Faculty of Social Sciences
Department of Economics
University of Copenhagen

Co-curricular Assignment

Marius Grünewald

Local Election Predictions
A machine-learning approach

Supervisor: Andreas Bjerre-Nielsen
ECTS points: 30
Date of submission: May 30, 2019
Keystrokes: 48000

Local Election Predictions

A machine-learning approach

Marius Grünewald

July 30, 2019

Abstract

This paper explores the potential for machine learning algorithms to predict local election outcomes. In order to do so, data from several sources is collected and web scraping applied. To measure social connectedness, Google search results of each candidate are used as a proxy. Gradient boosting seems to be the most promising method for prediction with an R^2 score of about 0.6. Evaluation on an aggregated seat level yields a score of 0.8. An ensemble estimation of two parts gradient boosting, two parts random forest and one part lasso returns similar results. While these scores are decent as a first attempt, future research could try to improve the results by optimizing the ensemble or test a neural network.

Contents

1 Introduction

Predicting national elections has a tradition dating back to 1916 in the United States (Walker (2006)) and 1945 in Great Britain (Whiteley (2005)). Nowadays, it is praticed frequently with a broad range of different methods. The most common ones are opinion polling, betting markets and predictive models.

While there is plentiful supply of either method in almost every developed country on the national and state level, local elections have not experienced the same trend yet. A telling example of this can be found when comparing some Wikipedia pages of the recent Spanish local elections, where few polling data can be found.[1] On a national level, the situation looks different with plenty polling data available.[2] Similar lack of thorough predictions can be found for local elections in the UK and Germany.

This shortcoming is rather intuitive due to several reasons. Firstly, it is expensive. Even exploiting correlations between similar entities, various different municipalities would have to be surveyed while just the process of identifying a representative sample might be even more intense due to restricted micro data and a smaller sample overall. Secondly, lack of interest of national groups. The limited spatial scope of local election may cause national players with the necessary funding to allocate money according to their preferences with are more aligned with national polling. Lastly, the highly personal relationship between candidates and voters in local elections is hard to capture. Since many countries, for instance Germany, Italy or Denmark, use some form of proportional representation with open lists voters can vote cross-party and accumulate more than one vote for a candidate. These factors make personal connection more important than in any other type of election. Further, the small size of some communities make personal exchange and contact more likely in the first place. This should enable voters to judge the character and policies better in local elections relative to any other election.

To address this, I collect data from various, decentralized sources and apply web scraping to get additional measures. This data allows to test and filter among different estimators. In a nutshell, I find that gradient boosting regression and an ensemble of regressor perform best in predicting local election outcomes.

This paper proceeds as follows. Section II explains the electoral regime for local election in the state of Baden-Württemberg followed by section III outlining the applied methodology more carefully. After that, section IV lays out the data that fuel the analysis. The results of the analysis are presented in section V. The conclusion in section VI sums up the entire study and looks for improvements in future research.

[1] Wikipedia contributors (2019b)
[2] Wikipedia contributors (2019a)

2 Local elections in Baden-Württemberg

Before examining the methodology more thoroughly, some discussion of the electoral system comes in handy. It explains why the fundamental problem is a regression problem and not a classification problem. Further, it yields insights regarding the set-up of the regression.

The following description relies heavily on a book by Quecke et al. (2019) which explains the elections procedure carefully.

In short, it can be described as a proportional voting system with open lists[3]. It means that a voter sees all lists with candidates participating in the election. In total, she has as many votes as there are seats in the local council, however less votes are feasible, too. The exact number of seats depends on the size of the municipality with largest in the sample being Stuttgart with 60 seats.

The process of voting follows a relatively complicated manner. First, one has to select a primary list. An uncommented, returned list gives each of the candidates on this list one vote. Subsequently, the amount of votes equals the number of candidates on that list - a list cannot hold more candidates than seats that are up for election. The second step allows for more individualization. Here, a voter can accumulate up to three votes for one candidate, effectively reshaping the (ranked) preferences of the list. Further, he can write in individual candidates from the other lists. The cross-voted votes later count to the original list of the cross-voted candidate. Accumulated and cross-voted votes are commonly known as the secondary vote.

The next step is to calculate the number of seats assigned to each list. Hereby, all votes cast for candidates on a list are added and quotients X calculated using equation (1)

$$X = \frac{V}{2s+1} \tag{1}$$

where V equals the tallied votes for a list and s the number of seats already secured by the same list. Table (1) helps to create an understanding of what this means in praxis. There are three parties with three candidates for three seats - a simplified example.

Table 1: Sainte-Lague Procedure

Denominator	A	B	C
$2*0+1$	33000	10000	51000
$2*1+1$	11000	3333	17000
$2*2+1$	6600	2000	10200

Now, the three biggest numbers are being selected, 51000, 33000 and 17000. Therefore, list **C** won two seats, list **A** one seat and list **B** is left with not a single seat. It is important to note that there is no minimum threshold to enter.

[3] A list is to be seen as a number of candidates that run for the same party/coalition

One particularity of the state of Baden-Württemberg is the concept of the *"Unechte[n] Teilortswahl"*[4] in which sub-lists for certain areas within one municipality are created. A fixed amount of seats is allocated to and reserved for these sub-lists. However, all the municipality can vote for the entire set of sub-lists. Whether a city applies this concept is within their own legal jurisdiction and subject to change.

In this sample there is one city which makes use of this procedure. However, the different structure of the data originated in cities using the *Unechte Teilortswahl* might be causing trouble and is therefore omitted.[5]

3 Methodology

Predictive modelling has risen to new popular heights, due to a surge of private and commercial modelling such as the projects of *FiveThirtyEight*[6], *Inside Elections*[7] or *Sabato's Crystal Ball*[8]. I will employ a, broadly speaking, similar approach.

The crucial difference to many higher-level predictions is the relevant dependent variable to predict. While the mentioned groups predict national elections on state-level to match the American electoral college, they do not go deeper than electoral districts to predict Congressional elections. To apply the same logic, one would have to predict on a municipal level to see how many votes/seats a party gets. However, this approach neglects the crucial personal components described earlier. Therefore, my approach is to forecast the number of votes for each candidate. From there, we can aggregate to see how many seats each party gets if a proportional voting system is employed.

Due to nature of the voting system, a probabilistic approach is harder to make us of. As explained above, the number of seats a party gets depends on the sum of all candidates' votes on the same list. The individuals on each list are sorted according to their votes. Therefore, the crucial measure is the aggregated sum of votes or, in other words, the strength of the list. Classifiers are not suitable for this type of problem making me chose regressors. Classification in election prediction would be to estimate the probability of getting elected. This is a sensitive approach in case of a first-past-the-post voting system. Here, the crucial component is the sum of all candidates on a list, which is not predictable by classification. The next paragraph will dive deeper into the choice of regressors.

