A Machine Learning Approach: Predicting Voter Turnout in the 2019 European Parliament Election

Amin Oueslati
a.oueslati@students.hertie-school.org

Oskar Krafft
b.krafft@students.hertie-school.org

Benedikt Korbach
b.korbach@students.hertie-school.org

Abstract

In our group project[1], we predict voting propensity in European elections from the Eurobarometer 91.5 European Parliament Post-Election Survey 2019. We find that our baseline Logistic Regression model, the SVM and the Random Forest perform almost equally with an F1 score of around 0.8, while the Naïve Bayes model predicts considerably worse (F1 score 0.6). The ambition of our final report is to additionally interpret our model substantively, examining both feature importance and their relationship to voting propensity.

1. Proposed Method

Our project pursues two goals. First, accurately predict subjects' propensity to vote from the other survey questions included in the Eurobarometer 91.5 European Parliament Post-Election Survey 2019 (EPPES)[2]. Two, interpret our model substantively by producing robust estimates for both feature importance and its relationship to voting propensity.

Our ambition for the midterm report, in line with the feedback we received on our proposal, was to make significant progress on our first goal. With this regard, we have successfully completed the following milestones: (1) cleaning the data, (2) exploring the data, (3) setting up the preprocessing pipeline, (4) tuning the hyperparameters across all models, (5) evaluating the results.

A challenge when predicting voter turnout from the vast number of questions collected in the EPPES relates to the relative sparsity of the data. After pre-processing, our dataset contains approximately 900 variables and 27,000 observations. Additionally, many questions related to respondents' opinions display a concentration of responses around few categories. Essentially, this pattern indicates a commonly observed response bias where participants avoid the extreme ends of the scale.

Considering these challenges, we expect a non-parametric random forest to outperform any parametric model like a logistic regression when confronted with our classification problem. Given the different trade-offs that come with non-parametric models, like the greater computational demands, we nevertheless pursue a modelling approach that follows the cascading complexity often observed in machine learning projects.

In total, we test four models. (1) As our baseline, we use a logistic regression model. (2) Subsequently, we employ a Naïve Bayes model. While the underlying assumptions, more specifically the independence of features, are almost certainly violated in our case, the same is true for any real-world classification problem. (3) Next, we use a Support Vector Machine (SVM). Generally, our binary classification problem and the still relatively small size of our dataset are well suited for a SVM. However, ex-ante, we do not expect the SVM to outperform Random Forest, as we arguably do not reach the levels of dimensionality and sparsity where SVM excels. Importantly, any minor outperformance would likely not justify the limited interpretability and high computational intensity. (4) Our last model is a Random Forest, which we ex-ante expect to perform the best, given its high degree of flexibility and general robustness.

2. Experiments

Data: In our project, we use the dataset of the Eurobarometer 91.5 European Parliament Post-Election Survey 2019 (EPPES). The Eurobarometer regularly surveys citizens on public opinion about a broad range of topics related to the EU and other political and social issues.

Apart from the core questions of the Standard Eurobarometer, the EPPES also includes questions about the public perception towards the EU, its institutions and, importantly, whether the respondent voted in the last European Parliament Election in May 2019. The survey is particularly

[1]Link to our repository and the data set: https://github.com/OskarKrafft/Machine-Learning-Project.git

suited to our goal of predicting voting propensity since the data was collected shortly after the election (07.06.2019 - 01.07.2019), working against issues like selective memory or winning bias, which often confound self-reported voting measures. While reduced, social desirability biases leading to inaccurate self-reporting will remain a confounder of our estimates. Additionally, statistical noise caused by random factors, for instance, weather, diminishes our predictive accuracy.

The poll was conducted among citizens aged 15 years and above in all EU member states and the five candidate countries at that time (Turkey, North Macedonia, Montenegro, Serbia and Albania). Around 1,000 interviews were conducted in every EU member state. However, only about 500 interviews were conducted in the Republic of Cyprus, Luxembourg, and Malta. Further, approximately 1,500 interviews were conducted in Germany. Survey questions cover a broad range of issues, such as attitudes towards the EU, perceptions of crucial policy issues, socioeconomic attributes, information on media consumption and respondents' value system. To achieve a balanced and accurate sample, the pollster applied a multi-stage, random sample design in which all sample points for a country were drawn with a probability proportional to population size and density.

