Predicting voter turnout: A comparative analysis of theory-driven and data-driven models

Grotto Tommaso

1. Introduction

Voter turnout is a cornerstone of democratic societies, as the legitimacy of governmental institutions relies heavily on popular consensus and participation. However, recent years have witnessed a decline in electoral participation during national elections across various European countries[1]. This trend has sparked significant concern among researchers and policymakers, prompting extensive discussions aimed at uncovering the factors contributing to low or high voter turnout.

This study seeks to investigate the determinants of voting behavior on the individual level by leveraging a combination of theory-driven and data-driven methodologies. We will analyze how demographic variables, social factors, political engagement, trust in institutions, economic conditions, religious belonging, and perceptions of discrimination impact the likelihood of an individual voting.

To achieve this, we will employ logistic regression, random forest, and gradient boosting models to identify and compare the most significant predictors of voting behavior. By examining the performance and accuracy of these models, we aim to determine which method provides the most reliable insights into voter turnout.

For the theory-driven approach, we will use logistic regression informed by existing literature to identify the most influential predictors. For the data-driven approach, we will apply logistic regression, random forest, and gradient boosting to a set of forty-six variables, allowing the models to autonomously determine the most impactful factors. By integrating both theory-driven and data-driven perspectives, we aim to provide a more nuanced and accurate analysis of the factors influencing voter turnout.

2. Literature Review and Research Question

Previous studies have identified a variety of factors that influence voting behavior at both macro and micro levels. At the micro level, demographic factors such as age, education, and income have been shown to significantly predict the likelihood of a person voting (Blais 2006; Matsusaka & Palda 1999). Specifically, older people tend to vote more frequently (Blais 2006; Smets & Van Ham 2013), and higher levels of education and income are associated with higher voter turnout (Cancela & Geys 2016; Franklin 1999; Lijphart 1997). Gender was considered influential, with men participating more in elections than women, but

---

[1] https://www.idea.int/data-tools/data/question-region?question_id=9188&continent=Europe&database_theme=293

this gap has been closing over time, to the point that some papers affirm that gender is not statistically significant anymore (Blais 2006; Smets & Van Ham 2013).

At the micro level, there are also some political aspects that the researchers indicate as important. If a person identifies with a political party the likelihood of voting increases, higher levels of interest in politics are associated with higher turnout, and the more knowledgeable the individuals are about politics the more they are likely to vote (Blais & Dobrzynska 1998; Smets & Van Ham 2013).

At the macro level, the type of electoral system is indicated as a significant factor, proportional representation systems tend to result in higher voter turnout compared to majoritarian systems due to their inclusive nature and the perception that every vote counts (Franklin 1999; Kostadinova 2003; Lijphart 1997). Countries that implement compulsory voting laws typically see higher turnout rates (Blais 2006; Franklin 1999; Geys 2006). The ease and accessibility of voter registration processes influence turnout, automatic registration methods are associated with higher participation rates compared to complex and restrictive registration requirements (Cancela & Geys 2016; Lijphart 1997; Powell 1986). Higher campaign spendings are associated with increased voter turnout (Cancela & Geys 2016; Franklin 1999).

This study focuses on the micro level, for this reason, the analysis that is being conducted uses individual-level variables to determine their impact on voter turnout.

The research question of this study is: which are the most influential variables for predicting voter turnout in national elections in European countries at an individual level? Additionally, which is the best analytical technique to use for predicting voter turnout in national elections in European countries?

To answer this question, we will employ a mixed-methods approach, utilizing both theory-driven and data-driven analytical strategies.

3. Data Description

The data used in this study comes from the European Social Survey (ESS), a cross-national survey conducted across Europe every two years. The sample comprises 7368 observations collected in 2020 from 18 European countries: Belgium, Bulgaria, Switzerland, Finland, Great Britain, Greece, Croatia, Hungary, Ireland, Iceland, Italy, Lithuania, North Macedonia, Netherlands, Norway, Portugal, Slovenia, and Slovakia.

The dependent variable in this study is whether an individual voted in the last national election or not. The independent variables, categorized by type, are:

Demography:
age (agea), gender (gndr), education level (eisced), country of residence (cntry), marital status (marsts), presence of children in the household (chldhhe), self-reported health status (health).