The most intuitive regressor is a linear regression. This is good start but might not be sufficient. Since local elections can be very specific, a researcher has to think in many facets. For instance, a local election can be influenced by national situations such as a weak incumbent government or an issue being mostly debated on a federal level, e.g. foreign policy (Thomsen (1998)). Local issues be a driving force just as well. A noteable example in the state of Baden-Württemberg is the new train station in Stuttgart (*Stuttgart 21*) (Keil and Gabriel (2012)). This complex structure of data requires

[4] No formal translation to English
[5] In more detail, it refers to the variable *"place on list"*
[6] Silver (2019)
[7] Gonzalez (2019)
[8] Sabato (2019)

multiple predictors to account for the potentially simultaneous trends exerting influence on voters. Subsequently, linear regression may be unsuitable. With many variables, OLS tends to have large variances and low bias as Tibshirani (1996) shows. Therefore, prediction accuracy can be improved by trading some bias for significantly lower variance and a regularized regression approach is likely helpful. Further, I consider different statistical methods for comparison. These are a random forest regressor, support vector machine regression and gradient boosting regressor and an ensemble estimator.
To fix ideas, equation (2) shows the baseline linear regression equation

$$Y_{it} = \beta X_{it} + Q'_{jkt}\delta + Z'_{jt}\gamma + V'_{kt}\pi + U'_t\nu + \epsilon \tag{2}$$

Here, i indexes the individual candidates and j the parties while k represents municipalities. Lastly, t stands for the election cycle. I choose the more detailed subscription to highlight the different sources of predictors and thus the high dimensionality. Y_{it} is the dependent variable. It captures the votes of each candidate gets in an election, by not including j and k, I make the implicit assumption that a candidate can not stand for two parties or in two localities at the same time. X_{it} stands for individual-level explanatory variables where the same assumption has been made. *Google search results* or *incumbency* can be mentioned as examples.
Z_{jt} collects all variable that are constant for individuals in the same party but constant across municipalities. Such a variable can be the popularity of parties at the national level that might influence voters' decision making.
V_{kt} gathers all factors that are constant for candidates and parties but vary across time and localities. Mainly, it consists of city demographics and socioeconomic data on the city level.
Some parties perform wildly different in particular cities than in the national environment. As of this, Q_{jkt} is included. It captures the performance of a party in a previous election in the city. The previous election is explicitly not a local election.
Ultimately, I collect variables that vary only across time. U_t represents it. Time trend is the most obvious example.
To get a better understanding about the entire set of variables and their properties provides the upcoming section a detailed overview.

4 Data

This section is structured in a way that it first explains the crucial variables and later provides descriptive statistics. Collecting the relevant data is a great challenge. To begin with, one has to collect all the voting records on a candidate level, the left-hand-side variable. In practice, there is no central record for municipal election results which in turn forces the researcher to access every city individually and search for election records.
The second, important measure newly introduced in this paper is Google search numbers aiming at

measuring social interconnectedness. To collect this data, I search for every candidate and retrieve the results via web-scraping. Due to common names that occur multiple times, the search is further restricted to city and year. To fix ideas, I search with the following algorithm

```
result = 'https://www.ecosia.org/search?q=' + names[i] + ' ' + year[i] + ' ' + city[i]
```

I expect this improved search strategy to be of better accuracy in measuring the social interconnectedness of candidates which is especially important in municipal elections.

Some fundamental concerns regarding this measure have tp be addressed, though. The first disadvantage points to the limited time, since the internet has not been accessible for the broad public before the late 1980's and early 1990's (Leiner et al. (2009)) it does not contain real time information about election before that time either. Subsequently, any analysis has to be performed for data starting in the 1990's for this technical reason.

This leads straight to the next problem with using Google search results. Especially in the early stages, there has been a dramatic increase in news reporting online as well as news consumption online (Greer and Mensing (2006)). Local newspaper are most likely to report frequently about municipal policies and elections with articles than local newspapers, according to Richardson and Franklin (2004), which generate coverage of these mostly via reader's letters. Further they find, that local newspaper do overall more reporting on local elections. However, digitization of these happened gradually so that a full picture is hard to draw in the '90s with many newspapers not starting publishing article online too before the early 2000's (Krueger and Swatman (2004)). Therefore, the earliest election considered is 2005. This should reduce any bias resulting from digitization of news reporting and measure social interconnectedness more accurately.

A last concern is whether Google searches might be biased towards a younger generation. The argument is that older voters have a set of activities and connections that is not accurately reflected in the online reporting of newspapers and social media. This is, at least partially, true. Mellon and Prosser (2017) find that the average Twitter user is significantly younger than the non-user. The same trend can be said for Facebook-users but to a lesser extent. Since social media profiles are also part of the search results, we may exaggerate the social connectedness of young people compared to old.

Besides, several other controls are included. These can be categorized as done in the equation in the previous section, candidate-level predictors, municipality-level predictors, county-level predictors, state predictors and national predictors.

This list is certainly not exhaustive but a reasonable subset of possible predictors. Individual-level predictors are variables that describe an individual candidate. All these variables are scraped or deducted from a party's nominated list.

$Woman$ (Friedhoff et al. (2015)), *Muslim background* and *(Non-Muslim) immigration background* (Schonwalder (2013)) account for negative discrimination against population groups that have been traditionally under-represented in local councils. Having a PhD is considered to make a positive

Table 2: Individual-level controls

Variable	Mean	Std. Dev.	Min.	Max.	N
Google search results (standardized)	4.24e-09	1	-.119657	88.01223	11567
Woman	0.374	0.484	0	1	11567
Place on list	23.158	13.953	1	60	11567
Incumbent	0.077	0.266	0	1	11287
Doctor	0.088	0.283	0	1	11567
Aristocracy	0.007	0.081	0	1	11567
Muslim background	0.033	0.179	0	1	11567
(Non-Muslim) immigration background	0.057	0.232	0	1	11567

difference in local elections and is therefore included (Friedhoff et al. (2015)). The incumbent effect has been proven important, for the case of local election consider Trounstine (2013), and is likely to help improve predictions. Lastly, *Place on list* captures on what number on the list a candidate is located. As explained earlier, this is particular to the proportional voting system of Sainte-Lague/D'Hondt which is applied in several German states and for other national assemblies.

Further, I start to collect data on candidates' age and occupation. However, historic voting records rarely contain these information limiting us to concurrent data. Restricting the sample to observations containing this information would reduce the over sample size dramatically nurturing significant doubt about the properties of estimator and regularization. Yet, future predictions might be able to include these measures.

National controls focus on parties and previous elections. From row 4 onwards, all rows in Table 3 are dummies that captures parties competing in national elections. Rows 2 and 3 report the performance of these parties at the last federal election (row 2). *Federal difference* is composed as the difference of a (national) poll conducted three months prior to the local elections and the last federal results of the party. Hereby, I attempt to measure the short-level trending of a party while the *federal elections*.

Table 3: National controls

Variable	Mean	Std. Dev.	Min.	Max.	N
Federal election	9.779	11.904	0	67.8	11567
Federal difference	0.507	4.785	-14	54.8	11567
CDU	0.105	0.307	0	1	11567
SPD	0.104	0.306	0	1	11567
FDP	0.099	0.299	0	1	11567
Grüne	0.1	0.3	0	1	11567
AfD	0.044	0.205	0	1	11567
Linke	0.09	0.287	0	1	11567
Partei	0.024	0.152	0	1	11567
Piraten	0.019	0.138	0	1	11567

Lastly, Table 4 shows the remaining variables from either of the two previous levels. They represent several demographic and socioeconomic predictors.

Additionally, I consider the interaction terms between all variables as well as their first polynomial. Subsequently, the total number of predictors is roughly 850.