We downloaded the data set from the GESIS Data Archive of the Leibniz Institute for the Social Sciences on 04.10.2022. The raw dataset contains the answers of 32,524 respondents across 872 variables.

Within the data cleaning stage, we exclude all observations from non-EU countries and territories as we are only interested in survey respondents who were able to vote in the last European Parliament elections. Further, we drop all variables referring to survey questions that were not answered by all respondents. These are mostly nested questions or ones that only apply to a substratum of respondents. We also discard some questions that we deem to have limited informative character, are closely related to other survey items or solely serve statistical purposes to decrease the computational complexity of our models. Lastly, all categorical variables are one-hot encoded, resulting in a cleaned data set containing 27,464 observations across 939 variables.

The conditional variable in our ML models refers to question QG1 of the EPPES: "[...] Did you yourself vote in the recent European Parliament elections?". Of all survey takers in our cleaned data set, 15,008 (54.6%) indicate they voted in the last parliamentary election, while 12,456 (45.2%) report not having voted (see Figure 1).

During the data exploration phase, we additionally examine the distributions of some informative socioeconomic variables such as age, gender and class affiliation (see Figure 2). The age of respondents ranges from 15 to 98 years,

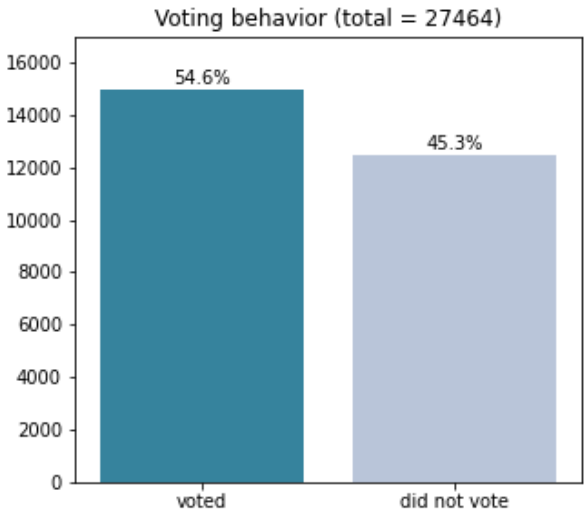


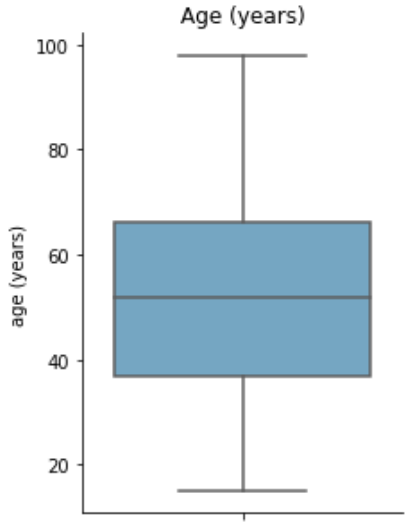
Figure 1: Voting Behavior

with a mean age of 51 years. Further, 54.2% of survey takers in our sample identify as women and 45.8% as men. Lastly, we observe that most respondents affiliate themselves with the middle class (47.3%), followed by the working class (26.4%) and the lower middle class (14.8%).

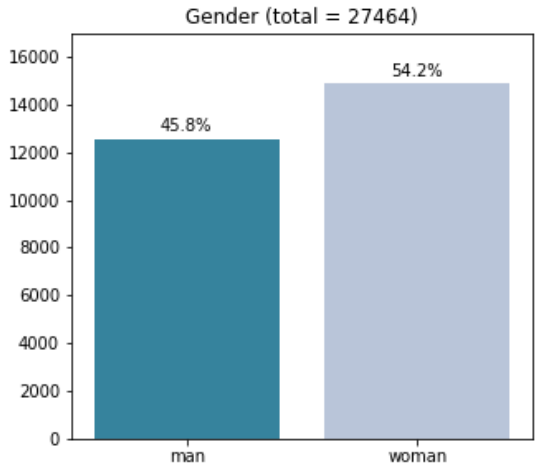
As the final step in the data exploration stage, we analyse how descriptive variables commonly used in the voting propensity literature are correlated with voter turnout in our data. For this purpose, we calculate and plot voting behaviour conditional on respondents' age, gender and class affiliation (see Figure 3). Additionally, we use survey question D71a2, which asks how frequently a survey taker discusses European political matters with friends or relatives as a proxy for interest in European politics (Question D71a2: "When you get together with friends or relatives, would you say you discuss frequently, occasionally or never about...?").