Trust in the institutions:
trust in parliament (trstprl), in the legal system (trstlgl), in the police (trstplc), in the politicians (trstplt), in political parties (trstprt), in the European Parliament (trstep), in the United Nations (trstun), in the scientists (trstsci).

Politics:
interest in politics (polintr), closeness to a party (prtdgcl), satisfaction with the way democracy works in the country (stfdem), importance of living in a democratically governed country (implvdm), left-right political alignment (lrscale), time spent reading, watching, or listening to news about politics (nwspol).

Social Factors:
agreement that gays and lesbians are free to live life as they wish (freehms), level of shame if a close family member is gay or lesbian (hmsfmlsh), agreement that gay and lesbian couples should have the right to adopt children (hmsacld), frequency of social meetings with friends, relatives, or colleagues (sclmeet), Overall life satisfaction (stflife), perception of safety in the neighborhood after dark (aesfdrk), experience of crime victimization (crmvct), frequency of internet use (netusoft).

Economic Factors:
satisfaction with the present state of the economy in the country (stfeco), agreement that the government should reduce differences in income levels (gincdif), household total net income (hinctnta), employment status (emplrel), total weekly hours worked (wkhtot).

Religion:
belonging to a particular religion or denomination (rlgblg), the religion or denomination belonging to at present (rlgdnm), frequency of pray outside of the religious services (pray).

Immigration:
view on allowing many or few immigrants of different races/ethnic groups (imdfetn), agreement that immigration is bad or good for the country's economy (imbgeco), agreement that the country's cultural life is undermined or enriched by immigrants (imueclt), agreement that immigrants make the country a worse or better place to live (imwbcnt).

Discrimination:
being a member of a discriminated group (dscrgrp), specifically being discriminated by race (dscrrce), by nationality (dscrntn), by religion (dscrrlg), by language (dscrlng).

4. Methodology

To explore the determinants of voting behavior, we will employ three analytical techniques: logistic regression, random forest, and gradient boosting. Each technique offers unique advantages and can provide complementary insights.

Logistic regression is a widely used statistical method for modeling binary outcomes. The advantages of this technique are that it provides interpretable coefficients that indicate the direction and strength of associations and that it is suitable for understanding the influence of individual predictors. However, this method assumes linear relationships between predictors and the log odds of the outcome, and it may not capture complex interactions between variables.

Random forest is an ensemble learning method that constructs multiple decision trees and combines their predictions. The advantages of this technique are that it can handle complex interactions and non-linear relationships and that it provides measures of variable importance. However, it is less interpretable than logistic regression and computationally intensive, especially with large datasets.

Gradient boosting is another ensemble learning technique that builds models sequentially, with each new model attempting to correct the errors of the previous one. The advantages of this technique include its high effectiveness for predictive modeling and its ability to handle large numbers of predictors and complex relationships. However, it is computationally demanding and interpreting its results can be challenging.

By employing these methods, we aim to leverage their respective strengths and mitigate their weaknesses to provide a comprehensive analysis of the factors influencing voting behavior.

The models created for this study were trained on 70% of the 7368 observations in the sample, with the remaining 30% used to test the accuracy of the models.

In our dataset, the variable of interest, "vote," shows a significant disparity between its categories, favoring those who voted in the last national election. Due to this, traditional machine learning models may skew towards the majority class, failing to accurately predict or capture the characteristics of the minority class. To address the issue, we employ the ROSE (Random Over-Sampling Examples) technique, which mitigates this problem by generating synthetic samples for the minority class, using the actual data as the base. It combines both over-sampling of the minority class and under-sampling of the majority class to create a more balanced dataset.

This approach ensures the model is trained on a dataset with equal representation of both classes, enhancing the model's robustness and generalizability. In our case, the training set balance improved from a 4759 - 399 ratio to a 2547 - 2611 ratio between the two classes. By applying ROSE to the training dataset, we aim to enhance the model's ability to detect and accurately classify the minority class, which is crucial for our analysis of voting behavior. This adjustment ensures that the insights derived from our models are reliable and not biased by class imbalance.

5. Results

We begin our analysis by displaying the result of the theory-driven logistic regression model. For this model, we employed the variables that are considered the most influential in the literature for predicting voting behavior. The independent variables are age, gender, education, income, interest in politics, closeness to a party, time spent on political news, and country of residence.

The vote variable is coded as 0 for people who voted in the last national elections and 1 for people who did not vote. Thus, a positive coefficient indicates that a high value of the predictor increases the likelihood of not voting.

