About the Project

General

By using individual-level data on demographics and Swiss voting in national elections, we train a random forest classifier that allows us to make predictions of the likelihood of a person to vote for given parties, depending on specific characteristics. As such, our interactive app informs both policymakers and political players about the specific demographic profile of voters.

Data (including Data Cleaning)

For this project, we use the cumulative dataset from the Swiss Election Study (SELECTS). It contains individual-level election data between the years 1971 and 2019. To predict our model, we include all available years. Additionally, our predictions focus on the 7 most important parties which are: CVP, EVP, FDP, GLP, GPS, SP and SVP. Missing observations of numerical variables are addressed by setting them to zero. An approach frequently applied to handle the absence of data. Our final data set comprises 23,843 observations and incorporates the following demographic variables: sex, age, education, income, religion, linguistic region, urban/rural, work situation, social sectoral occupation (classification by Kriesi), political interest, postmaterialist views, voter's participation in national elections and party attachment.

Model

For our classifier model, we compared the performance between a RandomForest Classifier model and a GradientBoosting model. In both cases, we reverted to various preprocessing and hyperparameter-tuning steps in order to increase not only the performance of the models but also their reliability. After several iterations, we opted for a GradientBoosting model. Precisely, in the first data preprocessing step, we generated polynomial features of degree 2 to capture non-linear relations between features using *PolynomialFeatures.* Then, to handle class imbalances, it was decided to apply *SMOTEENN*, which is a combination of *SMOTE* (Synthetic Minority Over-Sampling Technique) and *ENN* (Edited Nearest Neighbors), and then undersample the majority classes. The limitations of using such techniques are discussed below. To select the best hyperparameters we defined a hyperparameter grid and used *RandomizedSearchCV* with *StratefiedKFold* cross-validation. The model was then initialized and trained on the latter hyperparameters.

Model Performance

Our Gradient Boosting Classifier model, optimized through hyperparameter tuning, demonstrates moderate predictive power in forecasting individual's party support in the Swiss context. With an overall accuracy of 55% and a mean cross-validated F1 weighted score of 0.55, the model shows consistent but somewhat limited performance across different subsets of the data. It performs relatively well in predicting support for parties like

SP and FDP, with F1-scores of 0.66 and 0.59 respectively. The model also shows decent performance for CVP (F1-score 0.62), GLP (0.59), and EVP (0.53). However, it struggles more with predicting support for SVP (0.46) and GPS (0.39), suggesting that the model may have difficulty capturing certain political nuances or demographic patterns associated with these parties. The variation in performance across parties indicates that the model's predictions should be interpreted with caution, especially for parties with lower F1-scores. The precision and recall values also vary considerably between parties, ranging from 0.41 to 0.67 for precision and 0.37 to 0.72 for recall. The cross-validation scores (ranging from 0.53 to 0.57) suggest that the model's performance is relatively stable across different subsets of the data, but indicates that there is room for improvement. Limitations of this model include potential overfitting to the training data, as evidenced by the high max_depth of 20, and the possibility of underrepresenting or misclassifying support for parties with lower predictive scores, particularly GPS and SVP. The model's struggles with these parties might indicate a need for more features or different modeling approaches to capture their voter bases more accurately.
The best parameters found through hyperparameter tuning (learning_rate: 0.01, max_depth: 20, max_features: 'log2', min_samples_leaf: 3, min_samples_split: 18, n_estimators: 198) suggest a relatively complex model, which might be necessary given the intricacies of political preferences but could also contribute to overfitting.

Limitations

Finally, it should be noted that our analysis also contains some limitations. First, as mentionned, our Gradient Boosting model, while it offers valuable insights into Swiss voting patterns, shows variable performance across parties and potential overfitting. Second, demographics do not necessarily capture all the complex aspects influencing someone’s voting preferences. Context-specific factors, whether at the national or individual level such as an economic crisis or changes in personal life, as well as the media influence might play an important role in predicting voting behavior. Third, while back in the 1970s to 2000s religious affiliation, for instance, determined party vote largely (Lijphart, 1979), we see less such connections nowadays, also observable in the loss of power of religious parties like EVP or CVP. Thus, since our model uses data ranging back to the 1970s to predict voting behavior, predictions might not reflect current voting. Future predictions could aim to use only the most current data to provide forecasts that more accurately reflect present voting trends. Lastly, the most important limitations of using SMOTEENN are its potential to cause overfitting by generating overly similar synthetic samples, the introduction of noise in overlapping regions, and its sensitivity to parameter choices, which can make optimization challenging.

References

Lijphart, A. (1979). Religious vs. Linguistic vs. Class voting: The “crucial experiment” of comparing Belgium, Canada, South Africa, and Switzerland. American Political Science Review, 73(2), 442–458. https://doi.org/10.2307/1954890