Our observations are in line with the findings in the relevant literature: Voter turnout generally seems to increase with age, membership in a higher social class, and interest in politics. Additionally, our descriptive analysis suggests that men are more likely to vote than women.

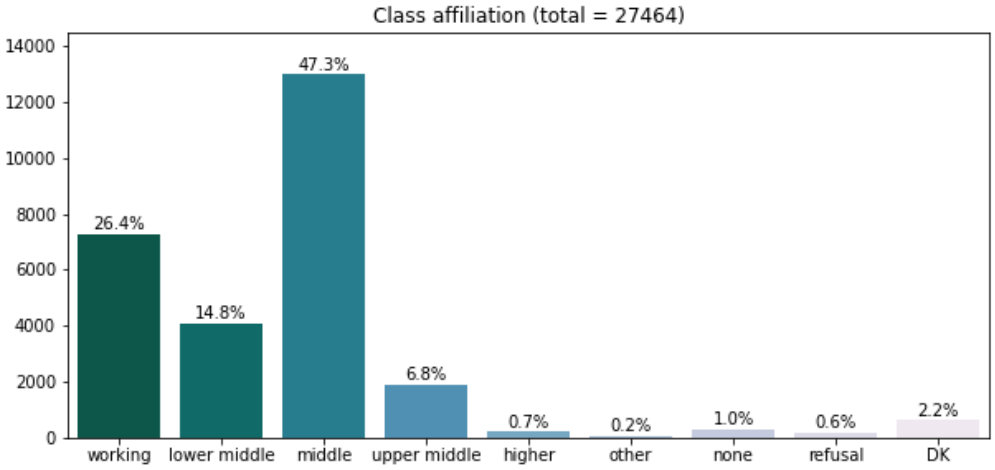
Evaluation method: Our evaluation strategy consists of four steps: (1) We use F1, the harmonic mean of precision and recall, as the key metric for tuning hyperparameters. We chose F1 given some imbalance in our data, in which case F1 is superior to other measures like accuracy. (2) Having identified the best-performing specifications for all models, we test the robustness of their relative performance on the training data through repeated cross-validation. Here, we



(a) Age



(b) Gender



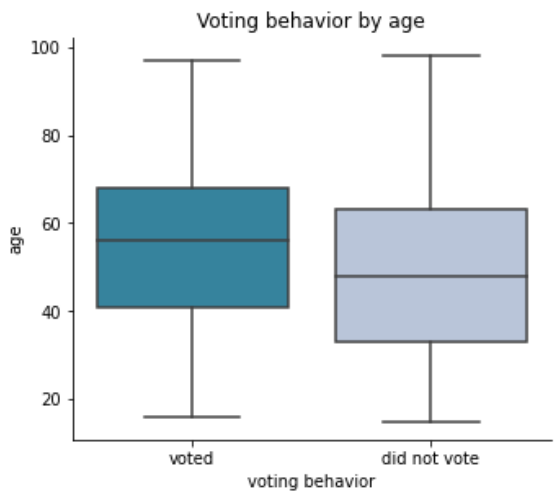
(c) Class affiliation

Figure 2: Descriptives

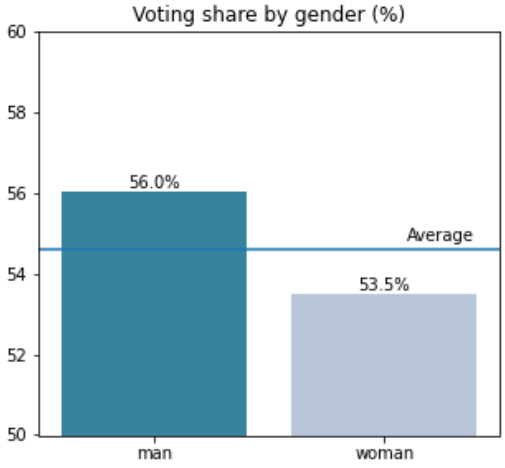
also rely on the F1 score. (3) We examine the learning curves of all hyperparameter-tuned models to explore any issues of over- or underfitting. (4) Lastly, we put the models to their final test by predicting voting behaviour from the test data. We compute all major evaluation metrics from the confusion matrix.

