Supervised learning with political data

Guillaume COQUERET

This version: 2018-03-09

Abstract The purpose of this article is to present a novel dataset based on the surveys carried out by the American National Election Studies (ANES). The data contains a wide range of fields and can be used as support for Data Science or Machine Learning applications, especially in class. As an illustration of its pedagogical benefits, I propose several classical tasks, such as regression and classification problems. In passing, I compare most widespread methods mentioned in supervised learning courses (regressions, trees, SVMs and neural networks).

Keywords Supervised learning · Political data · Support Vector Machines · Boosted trees · Neural networks

1 Introduction

When designing a Data Science or Machine Learning course, the instructor or professor is faced with the crucial choice of the data it will work with. While some options are obvious (e.g., images for convolution neural networks (CNN)), truly engaging databases are scarce and they are paramount because they will determine the commitment of students inside the classroom. For instance, one classical dataset cited in many tutorials is the Iris table, which is only truly appealing for flower lovers. Two (better, in my opinion) alternatives are the student alcohol consumption dataset and the adult income dataset.[1] Nevertheless, after a few hours of data exploration, I felt there was still room for improvement, even with these two good solutions at my disposal.

The quality of a dataset directly depends on its size in two obvious ways. First, there is the number of occurrences (usually, one occurrence corresponds to one line in a classical matrix formatting) and then there are the number of attributes that describe the occurrences (it is customary to use one column for each attribute). Very often, promising datasets can be frustrating because they lack some depth in one of these dimensions. A size of several hundred occurrences is often too small when it comes to training and testing a model on non-overlapping subsets and cross-validation is even more complicated. At a deeper level, the quality and interpretability of attributes matters too. For the educator, a large cross-section of criteria is much more effective if the audience can relate to it.

In the remainder of the paper, I address these issues through the presentation of a new dataset that I built essentially for the purpose of teaching data science and machine learning. The construction of the data is detailed in Section 2. Section 3 is dedicated to classification exercises while Section 4 is devoted to regression-based techniques. Finally, Section 5 conludes. All of the material (R scripts, heavily based on [8]) are available online.

The techniques listed in the paper are all described in detail in the textbook [5]. On the topic of Deep Learning, and neural networks more generally, I recommend the following references: [3], [6], [2].

2 Assembling the data

I discovered the datasets of the ANES when reading the paper on an estimation technique for political po-

G. Coqueret
E-mail: guillaume.coqueret@gmail.com

[1] These popular sets (and many others) can easily be found on data providers such as Kaggle, data.world or the UCI Machine Learning Repository.

larization [7]. Immediately, I understood their potential because their attributes are easy to interpret and connect to. Moreover, the full time-series file encompasses the years 1948 to 2012 and has more than 55 thousand lines and almost one thousand columns. The data is public and updated every two to four years.

Since 1948, the ANES has conducted surveys every presidential election year and often during midterm years as well. The institution asks respondents a long series of questions about themselves and how they feel on a large palette of topics. This produces an incredibly detailed series of snapshots of the US population - though it can be argued whether or not these snapshots are indeed representative of the US population as a whole.

I underline that the questions asked in the surveys may vary from one year to another, which makes data consistency complicated. Choosing a limited number of variables out of 950 is a difficult task. Obviously, out of the 950, some are redundant and some carry information of little value.[2] In order to filter out unnecessary columns, I used two criteria: the overall pertinence of the field and its availability in the cross-section of dates. A variable that is undefined or missing for most occurrences is not very useful.

The attributes of the main dataset (1948-2012) are listed in Tables 3 and 4 in the Appendix. Not all attributes are available at all dates, so I summarized their presence in recent years in Table 6. For the most recent database (2016), attributes were different because the ANES decided to change some questions and categories.[3] Also, there were two types of questionnaires: face-to-face and online, which creates some discrepancies between the answers of the respondents. The interested reader can have a closer look at the corresponding codebooks and the variable matching is provided in Table 5.

In Figure 1, I plot the distribution of demographic and feeling variables. The data consists of all surveys subsequent to 1986 (included).

I list a few comments below. The age of respondent peaks between 25 and 35 years. The military is the topic which is the most skewed to the right, hence for which the respondents have the highest average esteem. There are slightly more female respondents. Respondents own their home, by a factor 2-to-1 and a majority of them are married. An overwhelming proportion of respondents are white and 70% were actively working at the time of they answered the survey.

3 Classification tasks

A very tempting task given the dataset is to seek to predict party affiliation based on some other variables. Given the availability of the fields, the Party_simple variable seems adequate and leads to a ternary classification: Democrat, Independent or Republican. In the following sections, I investigate some issue related to diverse classification methods.

The explanatory variables are the following:

- Age,
- Gender,
- Education,
- Church_attendance,
- Income and
- Feeling_unions.

I restrict their number for two reasons. First for the sake of simplicity and second because I filter out non-ordered categorical variables. The justification for that lies in the requirement for some algorithms to work with fully numerical data (e.g., boosted trees and artificial neural networks). Hence, I keep ordered variables and digitize them into integer number when need be.[4]

As is customary in Machine Learning, I split the data in two. The training set consists of answers to the surveys that took place between 1986 and 2008. The testing set encompasses the 2012 and 2016 survey campaigns. Note that this chronological splitting implicitly assumes that past preferences or patterns hold in the future, which may not be true.[5] In order to avoid problems related to absent data, I filter out the occurrences for which one or more feature is missing. The final training set has 16,662 lines and the testing set, 7,482.