To make all data sources as transparent and detailed as research requires, the following Table

Table 4: Municipal and county controls

Variable	Mean	Std. Dev.	Min.	Max.	N
Share foreign tourists	0.275	0.074	0.097	0.404	11567
Unemployment	5.336	1.14	3.3	8.6	11567
Population	253828.873	177215.275	47164	632743	11567
Share migrants	0.181	0.039	0.106	0.246	11567
Share youth	26.997	2.966	22.4	35.7	11567
Share old	28.481	3.633	19.5	37.7	11567
Share students	0.128	0.067	0.011	0.327	11567
Share pupils	0.107	0.016	0.087	0.165	11567

To check for obvious correlations prior to combined attributes, Figure 1 shows the correlation coefficients of the basic predictors.

If we see plenty dark and very light colours on the map, correlations are generally very strong. This implies that the predictive power of the set of variables is weak since they all explain a similar variance. We see that most colours are somewhat purple-shaped. That implies a correlation coefficient of around 0. A very strong, positive correlation can be found between *share_youth* and *share_old*. The same on the other side of the correlation spectrum, indicated by black colour, can be seen for *share_old* and *share_students*. When performing the polynomials and interaction, this may multiply since there are several dummy variables. All in all, there seems to be sufficient diversity in the ground variables applied.

5 Results

Generally, the algorithms are evaluated on two levels. Since we are estimating the individual level of votes for each candidate, the strength of the estimators are being directly evaluated on the individual level. But equally interesting from a political point of view is the number of seats a party wins in an election. Almost universally, parties do celebrate their overall performance in terms of seats rather than in terms of which person wins a seat, at least when it comes to proportional voting systems.[9] Therefore, I will aggregate the results and see how the algorithms perform in this measure. Since it is a less detailed measure, chances are that the algorithm perform better there.

Before we turn to the results, I want to briefly explain the more advanced estimators and the different parameters needed to be set before estimating.

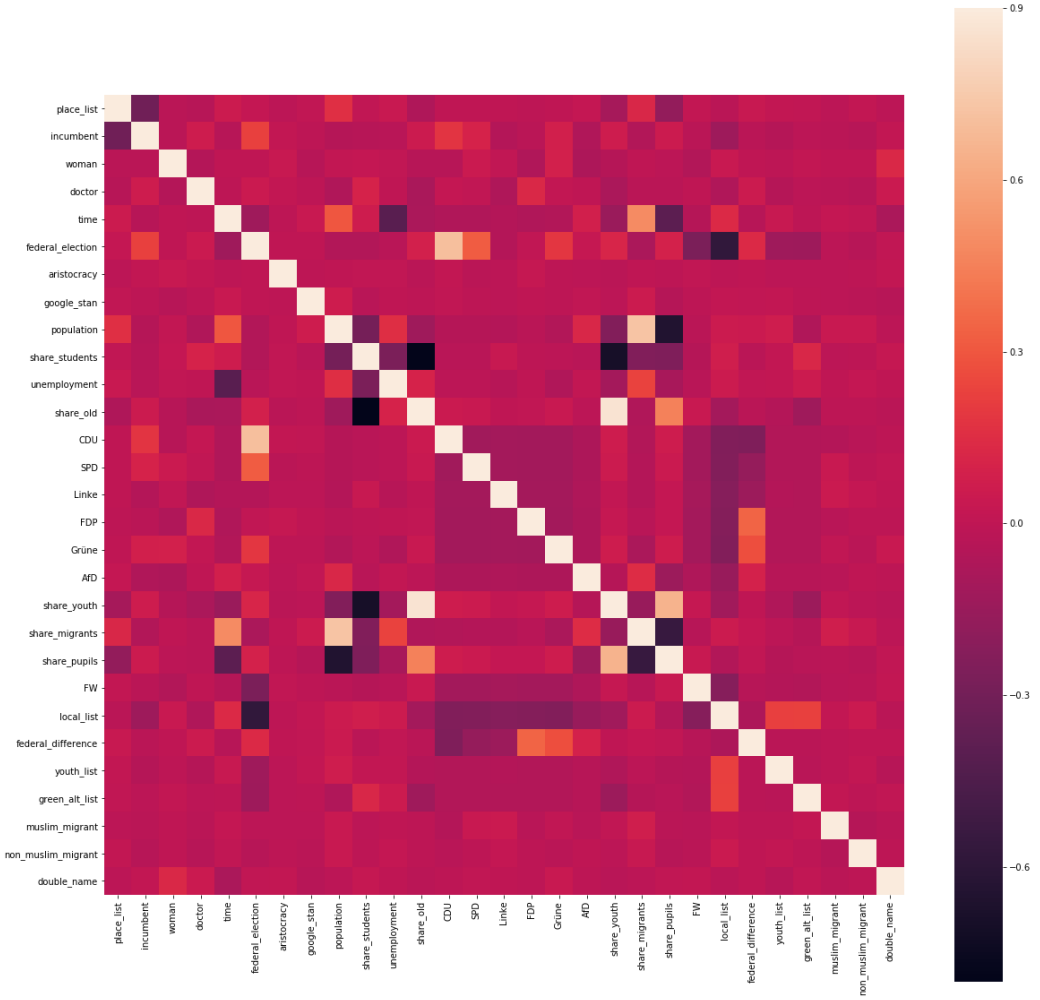
The first algorithm is the *Least Absolute Shrinkage and Selection Operator* (Lasso) first presented by Tibshirani (1996). Boiled down, the Lasso minimizes a loss function, similar to the OLS, with a penalization equal to the sum of the absolute values of all coefficients included in the estimation. Therefore, it may occur that sparser model are preferred and some coefficients set to zero. Mathematically, we

[9] Radio (2019)

Table 5: Data Sources

Variable	**Source**
Google search results (standardized)	Web scraping
Woman	Election lists from respective city
Place on list	Election lists from respective city
Incumbent	Previous election records from respective city
Doctor	Election lists from respective city
Aristocracy	Election lists from respective city
Muslim background	Election lists from respective city
(Non-Muslim) immigration background	Election lists from respective city
Share foreign tourists	Baden-Württemberg (2019)
Unemployment	Baden-Württemberg (2019)
Population	Baden-Württemberg (2019)
Share migrants	Baden-Württemberg (2019)
Share youth	Baden-Württemberg (2019)
Share old	Baden-Württemberg (2019)
Share students	Baden-Württemberg (2019)
Share pupils	Baden-Württemberg (2019)
Federal election	Cantow et al. (2019)
Federal difference	Cantow et al. (2019)
CDU	Election lists from respective city
SPD	Election lists from respective city
FDP	Election lists from respective city
Grüne	Election lists from respective city
AfD	Election lists from respective city
Linke	Election lists from respective city
Partei	Election lists from respective city
Piraten	Election lists from respective city

Figure 1: Correlation Heatmap



minimize the following equation with respect to β such that

$$\hat{\beta} = \arg\min \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \beta_j \right\} \tag{3}$$

It crucially depends on the λ parameter which assigns a weight to the penalization process. It must hold that $\lambda \geqslant 0$. Intuitively, a high lambda decreases variance and accepts a higher bias while a low lambda trades a higher variance for a lower bias. Table 6 reports the result of 10-fold cross-validated λ estimation.