Experimental details: Prior to any evaluation steps, the clean data is pre-processed before entering the models. (1) Training and test data are split 80/20. (2) Many features are measured on an ordinal scale. Importantly, these variables usually contain one or more "don't know" options, coded as the highest scale value. This is problematic since treating the feature as ordinal would turn "don't know" into the most extreme level of (dis)agreement for a given question, which is substantively nonsensical. Relevant literature offers two solutions to this problem. Either drop all rows which contain a "don't know" response or treat the variable as categorical. While the former shrinks the dataset and introduces bias, the latter eliminates the information on the variable's ordinal nature. After exploring both options, we decided to treat all ordinal variables as categorical since dropping rows with "don't know" would have reduced our data by more than 50%. Our pre-processing pipeline thus one hot encodes all ordinal variables.

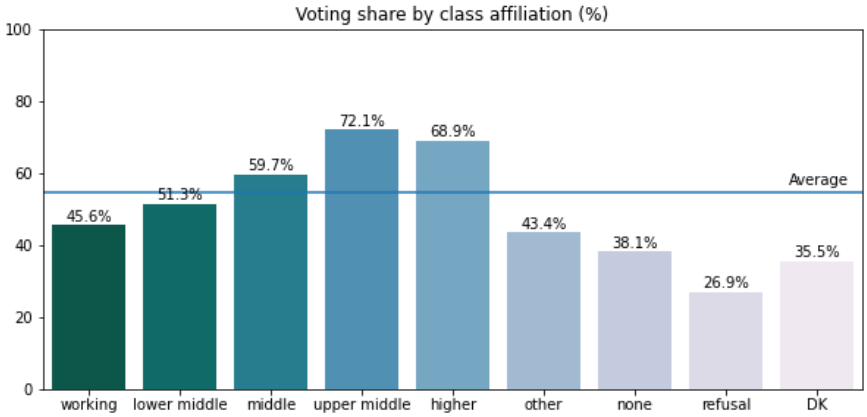
All model computations on training data are performed



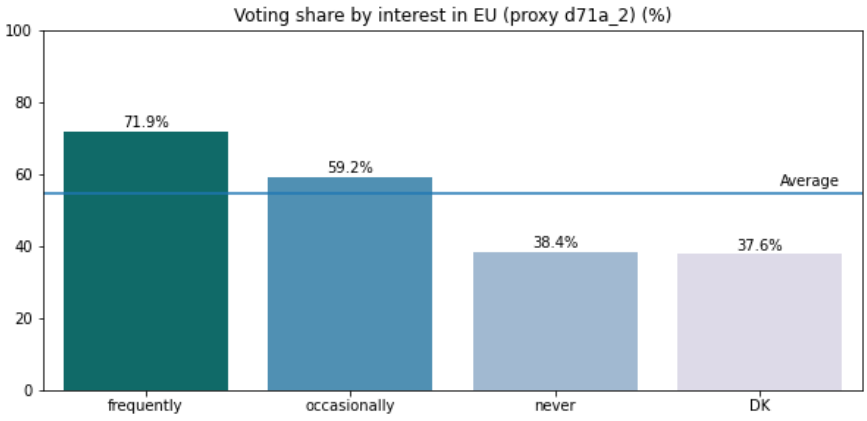
(a) Age



(b) Gender



(c) Class affiliation



(d) Interest in EU politics

Figure 3: Voting share by age, gender, class affiliation and interest in EU politics

with k-fold, stratified, repeated cross-validation. Whenever feasible, the data is split into 10 folds at 3 repetitions. We reduce the number of folds for SVM and Random Forest to manage computation time. By using stratified cross-validation, we account for the imbalances in our data.

The hyperparameters of the Logistic Regression model and SVM are tuned with grid search. In contrast, we rely on random search to explore a larger number of parameters for the Random Forest. Table 1 shows the best-performing parameters per model type. Overall, we have seen limited sensitivity across models to parameter variation, especially for the Random Forest. Naïve Bayes is implemented with a Bernoulli classifier, given that all variables apart from age are binary. To meet the classifier's assumptions, the variable age is divided into five bins and subsequently one-hot encoded.