I evaluate the performance of the models using the simple accuracy metric which captures the proportion of correct answers predicted over the testing sample. Obviously, other indicators[6] could complement the analysis, but I restrict the study to only one of them, for the sake of clarity and simplicity.

[2] The codebook for the 1948-2012 dataset is 456 pages long and not very entertaining to read. Many questions pertain to the feeling of the respondent towards past and current political figures, including local House and Senate candidates. The relevance of these particular items is at best of second order.

[3] As of February 2018, the 2016 survey had not yet been aggregated with the master Time Series Studies file.

[4] Non-ordered categorical variables are usually processed through one-hot encoding but I try to avoid this procedure in this paper.

[5] Partisan beliefs and attitudes can change through time. To cite only one example, [1] shows that Democrats' and Republicans' stance on abortion has shifted between 1972 and 1994.

[6] Precision, recall and Area Under the ROC curve, to names a few. Among these indicators, accuracy is probably the easiest to interpret.

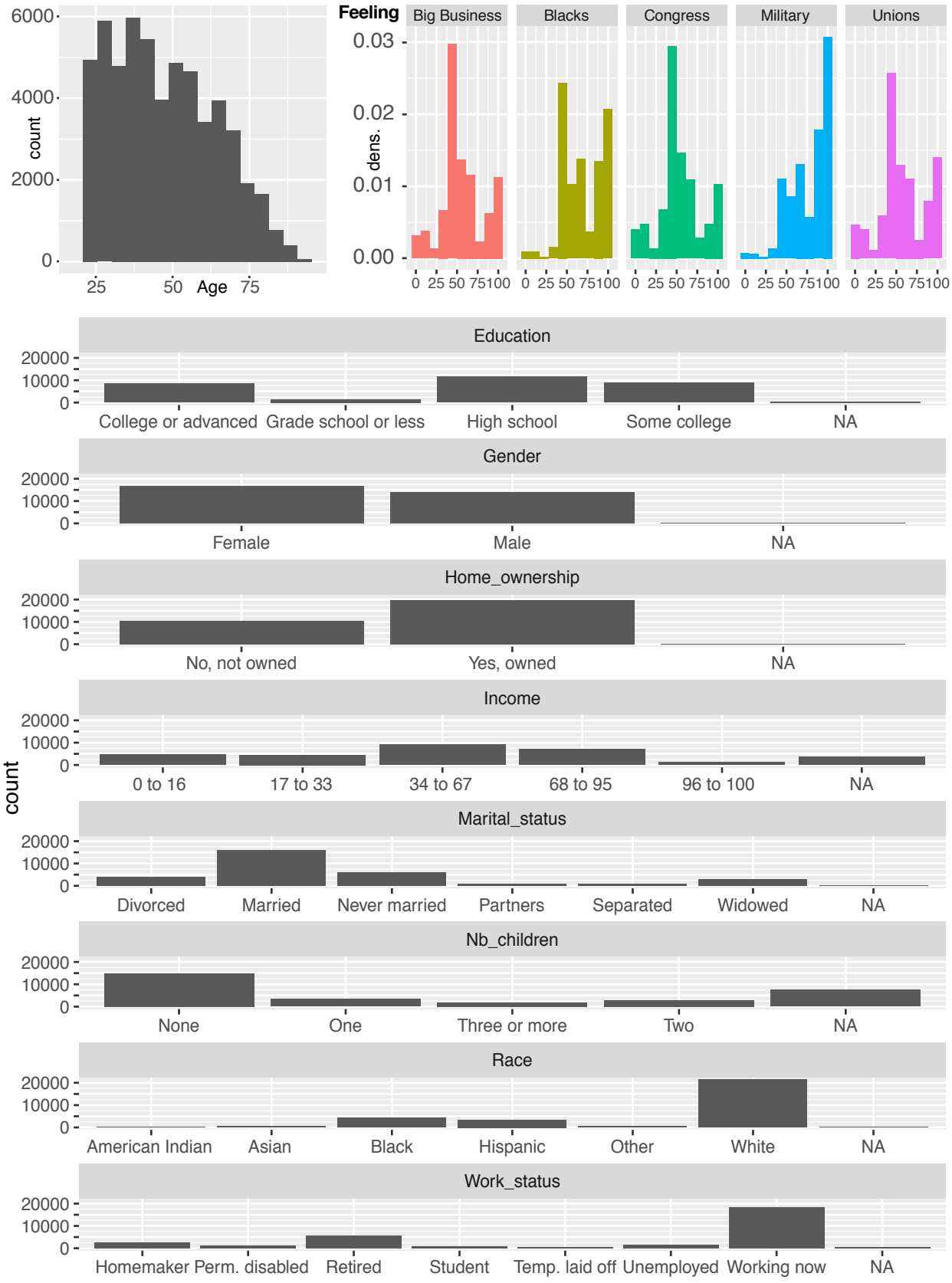

Fig. 1: **A glance at the data:** demography and feelings (1986-2016).