Table 1: Theory-driven logistic regression model between the vote variable and a series of predictors.

Predictor	Coeff.	Std. Error	P-value
(Intercept)	-0.4924795	0.2151284	0.022066 *
agea	-0.0008473	0.0004420	0.055243 .
gndr2	-0.0623555	0.0613319	0.309301
eisced2	0.1452807	0.1460953	0.320016
eisced3	-0.2894951	0.1561374	0.063724 .
eisced4	-0.2659942	0.1305615	0.041619 *
eisced5	-0.5959542	0.1549658	0.000120 ***
eisced6	-0.7281578	0.1437672	4.09e-07 ***
eisced7	-0.6932495	0.1466864	2.29e-06 ***
hinctnta	-0.0381729	0.0095450	6.35e-05 ***
nwspol	0.0004458	0.0001522	0.003412 **
polintr2	0.3448171	0.0992857	0.000515 ***
polintr3	0.816599	0.1067950	2.07e-14 ***
polintr4	1.1440480	0.1354560	< 2e-16 ***
prtdgcl2	0.4575202	0.0980388	3.06e-06 ***
prtdgcl3	0.7987101	0.1186868	1.70e-11 ***
prtdgcl4	1.6964940	0.2186992	8.68e-15 ***
cntryBG	-0.0526544	0.1626479	0.746140
cntryCH	0.9964578	0.1760575	1.52e-08 ***
cntryFI	0.3060867	0.1554297	0.048919 *
cntryGB	0.1165775	0.1810310	0.519599
cntryGR	-0.6362699	0.2215627	0.004082 **
cntryHR	0.5391301	0.1805960	0.002833 **
cntryHU	-0.3266037	0.1908263	0.086984 .
cntryIE	-0.3213536	0.2107728	0.127348
cntryIS	-0.6247072	0.2389318	0.008934 **
cntryIT	-0.6046000	0.1938218	0.001812 **
cntryLT	1.0203811	0.1868949	4.77e-08 ***
cntryMK	-0.1534673	0.2329859	0.510090
cntryNL	-0.4678692	0.1850496	0.011460 *
cntryNO	-0.1480856	0.1736649	0.393821
cntryPT	0.0633618	0.1867594	0.734407
cntrySI	0.0784715	0.2068293	0.704389
cntrySK	-0.4768699	0.2399494	0.046881 *

Significance codes: 0 < *** < 0.001 < ** < 0.01 < * < 0.05 < . < 0.1 < < 1

The results indicate that gender is not statistically significant, and age is not highly significant. The education is statistically significant, with individuals having higher levels of education tending to vote more. Similarly, people with higher income levels exhibit higher voter turnout. Interestingly, more time spent on political news is associated with a lower likelihood of voting. Both interest in politics and closeness to a party are associated with a higher probability of voting. Additionally, the country of residence is statistically significant for predicting voter behavior, Belgium is taken as reference and there are significant differences between it and some other countries like Switzerland, which has a higher probability of not voting, and Greece, which has a higher probability to vote.

These results align with the literature, except for age, which is not very significant, and political knowledge, proxied by the amount of time spent on political news, which is negatively associated with voting in elections. This may be because this variable is not a good proxy for political knowledge.

Table 2: Confusion matrix between predicted results and actual category of the test set for the theory-driven logistic regression model.

		Reference	
		Voted	Not Voted
Prediction	Voted	1379	59
	Not Voted	660	112

The accuracy of the logistic regression model on the test set is 0.67, a decent result. The sensitivity is 0.68 and the specificity is 0.65, both consistent with the overall accuracy.

We continue by presenting the results of the data-driven logistic regression model, utilizing 46 variables instead of the 8 variables used previously.

Table 3: Data-driven logistic regression model between the vote variable and a series of predictors.