To test their robustness and compare their performance, we evaluate our hyperparameter-tuned models on the training data in repeated, stratified cross-validation. As depicted in Figure 5, the F1 scores only exhibit small variance, which indicates robust model performance.

As Figure 4 displays, the learning curve for Logistic Regression shows a consistent yet minor convergence rate of training and validation F1 scores. Thus, the model would probably benefit from more training data to improve generalization. Additionally, the standard deviation of the validation score curve is slightly higher compared to the training curve. Hence, the Logistic Regression baseline suffers from a small error due to variance, therefore it is possible that the model is overfitting.

Contrarily, the Naïve Bayes learning curve shows that the training and validation curves have converged and subsequently, the model would not improve by adding more data. Furthermore, the SVM training and validation curves converge and stabilise beyond more than 8000 observations. Nevertheless, the standard deviation of the validation curve is is considerably larger than the training curve, pointing towards potential overfitting. Lastly, the Random Forest learning curve highlights that the model would not improve by adding more data as the two curves do not converge. Despite the relatively large gap between the two curves, the consistency of the validation and training curves potentially indicates a good fit.

Results: To evaluate the performance of our models, we fit our best-performing models on the unseen test data and report F1, MCC, accuracy, precision and recall in Table 2. Measured by F1, our key performance metric, Random Forest performed the best with an F1 score of 0.806, followed by Logistic Regression (0.803) and Support Vector Machine (0.796). As was the case for training data results (see Figure 5), the best three models show a near-equal capacity to predict voting behavior. Equally in line with the training results, the performance of the Naïve Bayes model was considerably worse, with an F1 score of 0.638.

Comment on your quantitative results: Contrary to our ex-ante expectation, the Random Forest has only weakly outperformed Logistic Regression. A possible explanation is offered by a large-scale benchmarking experiment from 2018[3]. The authors find that random forest outperforms Logistic Regression in 69.0% of all binary classification problems. Their analysis reveals further that if the number of features relative to observations were to increase in our model, this would likely amplify the outperformance of the Random Forest.

Similarly, we were surprised by the poor performance of Naïve Bayes, which clearly underperforms relative to the other models. This might suggest that the independence of features, the assumption underlying Naïve Bayes, is severely violated in our case.

While our models produce more accurate predictions for voter turnout than classical models in the voting propensity literature, they do fall behind the predictions from a US paper, representing the only comparable study to the best of our knowledge[1]. More specifically, the paper estimates voting propensity from the 2016 US Census with 374 questions and 75,000 respondents, achieving approximately 90% accuracy with both Random Forest and Logistic Regression. Considering the plateau in performance observed in the learning curves, for instance for Random Forest at an F1 score of around 0.80, the difference in sample size likely plays a limited role. However, substantively, the data in the Census might be more informative to voting decisions.

3. Future work

We plan to compute LIME and SHAP for our Random Forest model to tackle our second project goal relating to our model's global interpretability. Both methods are superior to alternative approaches like Permutation Feature Importance. Firstly, they establish the nature of the relationship between predictors and outcome variable, i.e., whether it is positively or negatively affecting voting propensity. Secondly, they are unaffected by multicollinearity between predictors, which we expect to observe [4]. SHAP will be our primary measure, given that it is generally more stable than LIME, which tends to vary greatly for different Kernel widths. However, we intend to compare SHAP and LIME, if we should encounter discrepancies between the two.

References

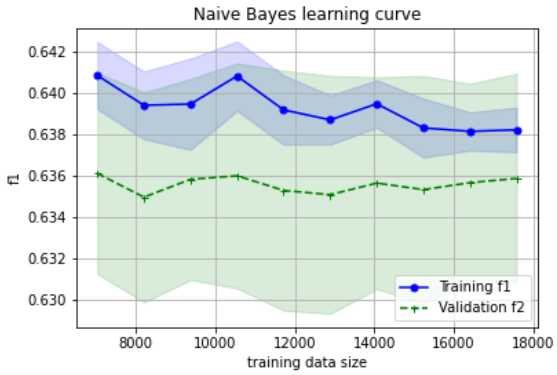
[1] T. Challenor. Predicting votes from census data. *Stanford University final reports*, Dec. 2017.