3.1 Simple trees

Beyond linear models (see Section 4.1 below), regression trees are another building block of machine learning tools. I refer to Section 9.2 in [5] for self-contained introduction on the subject. Given the large number of (unordered) categorical variables in the dataset, classification trees are probably the most obvious class of models to begin with. In Figure 11 in the Appendix, I plot the tree with Age, Gender, Race, Education, Income and Church_attendance as explanatory variables. The diagram shows that the feeling towards union and church attendance are the two most important variables in the determination of political identification. Typically, citizens with higher esteem for unions are, on average, more likely to be Democrats. Among Americans with inferior view on unions (feeling lower than 50), those who attend church regularly tend to feel closer to the Republican party.

The number of leaves of a decision tree is obviously an important parameter in the model. To test the sensitivity of the accuracy of the model, I tested a large number of possible values for the complexity of the tree. I summarize the results in Figure 2. First of all, there is a clear qualitative jump between the agnostic option[7] and simply two leaves. This means that the first splitting variable (likely the feeling towards unions) is valuable: the accuracy increases from 38% to 48%, which is significant. The best accuracy (49%) is obtained for 8 leaves. Beyond that, the model starts to overfit the training data and ultimately decreases to 45%.

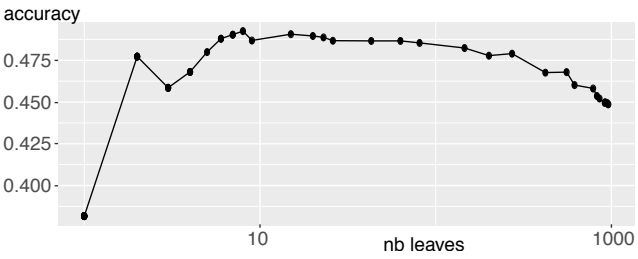

Fig. 2: **Accuracy of classification trees: impact of the number of leaves**. The dependent variable is Party_simple and the explanatory variables are Age, Gender, Education, Income, Church_attendance and Feeling_unions. The training set corresponds to the years between 1986 and 2008 and the test is performed on the occurrences of the 2012 and 2016 surveys. In the data, rows with missing values are discarded upfront. The x-axis has a logarithmic scale for the number of leaves in the trees and eta is the learning rate.

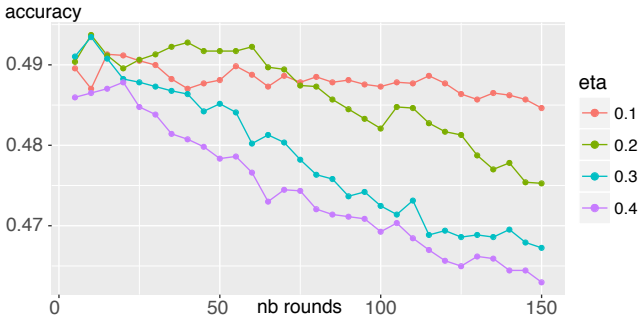

Fig. 3: **Accuracy of boosted trees: impact of the number of trees and learning rate**. The dependent variable is Party_simple and the explanatory variables are Age, Gender, Education, Income, Church_attendance and Feeling_unions. The training set corresponds to the years between 1985 and 2010 and the test is performed on the occurrences of the 2012 and 2016 surveys. In the data, rows with missing values are discarded upfront. The x-axis represents the number of trees used by the algorithm. Eta corresponds to the learning rate. The maximum depth of each layer (individual tree) is fixed to six.

3.2 Boosted trees

One of the most popular refinements of trees consists of successively aggregating (possibly optimized) trees (see Chapter 10 in [5]). Two of the most important parameters in boosted trees are the number of trees that are combined by the algorithm and the learning (or shrinkage) rate which determines the scale assigned to each new tree added to the model. In Figure 3 below, I show their impact on the accuracy of the prediction on the test set.

The curves confirm the well known fact that smaller learning rates usually perform better (e.g., Table 1 in [4]). Beyond 70 rounds of boosting, all configurations see their accuracy decrease. Moreover, overfitting occurs very early for the highest learning rates: their accuracies start declining before 30 rounds. Among all the combinations I tested, none surpasses 49.5% of correct predictions, which implies that boosting improves classical trees only marginally (the latter often reached accuracies in the 48%-49% range).

Finally, as for the basic decision trees, it is possible to point out which variables are the most impactful for the determination of party identification. Below, in Figure 4, I plot the relative contribution of each feature to one model of boosted tree for which there are 20 rounds and a shrinkage intensity of 0.1.

Plainly, the feeling towards unions is the most determinant variable. In unreported tests, I obtained very

[7] With only one node, the accuracy equals the proportion of Democrats in the testing sample because the prediction is always towards the Democratic class. Fortunately, this is not the worst choice because the proportion of Democrats in both samples is close: 37% (train) and 38% (test).