Support Vector Machine Regression (SVR) is the second approach.

$$J(\beta) = \frac{1}{2}\beta'\beta + \mathbf{C}\sum_{n=1}^{N}(\xi_n - \xi_n^*) \tag{4}$$

$$\text{s.t. } \forall n : y_n - (x_n'\beta + b) \leqslant (\epsilon + \xi_n) \tag{5}$$

$$\forall n : (x_n'\beta + b) - y_n \leqslant (\epsilon + \xi_n^*) \tag{6}$$

$$\forall n : \xi_n \geqslant 0 \tag{7}$$

$$\forall n : \xi_n^* \geqslant 0 \tag{8}$$

A SVR minimizes the number of predictors used for a given function, function (4). This given function has to make sure that no residual is greater than the initially defined tolerance margin (Cortes and Vapnik (1995)), as expressed in equations (5) - (8). I choose a linear relationship for the underlying kernel since the dataset is not too big. Intuitively, a linear kernel forces the function minimized to be linear and therefore be seen as a standard, linear regression line. Additionally, we have to hypertune the cost-function parameter C and the learning rate γ[10]. Again, the results can be seen in Table 6.

The third algorithm considered is a *Random Forest*. A random forest consists of several decision trees. A decision tree minimizes the number of splitting points and splitting variables from the entire dataset. The strength of variables and points is determined by the mean squared error of a regression with the specific set of variables/points minus the average of the outcome variable (Hastie et al. (2005)). Typically, a random forest bags many decision trees with an (un)weighted average to reduce variance and keep the bias low.

There are, however, some parameters that need fine tuning. They are presented in Table 6 alongside the estimated parameters.

Fourthly, the *Gradient Boosting Regression* is applied. In a nutshell, gradient boosting follows a sequential minimization strategy. A regression with a predictive feature[11] is estimated and the resulting error terms saved. (Hastie et al. (2009)) These error term is the new left-hand-side variable being explained by the same feature. The predicted error terms are, in response, added to the previous prediction of the output variable. This combination of predicted error terms and output values is the new explanatory variable. With this new variable the cycle starts again. Commonly, the average is used as the first prediction.

As with the other estimators, the fine-tuned parameter are reported in Table 6.

Lastly, I make use of an *Ensemble Estimator*. While technically random forests and gradient boosting are ensemble estimators too, the ensemble estimator combines different types of regressors not only the same in different shapes. I hope to achieve higher robustness and improved accuracy as argued by García-Pedrajas et al. (2005).

The weights are being tested by cross-validation from a subset of double and single weighting. Higher-

[10] The learning rate is not included in the simple formula presented above. It is only applied in the Dual Formula which is superior for computation but less intuitive

[11] The predictive feature can be thought of in the same fashion as the random forest. Each iteration, we can add another splitting point and/or splitting variable.

order-weights are not being considered due to limitation in computational power.

Table 6: Hypertuned Parameters

Estimator	Parameter	Value
Lasso		
	α	302.11
SVR		
	C	4150
	γ	0.0005
Random Forest		
	Max. Depth	90
	No. Estimators	180
	Min. Sample Leafs	3
	Min. Sample Split	10
Gradient Boosting		
	Max. Depth	5
	No. Estimators	180
	Min. Sample Leafs	60
	Min. Sample Split	300
Ensemble Regression		
Weights:	Lasso	1
	Random Forest	2
	Gradient Boosting	2
	SVR	0

As a next step, we can turn to the results and discuss the performance of the estimators for the training dataset. Generally, one does not put too much emphasize on the training scores since overfitting is looming above all high scores, even with cross-validated scores as I have done here. That is mainly because of the nature of election forecasts. Here, the training set captures data from 2009 and 2014. The test set, however, is exclusively made of 2019 data in order to mimic a real-world election forecast. The comparison between training and test scores may shed light on the out-of-sample power of each algorithm.

Cross-validated scores are not only correcting some overfitting, they also allow to compute standard errors. This can help address the precision of the estimator.

We can see that the linear regression does not do a very good job in predicting even within the training set. The negative R^2 arrives from the inclusion of non-explanatory variables. Gradient Boosting and Ensemble regression are the predicting algorithms for the training set just of a little of short of 70 % of an explained variance in the outcome variable. The three other estimators perform between 0.28 and 0.49.

The poor performance of the linear regression can likely be attributed to the complex nature of the data. As expected, there are multiple interdependencies of regressors as well as polynomial structures. The linear regression does not incorporate that and even if all interactions and polynomials were to be included, we would still expect a poor performance due to gross overfitting. The ability of the other estimators to handle larger sets of predictors and simultaneously control for the bias-variance trade-off pays off.

In terms of precision, it appears that the lasso estimator experiences significantly higher variances than

Table 7: Scores Training Set

Estimator	**RMSE**	R^2 **Score**
Linear Regression	26993.19	-56.24
	(53132.29)	(164.17)
Lasso with Cross Validation	5916	0.4182
	(4994.4046)	(0.4018)
Support Vector Machine Regression	6337.142	0.2830
	(3713.366)	(0.188)
Random Forest Regression	5193.19	0.49
	(3094.34)	(0.2971)
Gradient Boosting Regression	4257.50	0.6726
	(2536.60)	(0.1103)
Ensemble Regression	4052.69	0.686
	(2235.23)	(0.11)

Averages reported, standard errors in parentheses
All scores obtained by 10-fold cross validation
RMSE (Root Mean Squared Error) is defined as the square root of the negative error

the other estimators - the linear regression excluded. This might pose as a hint to a different score in the test set. The random forest behaves the same way, even though the reason is different. Potentially, the number of trees in the forest is too large relative to the data set since that might increase the variance of the set (Louppe (2014)).

As mentioned earlier, the score for the training sample, even though it is cross-validated, might be misleading. Therefore, Table 8 reports the scores for the test set, the election cycle 2019. In general, I followed the same procedure as for the previous evaluation. 10-fold cross validation is applied for the mean squared error and the R^2. The linear regression has been omitted since it's predictive power was far too low.

Table 8: Scores Test Set

Estimator	**RSME**	R^2 **Score**
Lasso with Cross Validation	6502.35	-0.130
	(4034.156)	(2.397)
Support Vector Machine	9217.69	0.2059
	(5116.909)	(0.1555)
Random Forest Regression	6540.61	0.42
	(3255.05)	(0.608)
Gradient Boosting Regression	5461.13	0.5941
	(1821.165)	(0.3089)
Ensemble Regression	6045.54	0.312
	(2708.90)	(1.066)

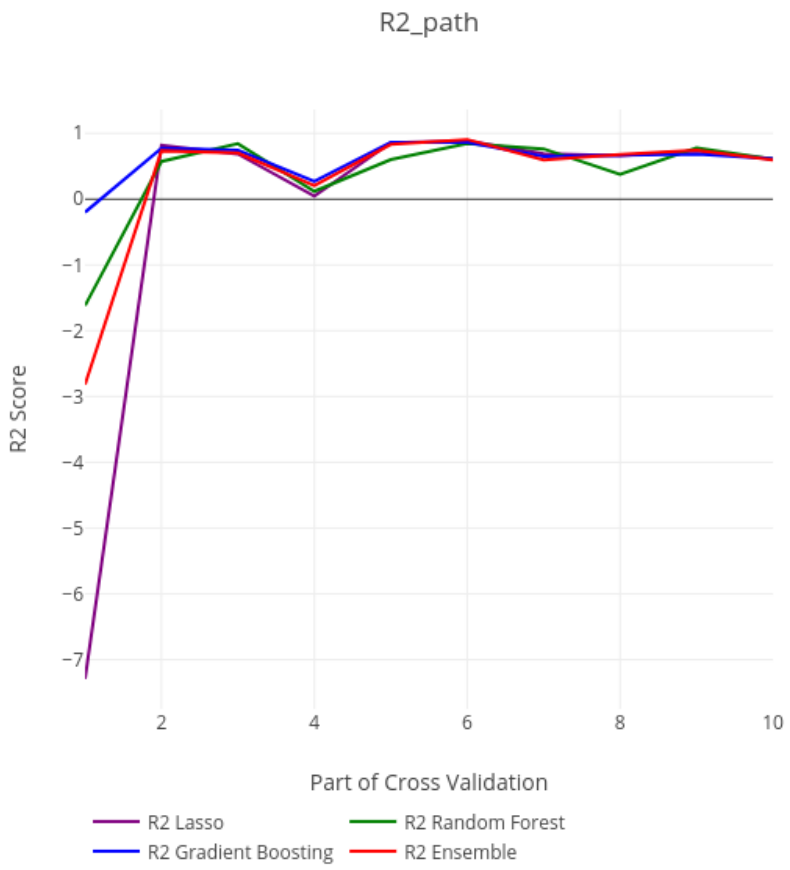
Averages reported, standard errors in parentheses
Linear Regression being dropped due to unreadable large RMSE
All scores obtained by 10-fold cross validation
RMSE (Root Mean Squared Error) is defined as the square root of the negative error