[2] E. Commission and B. European Parliament. Eurobarometer 91.5 (2019). GESIS Data Archive, Cologne. ZA7576 Data file Version 1.0.0, https://doi.org/10.4232/1.13393, 2019.

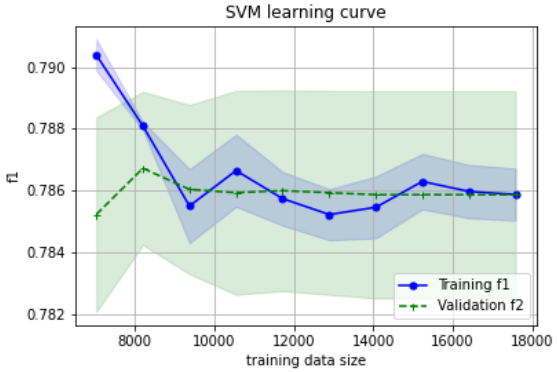
Model	Estimator	Hyperparameter
Logistic Regression	Logistic Regression	solver: L-BFGS, penalty: L2, C: 0.01
Naïve Bayes	Bernoulli	No hyperparameter tuning
Support Vector Machine	SVC	kernel: rbf, gamma: 0.01, C: 1.0
Random Forest	Random Forest Classifier	n_estimators: 944, min_samples_split: 2, min_samples_leaf: 1, max_features: auto, max_depth: 90

Table 1: Hyperparameter Tuning Overview

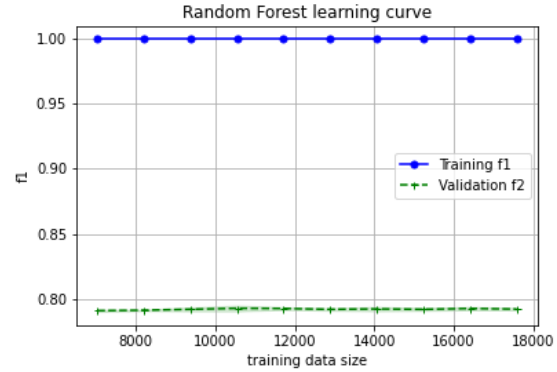
(a) Logistic Regression



(b) Naïve Bayes



(c) Support Vector Machine



(d) Random Forest

Figure 4: Learning Curves

Model	F1	MCC	Accuracy	Precision	Recall
Logistic Regression	0.803	0.645	0.824	0.826	0.781
Naïve Bayes	0.638	0.351	0.679	0.661	0.617
Support Vector Machine	0.796	0.640	0.821	0.836	0.759
Random Forest	0.806	0.650	0.827	0.825	0.788

Table 2: Classification Metrics

[3] R. Couronné, P. Probst, and A.-L. Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19(1), July 2018.

[4] C. Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

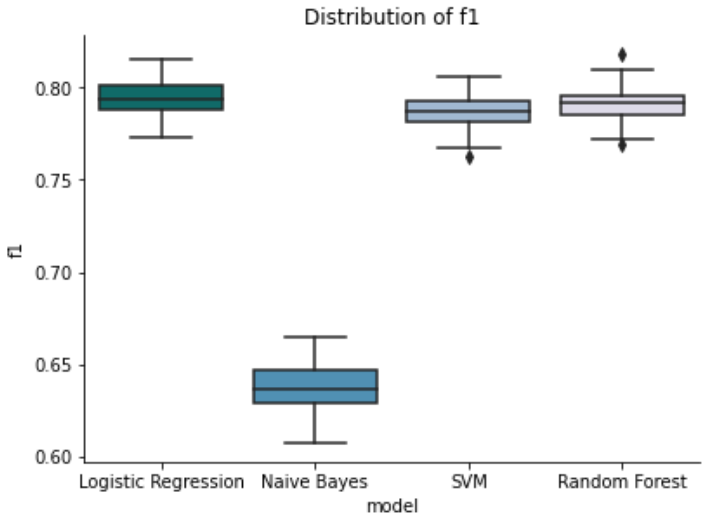


Figure 5: f1 scores per model (cv = 30)