Predictor	Coeff.	Std. Error	P-value	
(Intercept)	1.054e+00	5.944e-01	0.076334	.
agea	-7.226e-05	5.079e-04	0.886868	
gndr2	1.118e-01	7.836e-02	0.153541	
eisced2	4.460e-02	1.694e-01	0.792352	
eisced3	-3.850e-01	1.853e-01	0.037789	*
eisced4	-4.912e-01	1.614e-01	0.002335	**
eisced5	-1.030e+00	1.881e-01	4.35e-08	***
eisced6	-9.908e-01	1.776e-01	2.42e-08	***
eisced7	-9.493e-01	1.790e-01	1.14e-07	***
cntryBG	-4.190e-01	2.586e-01	0.105130	
cntryCH	1.716e+00	2.148e-01	1.34e-15	***
cntryFI	7.390e-01	1.949e-01	0.000150	***

cntryGB	3.114e-01	2.165e-01	0.150355	
cntryGR	-1.128e+00	3.346e-01	0.000751	***
cntryHR	1.245e+00	2.220e-01	2.03e-08	***
cntryHU	-4.405e-01	2.412e-01	0.067801	.
cntryIE	1.029e-01	2.473e-01	0.677257	
cntryIS	-6.438e-01	2.777e-01	0.020435	*
cntryIT	-2.831e-01	2.409e-01	0.239869	
cntryLT	1.727e+00	2.354e-01	2.20e-13	***
cntryMK	-7.202e-02	3.337e-01	0.829103	
cntryNL	-4.646e-01	2.152e-01	0.030860	*
cntryNO	1.360e-01	2.091e-01	0.515235	
cntryPT	6.612e-01	2.218e-01	0.002875	**
cntrySI	4.652e-01	2.452e-01	0.057774	.
cntrySK	-1.524e-01	2.828e-01	0.589994	
marsts2	-1.271e+01	4.526e+02	0.977600	
marsts3	2.786e-02	4.232e-01	0.947509	
marsts4	1.008e+00	2.731e-01	0.000222	***
marsts5	7.842e-01	2.834e-01	0.005654	**
marsts6	1.544e+00	2.697e-01	1.04e-08	***
marsts66	4.416e-01	2.607e-01	0.090318	.
chldhhe2	2.551e-01	1.093e-01	0.019630	*
chldhhe6	1.442e-01	9.231e-02	0.118330	
health2	2.283e-02	9.313e-02	0.806359	
health3	2.895e-01	1.119e-01	0.009694	**
health4	4.826e-01	1.725e-01	0.005155	**
health5	1.028e+00	3.029e-01	0.000690	***
trstprl	1.065e-02	1.379e-02	0.439993	
trstlgl	7.284e-03	1.280e-02	0.569391	
trstplc	-1.763e-02	1.369e-02	0.197735	
trstplt	-3.670e-02	1.447e-02	0.011210	*
trstprt	-5.334e-02	1.526e-02	0.000474	***
trstep	2.172e-03	1.352e-02	0.872363	
trstun	-1.196e-02	1.365e-02	0.380933	
trstsci	-2.217e-02	1.507e-02	0.141281	
polintr2	2.456e-01	1.138e-01	0.030930	*
polintr3	7.577e-01	1.231e-01	7.57e-10	***
polintr4	9.659e-01	1.590e-01	1.24e-09	***
prtdgcl2	3.759e-01	1.138e-01	0.000959	***
prtdgcl3	7.259e-01	1.372e-01	1.23e-07	***
prtdgcl4	1.307e+00	2.454e-01	1.00e-07	***
stfdem	-4.664e-03	1.343e-02	0.728457	
implvdm	-1.062e-01	1.612e-02	4.43e-11	***
lrscale	2.159e-02	1.144e-02	0.059147	.
nwspol	4.132e-04	1.753e-04	0.018452	*