When inspecting the results closer, we find some interesting changes. The lasso estimate is performing significantly worse than previously. Given that the ensemble estimate is just a linear combination of, among other, the lasso, it is unsurprisingly worse, too. Gradient Boosting, SVM and Random

Forest perform slightly worse but all within reasonable range that might be explained by the different data structure.

More interestingly still, the standard deviations have increased impressively, especially for the lasso and, again a direct follow-up of the lasso, the ensemble estimator. A higher variance means a higher range of scores and in combination with the lower average score likely to include more downward movement. To test this hypothesis, I plotted the path of the three estimators the ensemble is founded in as well as the combination itself as well. Figure 2 is that plot.

Figure 2: R^2 path of selected estimators



We can see that the first split in the cross-validation is hard to predict for all estimators, particularly bad for the lasso. Since the estimators behave similarly for all other parts, we can attribute the increased variance and the overproportional drop for the lasso to the first split in the cross validation. Overall, the ensemble estimator usually scores between 0.6 and 0.9 for the different parts of the cross validation. In about the same range is the gradient boosting, from 0.61 to 0.86.

The support vector machine regression generally performs worse than the other estimators. Potentially, we can attribute this to the linear structure of the underlying kernel. Future research could explore this speculative hypothesis by running a polynomial kernel.

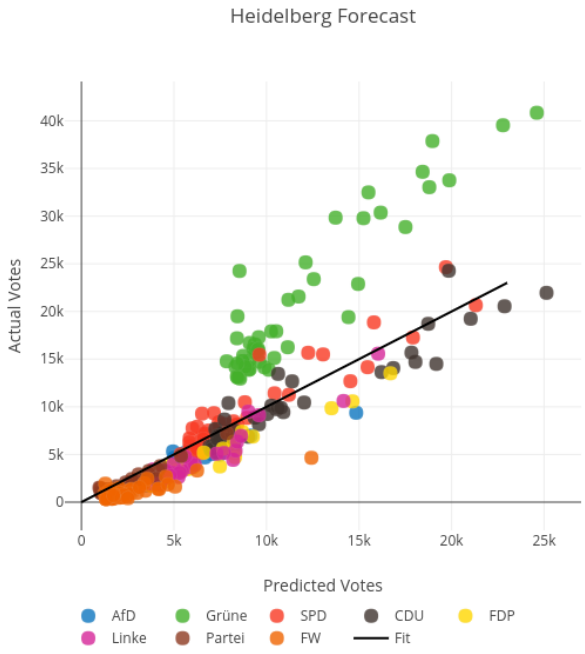
Before we turn to evaluate the estimator different level, it can prove insightful to plot predictions on

true values in order to see patterns that might be exploitable in future improvements. Therefore, I plotted the scatter plots of exactly this with three cities and the ensemble estimator. In all three cases, I omitted the several local lists since they would add an enormous amount of points that would make any distinction impossible. Further, most estimators perform most accurate with these local lists not giving much opportunity for constructive criticism.

Figure 3 plots the estimates for the city of Heidelberg.

The most obvious miscalculation lies within the green party. Despite the efforts of capturing all

Figure 3: Scatter plot Heidelberg



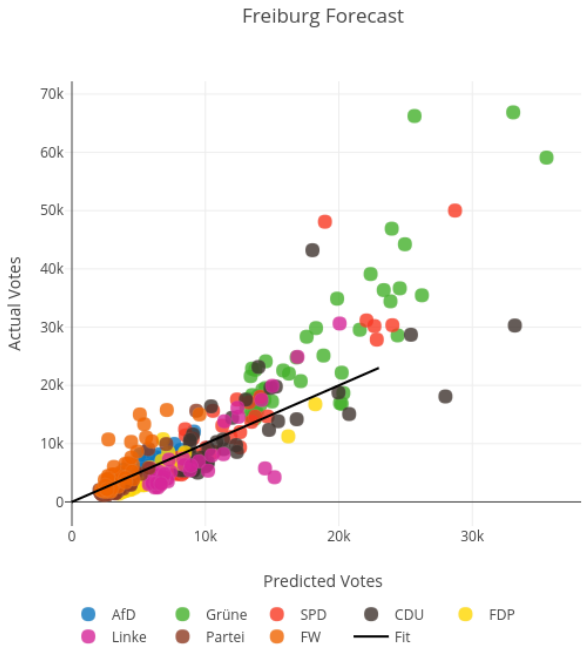
trends of parties in general it appears that this has not been successful. We can likely trace back the origins of the underestimation to a party and not to individual factors since the underperformance is consistent for all individuals. Additionally, it appears that a hypothetical linear line, slightly higher in slope of approximately 60 degrees, might be able to account for much in the variation not explained yet.

The picture looks somewhat similar for the city of Freiburg, as Figure 4 shows. While we can still draw a linear line to reduce the variance for the green points, the composition is more dispersed. Clearly, this points to individual factors not yet accounted for. Since this can be observed, in limited fashion, for the social democrats (red) as well as the conservatives (black), a reasonable hypothesis might be that their profession offers them a high degree of publicity. Potential professional careers could be politics, arts or entrepreneurship. Other categories are possible, too. One can think of time an individual has lived in a city or even the size of his family living in the city. While the latter two are hard to test empirically due to obvious data restrictions, the occupation might be testable and yet gathering employment records is time-consuming and certainly exceed the scope of this analysis.

Future research might be able to add this important component.

The last city discussed is Ulm. In Figure 5 it becomes evident that individual deviation is increasingly

Figure 4: Scatter plot Freiburg

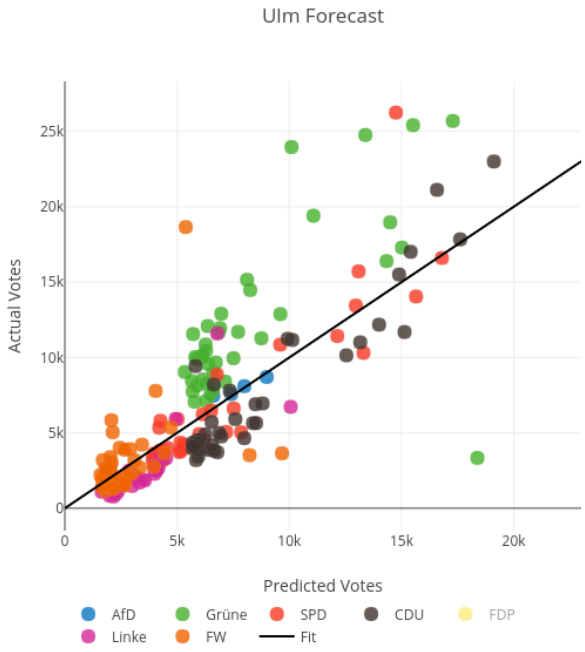


important compared to party deviation.