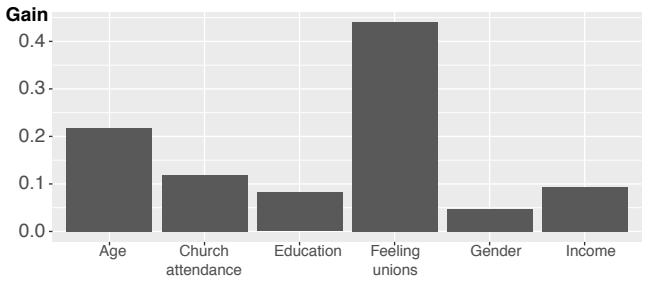

Fig. 4: **Boosted trees: importance of variables**. The dependent variable is Party_simple and the explanatory variables are Age, Gender, Education, Income, Church_attendance and Feeling_unions. The training set corresponds to the years between 1986 and 2008. In the data, rows with missing values are discarded upfront. The number of rounds of boosting is equal to 20 and the learning rate is set to 0.1. The maximum depth of each layer (individual tree) is fixed to six.

similar contributions when changing both the number of boosting rounds and the learning rate.

3.3 Support Vector Machines

I now turn to another popular family in the realm of Machine Learning tools, namely Support Vector Machines (SVMs). There are many options when resorting to SVMs and I cannot present all of them here. I will focus on two features: the regularizing term (cost), which tunes the penalization of the soft margin, and the kernel used to tranform the representation of the data. In Figure 5, I show the sensitivity of accuracy to these two key inputs.

Plainly, the cost regularization has no impact when the kernel is linear. Both radial and sigmoid kernels seem to improve when the cost increases, until it reaches a near unit value; then they decrease. The polynomial kernel has a somewhat erratic behavior with respect to the penalization. Fundamentally, these batches are disappointing because the maximum value (48.3%) does not seem to enhance the quality of prediction compared to the tree-based methods. To be fair, all kernels except the linear one rely on additional parameters that can be adjusted to improve the quality of fit.[8] An in-depth study of such fine-tuning extensions is beyond the scope of the present paper. Lastly, it is common practice to process categories through one-hot encoding before supplying the data to the algorithm. It is possible that this would further increase the accuracy.

[8] Also, in the e1071 package, it is possible to switch between two different classification and regression optimizations.

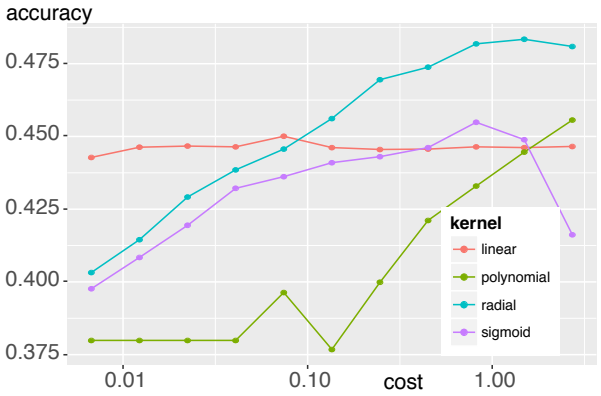

Fig. 5: **Support Vector Machines**. The dependent variable is Party_simple and the explanatory variables are Age, Gender, Education, Income, Church_attendance and Feeling_unions. The training set corresponds to the years between 1986 and 2008. In the data, rows with missing values are discarded upfront. The kernel specifications are the following: linear $(u'v)$, polynomial $((\gamma u'v)^3,)$ radial $(e^{-\gamma|u-v|^2})$ and sigmoid $(\tanh(\gamma u'v))$, with $\gamma = 1/6$. The x-axis has a log-scale.

3.4 Artificial Neural Networks

The final approach that I test on the sample is the simplest version of artificial neural networks (ANNs): the multilayer perceptron. The number of degrees of freedom in the construction of neural network models is incredibly large, and I stick here to the most basic configuration. I trained a network with two intermediate fully connected layers with 32 and 16 units with, respectively, the ReLU (rectified linear unit) and sigmoid activation functions. The categories were processed through one-hot encoding prior to running the back-propagation. Arguably, one essential exercise is to monitor the effect of successive rounds of back-propagation. The accuracy as a function of the number of epochs is shown below in Figure 6.

This graph illustrates the typical pattern of neural networks: an improvement in the first rounds of back-propagation followed by overfitting (and diminishing returns). A surprising feature is that the model performs better on the testing sample than on the training data. Note that none of the 30 epochs yields an accuracy above 50%, which appears as a ceiling I was not able to break. I have tested many other configurations: more layers, more units, other activation functions and optimizers, different batch sizes, etc. Nevertheless, I have not been able to (ever) reach 50% of correct predictions. It is unclear to me if further fine-tuning of hyperparameters or extensions such as recurrent or adversarial networks can generate accuracies that a substantially above 50%. This may require to increase the num-