freehms2	3.885e-02	1.073e-01	0.717397	
freehms3	2.350e-01	1.519e-01	0.121805	
freehms4	5.727e-02	1.798e-01	0.750036	
freehms5	6.701e-01	2.064e-01	0.001167	**
hmsfmlsh2	1.060e-01	1.932e-01	0.583030	
hmsfmlsh3	-2.370e-01	2.008e-01	0.237888	
hmsfmlsh4	-5.258e-01	1.889e-01	0.005379	**
hmsfmlsh5	9.715e-02	1.950e-01	0.618369	
hmsacld2	-1.209e-01	1.144e-01	0.290763	
hmsacld3	4.164e-04	1.362e-01	0.997561	
hmsacld4	4.000e-02	1.476e-01	0.786342	
hmsacld5	-6.759e-01	1.635e-01	3.58e-05	***
sclmeet2	-1.100e+00	2.718e-01	5.14e-05	***
sclmeet3	-1.709e+00	2.811e-01	1.20e-09	***
sclmeet4	-1.590e+00	2.684e-01	3.18e-09	***
sclmeet5	-1.852e+00	2.730e-01	1.16e-11	***
sclmeet6	-1.900e+00	2.721e-01	2.87e-12	***
sclmeet7	-2.385e+00	2.816e-01	< 2e-16	***
stflife	-3.657e-02	1.528e-02	0.016690	*
aesfdrk2	-5.920e-02	8.273e-02	0.474273	
aesfdrk3	1.561e-02	1.261e-01	0.901463	
aesfdrk4	7.069e-01	2.272e-01	0.001862	**
crmvct2	-2.186e-01	1.058e-01	0.038752	*
netusoft2	-6.679e-01	2.145e-01	0.001846	**
netusoft3	4.623e-01	2.006e-01	0.021164	*
netusoft4	-5.956e-02	1.676e-01	0.722356	
netusoft5	2.676e-02	1.301e-01	0.837049	
stfeco	4.345e-02	1.412e-02	0.002098	**
gincdif2	2.656e-02	8.561e-02	0.756395	
gincdif3	1.298e-01	1.180e-01	0.271295	
gincdif4	-1.936e-01	1.521e-01	0.203073	
gincdif5	3.327e-04	2.711e-01	0.999021	
hinctnta	-1.821e-02	1.158e-02	0.115828	
emplrel2	3.085e-01	1.115e-01	0.005643	**
emplrel3	5.221e-01	2.334e-01	0.025331	*
emplrel6	1.931e-01	2.568e-01	0.452104	
wkhtot	-2.444e-04	3.186e-04	0.442976	
rlgblg2	6.303e-01	1.151e-01	4.34e-08	***
rlgdnm2	-8.981e-02	1.479e-01	0.543742	
rlgdnm3	9.939e-01	2.243e-01	9.39e-06	***
rlgdnm4	3.170e-01	5.291e-01	0.549018	
rlgdnm5	-1.326e+01	4.908e+02	0.978451	
rlgdnm6	9.105e-01	2.961e-01	0.002103	**
rlgdnm7	-3.258e-01	6.607e-01	0.621876	

rlgdnm8	5.293e-01	5.263e-01	0.314554
pray2	1.600e-01	1.631e-01	0.326608
pray3	3.274e-02	1.759e-01	0.852398
pray4	7.567e-01	1.596e-01	2.13e-06 ***
pray5	-9.306e-02	1.817e-01	0.608627
pray6	1.307e-01	1.329e-01	0.325380
pray7	1.869e-01	1.307e-01	0.152510
imdfetn2	-4.518e-01	1.059e-01	2.00e-05 ***
imdfetn3	-3.441e-01	1.261e-01	0.006346 **
imdfetn4	-6.998e-01	1.770e-01	7.72e-05 ***
imbgeco	-2.781e-02	1.373e-02	0.042831 *
imueclt	2.251e-02	1.325e-02	0.089306 .
imwbcnt	-2.178e-02	1.492e-02	0.144349
dscrgrp2	6.666e-01	1.591e-01	2.80e-05 ***
dscrrce1	8.920e-01	3.610e-01	0.013469 *
dscrntn1	-6.202e-01	5.363e-01	0.247514
dscrrlg1	1.578e+00	4.237e-01	0.000195 ***
dscrlng1	-1.422e+01	2.327e+02	0.951297

Significance codes: 0 < *** < 0.001 < ** < 0.01 < * < 0.05 < . < 0.1 < < 1

In this model, age and gender are not statistically significant predictors of voting behavior. Higher education levels are associated with a higher likelihood of voting, with significant differences between categories. The country of residence has a significant impact on voting behavior. The marital status has significant differences between its categories, being legally married is taken as reference and being divorced, widowed, or never married increases the probability of not voting. Worse health conditions are associated with a decreased likelihood of voting.

The trust in parties is statistically significant, higher trust is associated with a higher probability of voting. Political interest and closeness to a party are also significant, less interested individuals and those not close to a party are less likely to vote. Individuals who believe that it is important to live in a democratically governed country are more likely to vote. Individuals who strongly oppose the right of gay and lesbian couples to adopt children are statistically more likely to vote than those who strongly support this right. Frequent social interactions with friends, relatives, or colleagues increase the likelihood of voting.

Belonging to a particular religion reduces the probability of voting and being orthodox lowers the probability compared to being from any other religion. The frequency of praying, which can range from every day to never, has no significant differences except for those who pray once a month, who have a lower probability of voting.