Not only are the green data points dispersed across the scatter plot but also the black (conservatives) and red (social democrats) estimates are less dense. Interestingly, it appears that cities with smaller populations are more individualized in their behaviour. Ulm is significantly smaller than Freiburg or Heidelberg. Potentially unobserved characteristics, such as an occupation that is open to contact with the electorate, might be more important in smaller populations. The next step is to evaluate the estimators on the party level. Again, I computed cross-validated mean squared errors and R^2 scores for all estimators. The picture that is drawn is slightly different, however. As can be seen in Table 9, the received scores are generally higher and further experience higher stability. That is not surprising since the aggregated data is divided by the sum of votes of the party. Outliers are flatted by the number of people on a list, if they aren't significant outliers themselves.

In essence, the most promising strategy appears to be the ensemble regression. It reports an average R^2 of 80% of explained variance with a standard deviation of 10%. It is trailed slightly by the gradient boosting with a lower score (75%) but higher precision (standard deviation of 7.5%). The lasso estimator does a decent job, too. Its R^2 score is about 0.71 with only 0.19 as the standard deviation. The standard errors for the three remaining estimators are evidently higher (0.29, 0.41 & 0.64) even though their prediction has been improved most compared to the candidate-level predictions. The linear regression, for instance, has now a positive score and the support vector machine regression

Figure 5: Scatter plot Ulm



doubles its previous score to 0.44.

The R^2 path during the cross validation is shown in Figure 6. We can observe that all estimators, except the gradient boosting and the ensemble estimators are dropping significantly in the last (two) cross-validating regressions. That is likely to be the major part in explaining the higher score as well as the higher precision for these two estimators compared to the other. Once again, the vector support machine does not the best job in explaining the variance, in effect it performs worse than the linear regression in many instances.

The ensemble estimator has a higher variance than the gradient boosting regressor (see 6). However, that seems to be caused by the upward defection of the ensemble estimator in 7th and 8th regression.

To filter a potential reason of inaccuracy, I take a closer look at several estimators and split the prediction according to party lines. In Figure 7, the seat prediction are plotted on the true seats. The 45-degree line stands for the perfect prediction. That being said, we would like to see all points on the line. Any deviation is simply a misprediction. Already a quick look makes clear that most predictions are either on the line or close by. The plot covers all cities which explains why colours that are associated with a particular party are appearing more than once.

As Figure 7 clearly demonstrates however, green party is generally underestimated.

Since it occurs in the same manner for the gradient boosting (Figure 8) as well as the support vector machine which is not part of the ensemble (Figure 9), it likely that this is an issue of missing data. The data fails to capture the trend of an increased popularity of the green party (*Die Grüne*) in Germany[12]. To correct for this a new measure has to be included. Future research could try to find a

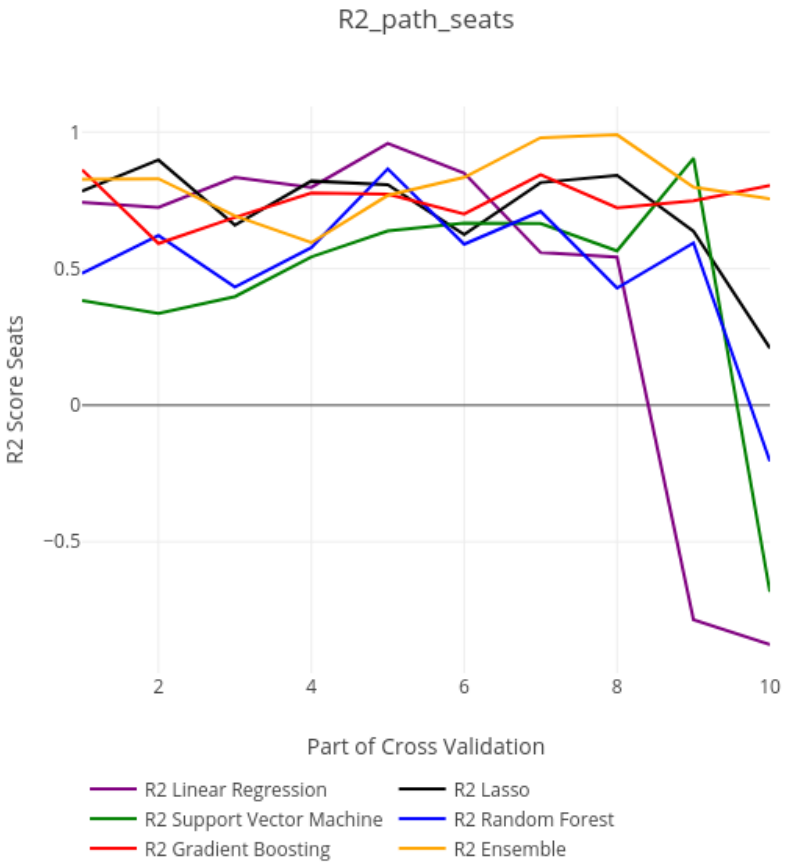
[12]Schwenck (2019)

Table 9: Party Level Evaluation

Estimator	RMSE	R^2
Linear Regression	1.974	0.4345
	(1.04)	(0.644)
Lasso Regression	1.670	0.7097
	(0.689)	(0.1889)
Support Vector Machine Regression	2.2048	0.4414
	(0.8579)	(0.4074)
Random Forest Regression	2.1609	0.5094
	(0.82)	(0.2687)
Gradient Boosting Regression	1.6009	0.7509
	(0.5997)	(0.075)
Ensemble Regression	1.349	0.80
	(0.77)	(0.11)

Averages reported, standard errors in parentheses
All scores obtained by 10-fold cross validation
RMSE (Root Mean Squared Error) is defined as the square root of the negative error

Figure 6: Scores estimators seats



variable that relies on media analysis, such that it calculates the minutes of a the federal TV spent on each parties up to some point prior to the election.