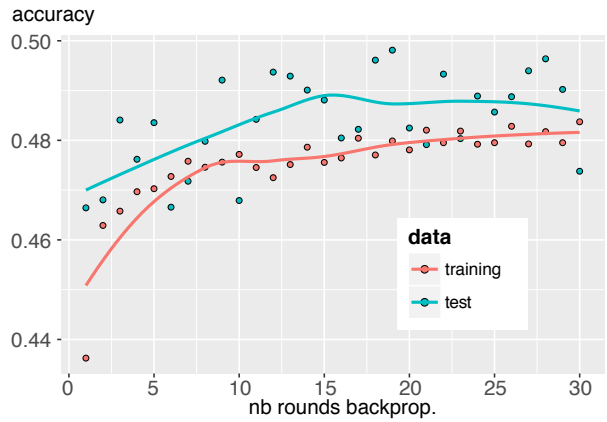

Fig. 6: **Multi-layer perceptron**. The dependent variable is Party_simple and the explanatory variables are Age, Gender, Education, Income, Church_attendance and Feeling_unions. The training set corresponds to the years between 1986 and 2008. In the data, rows with missing values are discarded upfront. The batch size is 16, the loss function is categorical cross-entropy and the optimization is performed using the root-mean square propagation algorithm.

ber of explanatory variables because the six used in the exercise are probably not nearly enough.

4 Regression tasks

In this section, I aim at explaining and predicting numerical variables. A popular choice is income (as in Section 9.2 of [4]), but even though I could digitize the five categories of the corresponding item, an appealing alternative stands out: the feeling towards unions. The choice of explanatory variables is driven by two criteria: availability though time and the requirement that the variable be numerical or ordered. The final list is the following:

- Age,
- Gender,
- Education,
- Income,
- Church_attendance,
- Home_ownership,
- Abortion,
- Foreign_policy, and
- Government_economy

In order to assess the goodness-of-fit of the models, I will resort to the Mean Absolute Error (MAE), computed as

$$\text{MAE}(\boldsymbol{f}, \hat{\boldsymbol{f}}) = \frac{1}{T}\sum_{t=1}^{T} |f_t - \hat{f}_t|,$$

where $\boldsymbol{f} = f_t$ is the collection of the values of Feeling_unions in the test sample and and $\hat{\boldsymbol{f}} = \hat{f}_t$ aggregates the predictions stemming from the model fitted on the training sample. T is the number of occurrences in the test sample. The MAE is straightforwardly the average (absolute) deviation on the Feeling_unions score implied by the model. Again, I filter missing values upfront and the resulting training sample has 9,816 lines while the testing set has 6,839 occurrences.

4.1 Linear regression

In Table 1, I provide the summary of the linear regression over the whole sample from 1986 to 2016. All variables appear to significantly impact the feeling towards unions. According to the estimates, this feeling is the highest when combining the following characteristics: young, women, less educated, lower income, religious, with a positive (above average) feeling towards the black population.

If I then estimate the coefficients on the sample from 1986 to 2008 and compute the MAE on the test sample consisting of the surveys of 2012 and 2016, I get an average error of 21.2, which seems rather large.[9] The purpose of the subsequent section will be to improve this figure.

4.2 Regression trees

As in the previous section, I start by growing a simple tree on the whole sample from 1986 to 2016. The corresponding diagram in presented in Figure 12 in the Appendix. In the determination of the feeling towards unions, the feeling towards the role of government in the economy is the primary variable and education as a secondary adjustment.

As with the classification tree, I investigate the impact of the depth of the tree on the performance metric (MAE). The sensitivity to the number of leaves is detailed in Figure 7. Similarly to the classification trees, the regression trees start to overfit quite early, beyond 15 leaves. The overall minimum is 21.2, which means that simple regression trees do not improve on the linear regression.

The question is now whether or not boosted trees can do better. I replicate the protocol of classification trees and produce Figure 8 below. Again, the minimum MAE is equal to 19.9, which means that there is no further reduction of MAE when switching to boosted

[9] Using the training sample average as unique prediction gives an MAE of 22.4.

	Estimate	Std. Error	t-stat.	Pr($>\|t\|$)	Signif.
(Intercept)	84.94495	2.27960	37.263	< 2e-16	***
Age	-0.02701	0.01551	-1.742	0.0816	.
Gender	2.12366	0.50085	4.240	2.25e-05	***
Education	-4.07718	0.30674	-13.292	< 2e-16	***
Income	-2.19488	0.25863	-8.487	< 2e-16	***
Church_attendance	-0.02960	0.16862	-0.176	0.8607	
Home_ownership	0.82113	0.58431	1.405	0.1600	
Abortion	0.20731	0.25060	0.827	0.4081	
Foreign_policy	-0.38553	0.57885	-0.666	0.5054	
Government_economy	-2.86120	0.13734	-20.833	< 2e-16	***

Table 1: **Linear regression**: The dependent variable is Feeling_unions and the explanatory variables are Age, Gender, Education, Income, Church_attendance, Home_ownership, Abortion, Foreign_policy and Government_economy. The tree is built on all data from 1986 to 2016. The level of significance of the p-values is classified as follows: $0 < *** < 0.001 < ** < 0.01 < * < 0.05 < . < 0.1$. The Adj. R^2 is 0.106.