People opposed to the immigration of different ethnic groups are more likely to vote than those supportive of immigration. People who indicate themselves as discriminated in the country they live in have a higher probability of voting, the only exception is for those who feel discriminated for their religion, who are less likely to vote.

Table 4: Confusion matrix between predicted results and actual category of the test set for the data-driven logistic regression model.

		Reference	
		Voted	Not Voted
Prediction	Voted	1456	61
	Not Voted	583	110

The accuracy of the data-driven logistic regression model on the test set is 0.71, which is good but not optimal. The sensitivity is 0.71 and the specificity is 0.64, both indicating decent performance.

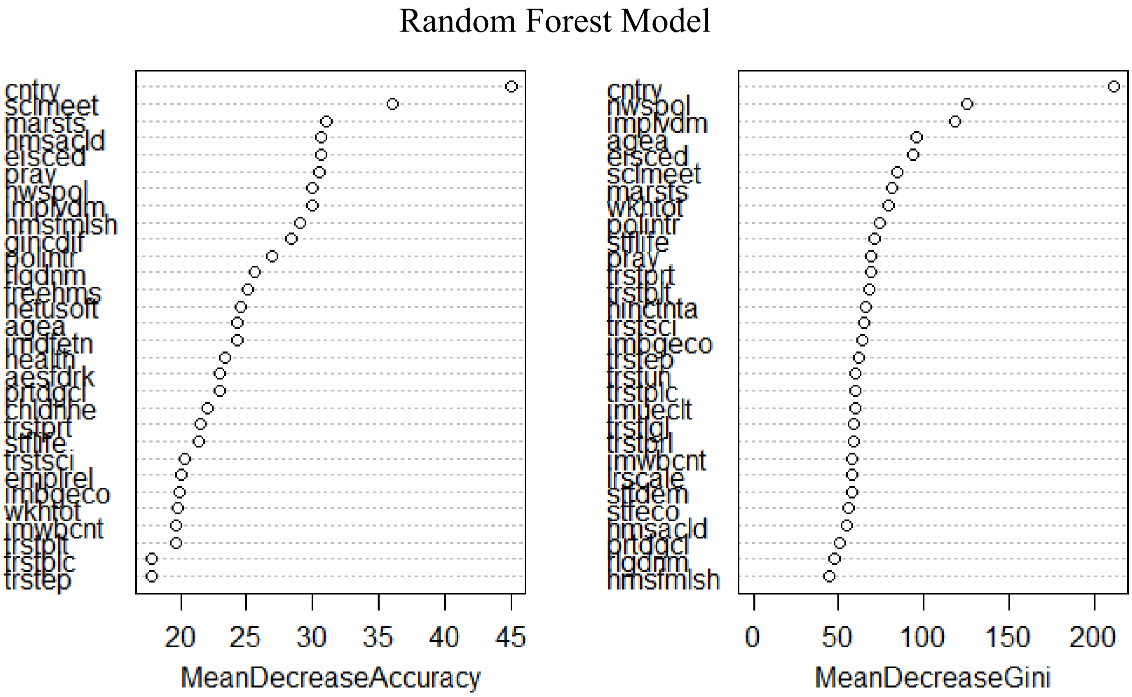
We continue by presenting the results of the random forest model, displaying the importance of features using the mean decrease accuracy (MDA) and mean decrease Gini (MDG). The former measures the impact of each feature on model accuracy by shuffling its values and observing the change in accuracy. The latter evaluates feature importance based on the reduction in Gini impurity, indicating the feature's role in making accurate splits in the decision trees.

Table 5: Mean decrease accuracy and mean decrease Gini for the predictors of voting behavior in the random forest model.

Predictor	Mean Decrease Accuracy	Mean Decrease Gini
agea	24.329581	95.4775590
gndr	17.552631	14.1522725
eisced	30.604529	93.0540658
cntry	45.042295	210.9978410
marsts	30.985632	81.0287464
chldhhe	22.003832	39.6307314
health	23.347174	44.7614602
trstprl	16.508326	58.4969396
trstlgl	16.762464	58.8081770
trstplc	17.750020	59.4192539
trstplt	19.617507	67.9945247
trstprt	21.437622	68.4761225
trstep	17.743127	61.9896929
trstun	16.918184	59.5315930
trstsci	20.308462	64.9556288
polintr	26.881258	74.0785597
prtdgcl	22.891615	51.0014946