Overall, Figure 8 paints the same picture with the only notable difference being some overestimation

Figure 7: Ensemble estimator seat prediction

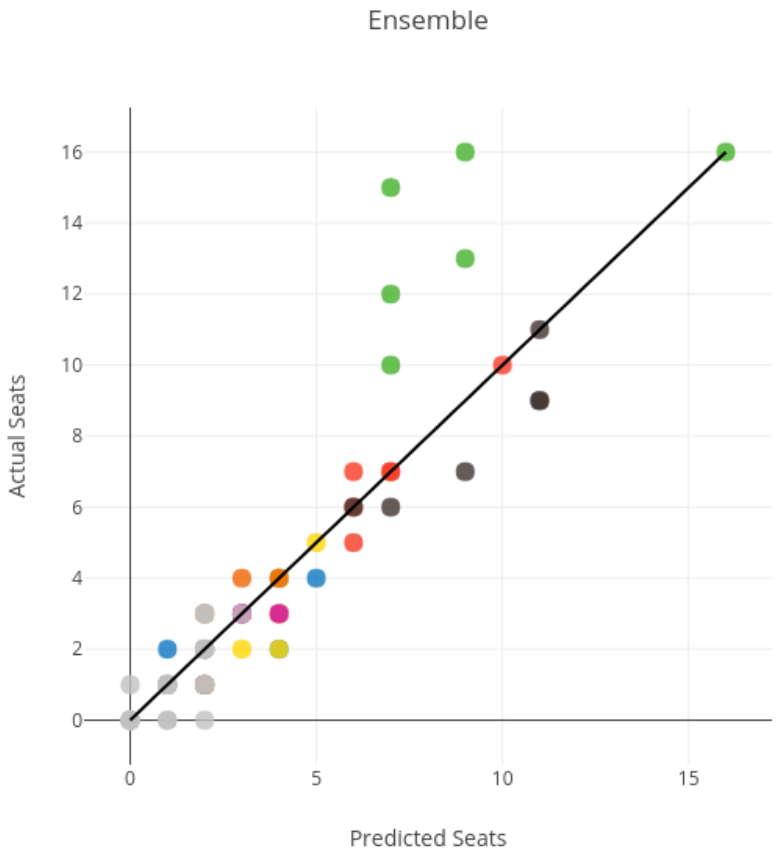


Figure 7 shows a scatterplot of true and predicted seats. The green colour represents *Die Grünen*, black the *CDU*. Red represents the *SPD*, purple *Die Linke* and blue the *AfD* while the *FDP* is yellow and *FW* are orange. All remaining parties are grey. A mixture of colours represents several parties on the same coordinates.

of the conservatives (*CDU*). Further, there is a blue outlier meaning that the right-wing populists (*AfD*) have been overestimated in one occasion.

In contrast to this, a look at the less-nuanced SVM regressor is valuable. Figure 9 shows a more dispersed pictured.

Not only is there a big discrepancy for *Die Grüne* but also for the *CDU* and the red points, the social democrats (*SPD*) but some *AfD* estimates are quite off the 45-degree line, too. However, it shows that *Die Grüne* are not consistently underestimated any more. A possible explanation is that there is a combination of parameters that is able to account for the recent hype that has been dropped by the regularization of the other estimators. Yet, it appears that the price a correcting for a high share of votes for the green party is to constantly underestimate the *SPD* and at the same time overestimate the conservatives as well as the right-wing populists. Given the scores earlier, the trade-off is not beneficial for the accuracy of the estimator. Therefore, new data is likelier to solve this problem.

The results of the other estimators can be forwarded by the author.

Overall, the ensemble regressor consisting of two parts gradient boosting, two parts random forest and

Figure 8: Gradient boosting seat prediction

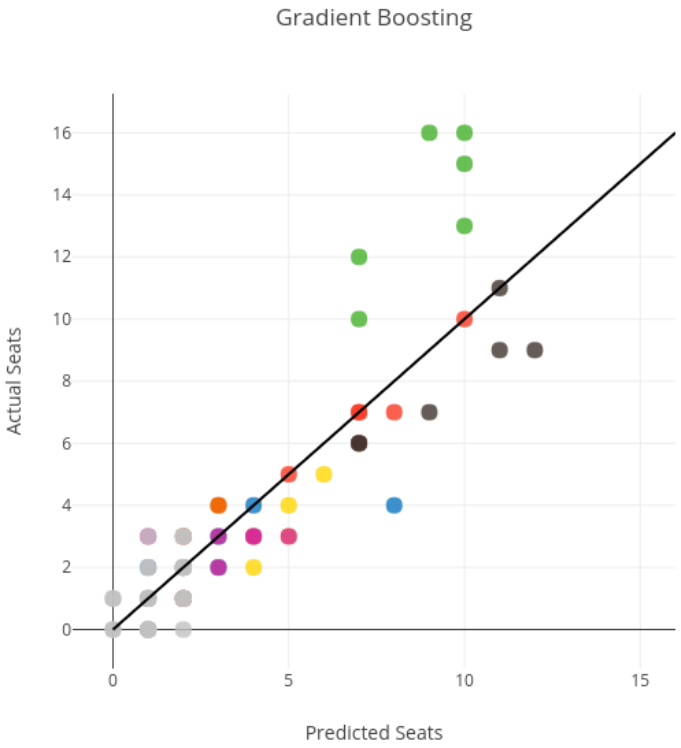


Figure 8 shows a scatterplot of true and predicted seats. The green colour represents *Die Grünen*, black the *CDU*. Red represents the *SPD*, purple *Die Linke* and blue the *AfD* while the *FDP* is yellow and *FW* are orange. All remaining parties are grey. A mixture of colours represents several parties on the same coordinates.

one part lasso seems to be the most promising estimator but with the simple gradient boosting almost being on par. Other estimators such as the lasso or the random forest provide decent prediction that do not reach the accuracy and precision of the former two. The SVM and the linear regression fail to predict the results credibly with too much variation in the quality of the forecast depending on the sample split. Even besides the high standard errors, the average accuracy is not great either.

Regarding the measure we newly introduced, Google search results, to capture social connectedness, the usefulness for estimators is mixed, as Table 10 demonstrates. While we cannot take the values obtained from either regression literally, they are some first indicators whether or not there is any relevance to them.

Any OLS might suffer from omitted variable or reverse causality bias, to give just two examples of possible biases. Both of them have not been addressed since this it would exceed the scope of this analysis. Feature importances of the random forest might be biased, too. This is due to scikit default estimation of feature importances which may well be problematic (Strobl et al. (2007)).

It has been dropped by the lasso estimator, as have been all linear combinations and the polynomial.[13]. The picture is different for the random forest where it resides as the 10th most important feature, likely surviving permutation estimates. Similar results can be found for the gradient boosting

[13]Results can be received from the author

Figure 9: SVM seat prediction

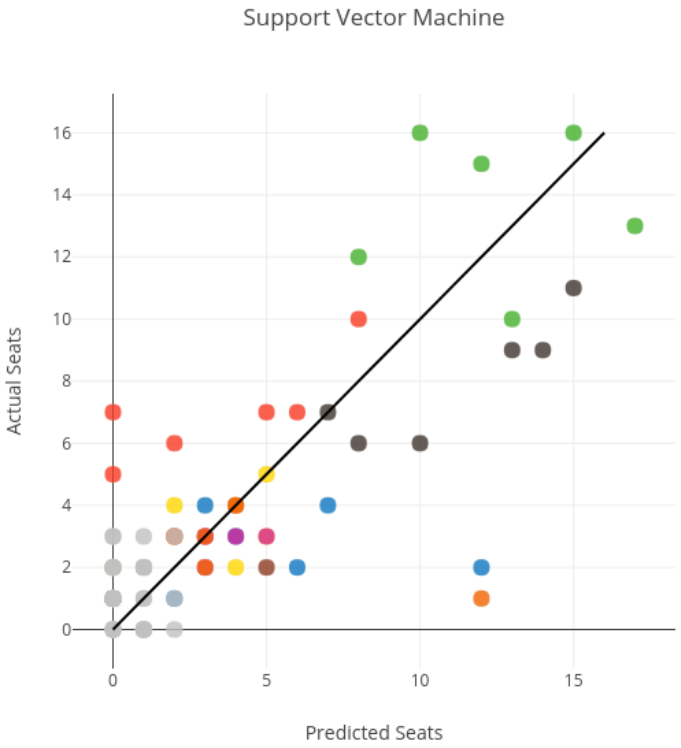


Figure 9 shows a scatterplot of true and predicted seats. The green colour represents *Die Grünen*, black the *CDU*. Red represents the *SPD*, purple *Die Linke* and blue the *AfD* while the *FDP* is yellow and *FW* are orange. All remaining parties are grey. A mixture of colours represents several parties on the same coordinates.

regression. Here, the feature as well as many interaction terms are included and even in the upper half in terms of magnitude. In the case of the linear regression, the feature exhibits weak statistical significance at the 10% level. As being said before, this should not be taken literally but rather as an indicator.