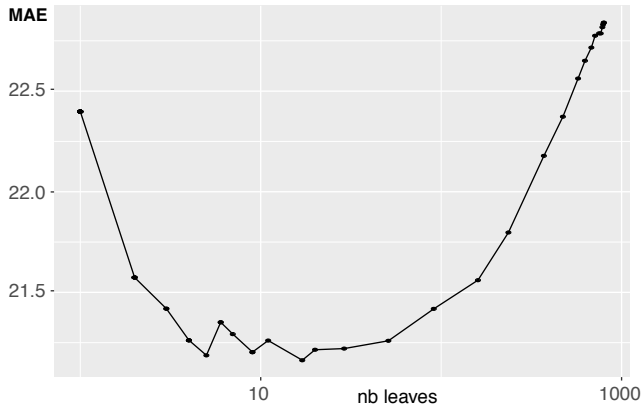

Fig. 7: **MAE of regression trees: impact of the number of leaves**. The dependent variable is Feeling_unions and the explanatory variables are Age, Gender, Education, Income, Church_attendance, Home_ownership, Abortion, Foreign_policy and Government_economy. The training set corresponds to the years between 1986 and 2008 and the test is performed on the occurrences of the 2012 and 2016 surveys. In the data, rows with missing values are discarded upfront. The x-axis has a logarithmic scale for the number of leaves in the trees and eta is the learning rate.

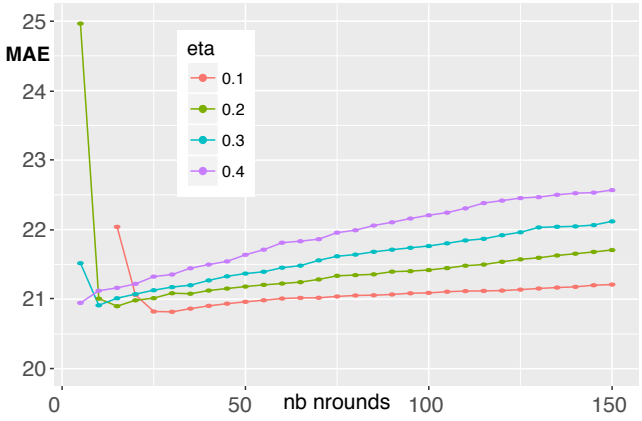

Fig. 8: **MAE of boosted trees: impact of the number of trees and learning rate**. The dependent variable is Feeling_unions and the explanatory variables are Age, Gender, Education, Income, Church_attendance, Home_ownership, Abortion, Foreign_policy and Government_economy. The training set corresponds to the years between 1985 and 2010 and the test is performed on the occurrences of the 2012 and 2016 surveys. In the data, rows with missing values are discarded upfront. The x-axis represents the number of trees used by the algorithm. Eta corresponds to the learning rate (the first values for $\eta = 0.1$ are cut from the graph to ease readability). The maximum depth of each layer (individual tree) is fixed to six.

alternatives. In unreported trials, I tested lower values of η (e.g. 0.01) combined to a higher number of rounds (300), but it did not change the minimum floor for the MAE.

In terms of relative importance, there is a clear dispersion across variables: the Government_economy (43%) variable, along with Education (20%), Age (14%) and Income (10%) are those that stand out. These order of magnitudes are very consistent with the graphical representation of Figure 12. This implies that views on foreign policy for instance have little predictive power over views on unions - according to the dataset.

4.3 SVM

To illustrate the use of SVM, I proceed differently than in the previous section. Here, I fix the kernel to the radial alternative and I plotted the sensitivity of the MAE to its unique parameter γ. In Figure 9, I show this impact for different values of the penalization of the model.

For this choice of (radial) kernel, the results are not as promising as expected: the best error I could reach is 20.7%, which is above those obtained with trees. Again, it is likely that changing the kernel or other parameters

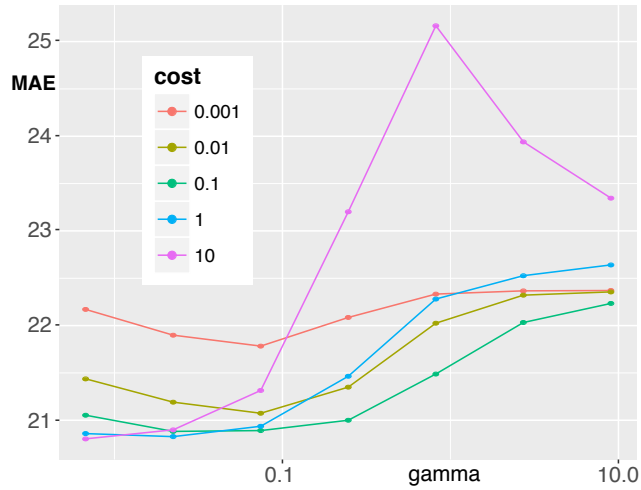

Fig. 9: **Support Vector Machines**. The dependent variable is Feeling_unions and the explanatory variables are Age, Gender, Education, Income, Church_attendance, Home_ownership, Abortion, Foreign_policy and Government_economy. The training set corresponds to the years between 1986 and 2008. In the data, rows with missing values are discarded upfront. The kernel is radial ($e^{-\gamma|u-v|^2}$). The x-axis has a logarithmic scale.