stfdem	15.941027	57.4123207
implvdm	29.925599	118.3378017
lrscale	15.008227	57.6999224
nwspol	29.946748	124.7799388
freehms	25.098925	41.3459973
hmsfmlsh	29.054798	44.7671155
hmsacld	30.662175	54.7798353
sclmeet	36.097384	83.8111038
stflife	21.339440	70.7206653
aesfdrk	22.971652	32.2192424
crmvct	13.835991	12.3298979
netusoft	24.535845	37.3223638
stfeco	17.139921	55.6555708
gincdif	28.342983	43.0788818
hinctnta	16.373521	65.8962225
emplrel	20.084534	21.0605925
wkhtot	19.843528	79.0500416
rlgblg	15.249218	16.2539650
rlgdnm	25.596348	47.3315584
pray	30.555774	68.7458515
imdfetn	24.237098	37.9356186
imbgeco	19.879557	64.1193449
imueclt	17.229829	59.2939932
imwbcnt	19.667731	57.8407241
dscrgrp	11.463363	6.3140628
dscrrce	7.977715	2.1919835
dscrntn	6.609918	1.1116172
dscrrlg	5.754533	1.5891283
dscrlng	3.158482	0.3712079

Graph 1: Mean decrease accuracy and mean decrease Gini for the predictors of voting behavior in the random forest model.



In this model, the country where a person lives is the most crucial factor for predicting voting behavior, with a mean decrease accuracy of 45 and a mean decrease Gini of 210. Beyond the top spot, there are some differences in the ranking of predictor importance between mean decrease accuracy and Gini. For MDA, the frequency of meetings with friends, the marital status, the opinion on gay right to adopt children, and the education level are the other most important predictors. Whereas for MDG, the amount of time spent on political news, the importance given to living in a democracy, the age, and the education level are the other important variables for predicting voting behavior.

Table 6: Confusion matrix between predicted results and actual category of the test set for the random forest model.

		Reference	
		Voted	Not Voted
Prediction	Voted	1871	112
	Not Voted	168	59

The accuracy of the random forest model on the test set is 0.87, which is good. The sensitivity is 0.92, indicating very good performance, but the specificity is 0.34, which is poor.

We continue by presenting the results of the gradient boosting model, highlighting the relative influence of the variables in predicting voting behavior.

Table 7: Relative influence of the predictors in the gradient boosting model.

Predictor	Relative Influence
nwspol	19.56504868
implvdm	16.06362586
cntry	15.41644215
agea	10.20728962
marsts	9.06179933
polintr	5.91965797
wkhtot	5.23695893
eisced	3.32898026
sclmeet	2.36325708
prtdgcl	2.30112323
trstprt	2.20567195
stflife	2.13593771
rlgdnm	1.89074523
trstplt	0.83713045
health	0.60258389
imbgeco	0.57509097
pray	0.56381587
imdfetn	0.33341225
trstsci	0.29967379
chldhhe	0.23627904
hinctnta	0.23546970
trstep	0.12753178
netusoft	0.11364212
hmsfmlsh	0.07760212
imwbcnt	0.05580860
crmvct	0.05313471
hmsacld	0.04694454
gincdif	0.03427033
aesfdrk	0.03311773
trstprl	0.02860712
trstplc	0.02742474
trstun	0.02192224
gndr	0.00000000
trstlgl	0.00000000

stfdem	0.00000000
lrscale	0.00000000
freehms	0.00000000
stfeco	0.00000000
emplrel	0.00000000
rlgblg	0.00000000
imueclt	0.00000000
dscrgrp	0.00000000
dscrrce	0.00000000
dscrntn	0.00000000
dscrrlg	0.00000000
dscrlng	0.00000000

The variables with the highest relative influence are the time spent on political news, the importance given to living in a democratically governed country, the country where you live, the age, the marital status, the level of interest in politics, and the total amount of hours worked per week.

Table 8: Confusion matrix between predicted results and actual category of the test set for the gradient boost model.

		Reference	
		Voted	Not Voted
Prediction	Voted	1819	111
	Not Voted	220	60

The accuracy of the gradient boosting model on the test set is 0.85, which is good. The sensitivity is 0.89, indicating very good performance, while the specificity is 0.35, which is poor.