In total, the hypothesis of relevance when predicting future local election outcomes seems to be supported by more evidence than not.

Table 10: Google Search Result Coefficients

Estimator	Value
Lasso	0
SVR	-891
Random Forest	0.013
Gradient Boosting	0.0007
Linear Regression	-166

6 Conclusion

This paper aims at offering a start for local election predictions. The main rationale for doing so, is the lack of frequent local election predictions that quantify results. Further, the lack of consistent polling on the municipal level adds to the shortage of accurate prediction and even if there was polling across cities - the electoral system in many states of Germany will cast doubt on the polls collected.

In order to achieve credible, quantifying predictions, I collect decentralized election results from several cities, demographic and socioeconomic statistics as well as political trend data. Further measures on the candidate level are constructed. Lastly, I propose to measure social interconnectedness with Google search results. This appears to be a justifiable predictor for most estimators.

Overall, the training data covers the state of Baden-Württemberg for the years 2009 and 2014. The test data are the recent election results of 2019. All data is being standardized, linear combinations are being calculated and polynomials constructed.

When forming the predictions, this paper considers several regressors. The most valuable proves to be a simple gradient boosting as well as a blend of two parts gradient boosting, two parts random forest and one part lasso. Most of the times, both estimator reach R^2 scores between 0.6 and 0.8 when predicting the amount of votes a candidate receives.

Many forecasts care in addition to that about the number of seats a party gets in the local council. Subsequently, I aggregate the votes and calculate the seats for each party in each city according to the procedure by Webster/Sainte-Lague.

After cross validation of the results on the seat level, the ensemble and gradient boosting estimator are the best predictors. They achieve average R^2 scores of 0.75-0.8 with little standard deviation (around 0.1).

Yet, there is plenty of room for improvement in future research. In terms of data, I lack data on income as well as data on recent trends of parties. Web-scraped house prices and web-scraped media meta-data might provide proxies to this problem. Further, the weights of the ensemble could be improved by expanding the weight-matrix to higher numbers for individual estimators. Gradient boosting itself could be improved by using extreme gradient boosting. Additionally, neural networks might be able help proceed this project and may even be included in the ensemble estimation.

References

Baden-Württemberg, S. L. (2019): "Regionaldaten," https://www.statistik-bw.de/SRDB/?R=GS117019, [Online; accessed 30-July-2019].

Cantow, M., M. Fehndrich, A. Schneider, and W. Zicht (2019): "Wahlrecht.de," https://www.wahlrecht.de/, [Online; accessed 30-July-2019].

Cortes, C. and V. Vapnik (1995): "Support-vector networks," *Machine learning*, 20, 273–297.

Friedhoff, C., L. Holtkamp, and E. Wiechmann (2015): "Frau Doktor steht zur Wahl. Eine quantitative Analyse des bundesdeutschen Wahlverhaltens auf lokaler Ebene aus der Genderperspektive," *GENDER–Zeitschrift für Geschlecht, Kultur und Gesellschaft*, 8.

García-Pedrajas, N., C. Hervás-Martínez, and D. Ortiz-Boyer (2005): "Cooperative coevolution of artificial neural network ensembles for pattern classification," *IEEE Transactions on evolutionary computation*, 9, 271–302.

Gonzalez, N. L. (2019): "Inside Elections - Nonpartisan Analysis," http://insideelections.com/, [Online; accessed 21-July-2019].

Greer, J. and D. Mensing (2006): "The evolution of online newspapers: A longitudinal content analysis, 1997-2003," *Internet newspapers: The making of a mainstream medium*, 13–32.

Hastie, T., R. Tibshirani, and J. Friedman (2009): "Boosting and additive trees," in *The elements of statistical learning*, Springer, 337–387.

Hastie, T., R. Tibshirani, J. Friedman, and J. Franklin (2005): "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, 27, 83–85.

Keil, S. and O. Gabriel (2012): "The Baden-Wurttemberg State Election of 2011: A Political Landslide," *German Politics*, 21, 239–246.

Krueger, C. C. and P. M. Swatman (2004): "Developing e-business models in practice: the case of the regional online newspaper," *International Journal of Information Technology and Management*, 3, 157–172.

Leiner, B. M., V. G. Cerf, D. D. Clark, R. E. Kahn, L. Kleinrock, D. C. Lynch, J. Postel, L. G. Roberts, and S. Wolff (2009): "A brief history of the Internet," *ACM SIGCOMM Computer Communication Review*, 39, 22–31.

Louppe, G. (2014): "Understanding Random Forests: From Theory to Practice," Ph.D. thesis, University of Liege, Belgium, arXiv:1407.7502.

Mellon, J. and C. Prosser (2017): "Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users," *Research & Politics*, 4, 2053168017720008.

Quecke, A., I. Bock, and H. Königsberg (2019): *Das Kommunalwahlrecht in Baden-Württemberg*, Kohlhammer Verlag.

Radio, D. (2019): "Kommunalvalg 2017," https://www.dr.dk/nyheder/politik/kv17/resultater/, [Online; accessed 21-July-2019].

Richardson, J. E. and B. Franklin (2004): "Letters of intent: Election campaigning and orchestrated public debate in local newspapers' letters to the editor," *Political Communication*, 21, 459–478.

Sabato, L. J. (2019): "Sabato's Crystal Ball," http://crystalball.centerforpolitics.org/crystalball/about/, [Online; accessed 21-July-2019].

Schonwalder, K. (2013): "Immigrant representation in Germany's regional states, the puzzle of uneven dynamics," *West European Politics*, 36, 634–651.

Schwenck, V. (2019): "Der Erfolg ist auch ein Risiko," https://www.tagesschau.de/inland/gruene-umfragehoch-101.html, [Online; accessed 21-July-2019].

Silver, N. (2019): "FiveThirtyEight," https://fivethirtyeight.com/, [Online; accessed 21-July-2019].

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007): "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, 8, 25.

Thomsen, S. R. (1998): "Impact of national politics on local elections in Scandinavia," *Scandinavian Political Studies*, 21, 325–345.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Trounstine, J. (2013): "Turnout and incumbency in local elections," *Urban Affairs Review*, 49, 167–189.

Walker, D. A. (2006): "Predicting presidential election results," *Applied Economics*, 38, 483–490.

Whiteley, P. F. (2005): "Forecasting seats from votes in British general elections," *The British Journal of Politics and International Relations*, 7, 165–173.

Wikipedia contributors (2019a): "2019 Spanish general election — Wikipedia, The Free Encyclopedia," https://en.wikipedia.org/w/index.php?title=2019_Spanish_general_election&oldid=906820360, [Online; accessed 21-July-2019].

——— (2019b): "Opinion polling for the 2019 Spanish local elections — Wikipedia, The Free Encyclopedia," https://en.wikipedia.org/w/index.php?title=Opinion_polling_for_the_2019_Spanish_local_elections&oldid=901001046, [Online; accessed 21-July-2019].