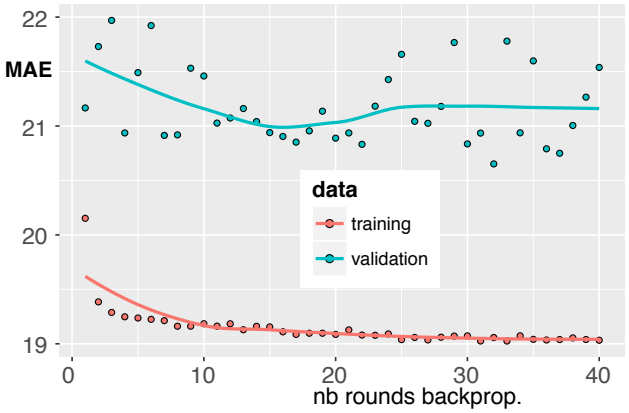

Fig. 10: **Multi-layer perceptron**. The dependent variable is Feeling_unions and the explanatory variables are Age, Gender, Education, Income, Church_attendance, Home_ownership, Abortion, Foreign_policy and Government_economy. The training set corresponds to the years between 1986 and 2008. In the data, rows with missing values are discarded upfront. The batch size is 4, the loss function is MAE and the optimization is performed using the root-mean square propagation algorithm.

and resort to one-hot encoding of categorical variables would improve the results. Nevertheless, whether these gains can be substantial remains uncertain.

4.4 ANN

Finally, I train a neural network in the same configuration as the one in Section 3.4: two intermediate fully connected layers with 32 and 16 units with, respectively, the ReLU (rectified linear unit) and sigmoid activation functions. The corresponding plot is located in Figure 10 below.

The learning phase lasts between the first and twenty-fifth round of back-propagation. The minimum MAE is 20.6%, which does not improve on the previous results.

5 Conclusion

The purpose of the present paper is to introduce a new dataset based on survey results released by the ANES. Thanks to the numerous features of the database, it is both easy and intuitive to present and illustrate data mining techniques that rely on this database. I provide a few examples and I summarize the findings in Table 2 below.

The most impressive results are obtained for the classification task. This is where neural networks perform best, even though the incremental improvement compared to simple trees is not sizable. With regard to regression analysis, the enhanced methods provide only marginal gains on the MAE. The best results are obtained from regression trees.

Model	Accuracy. (class.)	MAE (reg.)
No model	0.383	22.2
Lin. regression	-	21.2
Simple trees	0.488	21.2
Boosted trees	0.494	20.8
SVMs	0.483	20.8
Neural networks	0.498	20.7

Table 2: **Summary of best results**.

Overall, the examples I illustrated are just the tip of the iceberg of the capabilities of the dataset. Typically, I only used a subset of all variables while the number of permutations for dependent and explanatory variables is very large. Furthermore, the database's potential for datavisualization is almost limitless and its rich mixture of numerical, ordered categorical and unordered categorical variables will always make it a fertile ground for applications in artificial intelligence.

6 Disclaimer

The original collector of the data, ANES, and the authors bear no responsibility for use of the data or for interpretations or inferences based upon such uses.

References

1. Adams, G.D.: Abortion: Evidence of an issue evolution. American Journal of Political Science pp. 718–737 (1997)
2. Chollet, F.: Deep Learning with Python. Manning (2018)
3. Du, K.L., Swamy, M.N.: Neural networks and statistical learning. Springer (2013)
4. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
5. Friedman, J.H., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer (2009)
6. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning. MIT Press (2016)
7. Krasa, S., Polborn, M.: Policy divergence and voter polarization in a structural model of elections. Journal of Law and Economics **57**(1), 31–76 (2014)
8. Wickham, H., Grolemund, G.: R for data science: import, tidy, transform, visualize, and model data. O'Reilly (2016)