6. Conclusion

This study investigates the determinants of voting behavior using a combination of theory-driven and data-driven approaches. By comparing logistic regression, random forest, and gradient boosting models, we identify key factors that influence voting behavior.

In the theory-driven approach, we found that the variables used are statistically significant except for age and gender. Specifically, individuals with high levels of education and income are more likely to vote, and the interest in politics and the closeness to a party are associated with a higher probability of voting. Surprisingly, more time spent on political news is associated with a lower likelihood of voting, which is the only result that do not align with the literature, it may be because this variable is not a good proxy for political knowledge.

In the data-driven approach, we discovered that the country of residence is a crucial predictor of voting behavior. This variable shows a 99.9% statistically significant difference among its categories in logistic regression, is the most important feature in the random forest model, for both mean decrease accuracy and Gini, and holds the third highest relative influence in the gradient boosting model.

Other important predictors include the time spent on political news, which has the highest relative influence in the gradient boosting model, ranks second for mean decrease Gini and seventh for mean decrease accuracy in the random forest model, and holds a 95% significance level in the logistic regression model. The importance placed on living in a democratically governed country is another critical predictor, with the second highest relative influence in the gradient boosting model, third for mean decrease Gini and eighth for mean decrease accuracy in the random forest model, and a 95% significance level in the logistic regression model. Marital status also emerged as important, with the fifth highest relative influence in the gradient boosting model, third for mean decrease accuracy and seventh for mean decrease Gini in the random forest model, and a 99.9% significance level for some categories in the logistic regression model.

Some variables showed different levels of importance across models. For instance, age is the fourth most influential predictor in the gradient boosting model and ranks fourth for mean decrease Gini, but is fifteenth for mean decrease accuracy and is not statistically significant in the logistic regression model. Another example is the opinion on whether gay and lesbian couples should have the right to adopt children, which ranks fourth for mean decrease accuracy and shows a 99.9% significance level for one of its categories, but ranks lower in mean decrease Gini and relative influence.

Overall, the results from the theory-driven and data-driven approaches are generally coherent, with variables that are statistically significant in the former also being significant in the latter.

The model with the best accuracy is the random forest model, followed by the gradient boosting model, then by the data-driven logistic regression model, and lastly by the theory-driven logistic regression model. However, the theory-driven logistic regression model has the best specificity, with the data-driven logistic regression model as a close second, while the random forest and gradient boosting models are very lacking in this aspect.

Future research should continue to explore these relationships between voter turnout and the various predictors using diverse datasets and advanced analytical techniques to build on the findings of this study and further elucidate the drivers of voting behavior.

7. References

Blais, A., & Dobrzynska, A. (1998). Turnout in electoral democracies. European journal of political research, 33(2), 239-261.

Blais, A. (2006). What affects voter turnout? Annu. Rev. Polit. Sci., 9(1), 111-125.

Cancela, J., & Geys, B. (2016). Explaining voter turnout: A meta-analysis of national and subnational elections. Electoral Studies, 42, 264-275.

Fornos, C. A., Power, T. J., & Garand, J. C. (2004). Explaining voter turnout in Latin America, 1980 to 2000. Comparative political studies, 37(8), 909-940.

Franklin, M. N. (1999). Electoral engineering and cross-national turnout differences: what role for compulsory voting? British Journal of Political Science, 29(1), 205-216.

Geys, B. (2006). Explaining voter turnout: A review of aggregate-level research. Electoral Studies, 25(4), 637-663.

Geys, B. (2006). 'Rational'theories of voter turnout: a review. Political Studies Review, 4(1), 16-35.

Kostadinova, T. (2003). Voter turnout dynamics in post‐Communist Europe. European journal of political research, 42(6), 741-759.

Lijphart, A. (1997). Unequal participation: Democracy's unresolved dilemma presidential address, American Political Science Association, 1996. American Political Science Review, 91(1), 1-14.

Matsusaka, J. G., & Palda, F. (1999). Voter turnout: How much can we explain? Public Choice, 98(3), 431-446.

Powell, G. B. (1986). American voter turnout in comparative perspective. American Political Science Review, 80(1), 17-43.

Smets, K., & Van Ham, C. (2013). The embarrassment of riches? A meta-analysis of individual-level research on voter turnout. Electoral Studies, 32(2), 344-359.

The codebook and code used for this analysis are available on:
https://github.com/TommasoGrotto2/CCS-Assignment