prop. of private schools (2021)	1
prop. of women (1980)	1
prop. of 15-39 (1980)	0
prop. of 40-59 (1980)	0
prop. of 60+ (1980)	1
gdp ratio (1980)	3

Table 2: Number of times each feature is chosen as the splitting variable

Figure 3a and 3b display the boundary decision and the value of the leaf prediction for two regressor trees. The decision tree using GDP ratio (figure 3a) captures relationship between votes and wealth better than the regression line (figure 2), which was nearly flat.

[4]Note that the algorithm used only 14 trees.

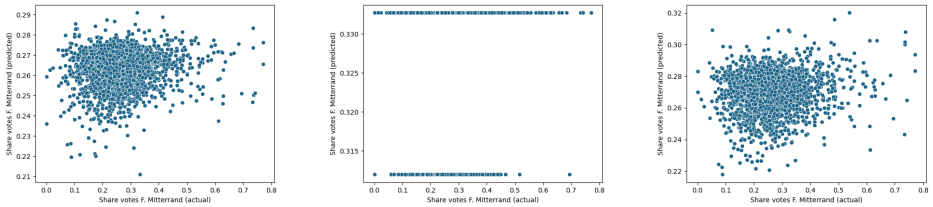


(a) Decision boundary for a decision stump that splits on GDP ratio.

(b) Decision boundary for a decision stump that splits on the number of voters.

Multi-layer perceptron I use cross validation to tune the hyperparameter λ, which penalizes the magnitude of the weights w_i. The parameter λ is in the range $\{0.1, 0.5, 1.0, 2.0, 5, 10\}$. The best regularization parameter λ equals 0.4. I used the sigmoid activation function but other functions ($ReLu$, hyperbolic tangent) gave approximatively the same MSE. The MSE (training) is MSE $MSE_N^T = 0.0081$ and the test data one is $MSE_N^T = 0.0084$.

Discussion The three algorithms have a MSE on the testing data that are very similar, around 0.09. This may be due to the fact that the variables don't have a sufficient explanatory power. Figure 4a, 4b, and 4c display the actual outcomes (shares for F.Mitterrand) versus the predicted outcomes by the three models on the test data. The prediction is noisy: for actual shares of 0.2, the NN algorithm predicted values range from 0.22 to 0.3. A same commentary can be made about the Ridge and Adaboost models.



(a) Ridge predictions.

(b) Adaboost predictions.

(c) Neural network predictions.

6 Conclusion / Future Work

Are socio-demographic variables good predictors of electoral outcomes? I tried to answer this question using electoral data from the 1981 French presidential election. The goal is to predict the vote shares of the leading candidate, F. Mitterrand, using Ridge regression, Adaboost, and Multi-layer perceptron. The most important features changed by model, but the wealth of a municipality was always predictive of vote shares. This being said, the three models don't differ much one from another and are delivering noisy predictions. This suggests that the available input variables are not explanatory enough.

Future work would get access to richer covariates data and other models (for instance, using beta regression instead of linear regression as the outcome is between 0 and 1). Another route would be to consider other determinants of voting, such as economic downturns (voters often punish the incumbent during economic shocks) or political campaigns.

7 Contributions

Since I am doing the probject by myself, I am the only contributor.

References

H. Ali, H. Farman, and H. Yar. 2022. Deep learning-based election results prediction using twitter activity. *Soft Computing*, page 7535–7543.

Kellyton Brito and Paulo Jorge Leitão Adeodato. 2023. Machine learning for predicting elections in latin america based on social media engagement and polls. *Government Information Quarterly*, 40(1):101782.

Julia Cagé and Thomas Piketty. 2023. *Une histoire du conflit politique. Elections et inégalités sociales en France, 1789-2022*. Le Seuil.

Skyler J Cranmer. 2019. Introduction to the virtual issue: Machine learning in political science. *Political Analysis*.

Paul F. Lazarsfeld, Bernard Berelson, and Hazel Gaudet. 1968. *The People'S Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Columbia University Press.

Pankaj Sinha, Aniket Verma, Purav Shah, Jahnavi Singh, and Utkarsh Panwar. 2020. Prediction for the 2020 united states presidential election using machine learning algorithm: Lasso regression. *Munich Personal RePEc Archive*.