A Trees

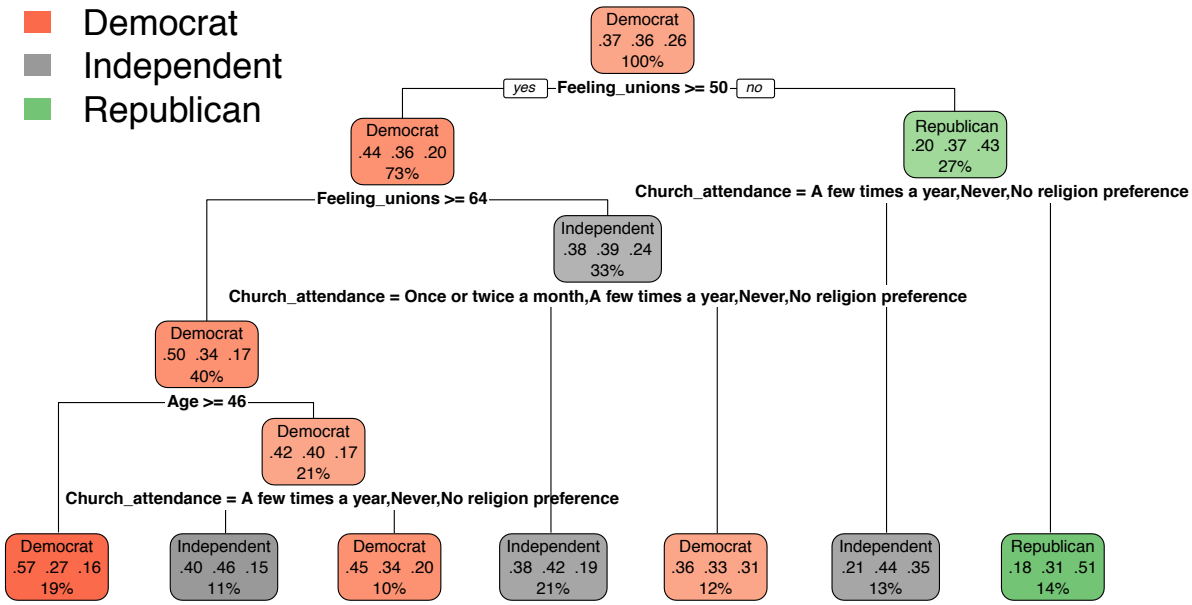

Fig. 11: **Simple classification tree**. The dependent variable is Party_simple and the explanatory variables are Age, Gender, Education, Income, Church_attendance and Feeling_unions. The cp parameter in the rpart function is equal to 0.005. The tree is built on all data from 1986 to 2016.

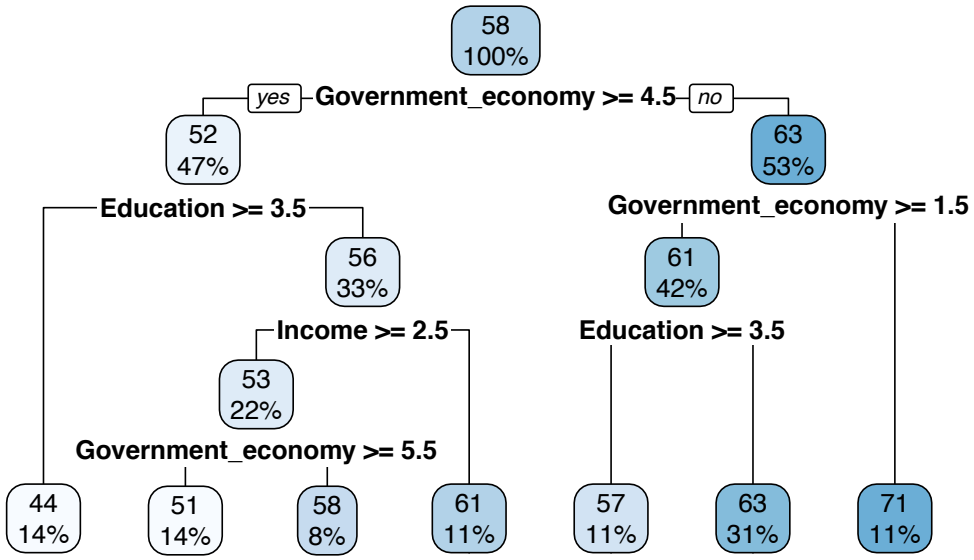

Fig. 12: **Simple regression tree**. The dependent variable is Feeling_unions and the explanatory variables are Age, Gender, Education, Income, Church_attendance, Home_ownership, Abortion, Foreign_policy and Government_economy. The cp argument in the rpart function is equal to 0.003.

B Data description

Name	Topic	Values
Date	Year of the survey	1948 1952 1954 1956 1958 1960 1962 1964 1966 1968 1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2008 2012 2016
Gender	Gender of respondent	Male / Female
Race	Race of respondent	White Black Asian or Pacific Islander American Indian or Alaska Native Hispanic Other or multiple races
Education	Highest degree earned by respondent	Grade school or less High school Some college College or advanced degree
Income	Total annual income for respondent's family (Data processed into discretized distribution)	0 to 16 percentile 17 to 33 percentile 34 to 67 percentile 68 to 95 percentile 96 to 100 percentile
Work_status	Respondent's work status	Working now Temporarily laid off Unemployed Retired Permanently disabled Homemaker Student
Religion	Respondent's major religion group	Protestant Catholic Jewish Other and none
Church_attendance	Frequency at which respondent goes to church	Every week Almost every week Once or twice a month A few times a year Never No religion preference
Religion_important	Is religion important to respondent?	Yes / No
Nb_children	Nb of children below 18 in respondent's household	None One Two Three or more
Home_ownership	Does respondent owns his/her house?	Yes, owned / No, not owned
Marital_status	Situation of respondent regarding his/her spouse	Married Never married Divorced Separated Widowed Partners, not married

Table 3: Items: questions and possible answers.

Name	Topic	Values
Party_identification	Political category that best describes respondent	Strong Democrat Weak Democrat Independent - Democrat Independent - Independent Independent - Republican Weak Republican Strong Republican
Party_simple	Same as above, in short	Democrat / Independent / Republican
Level_info	Respondent's level of information about public affairs	Very high Fairly high Average Fairly low Very low
Abortion	Respondent's view on when abortion should be allowed	Never In case of rape, incest, danger If need clearly established Personal choice
Feeling_blacks	Respondent's feeling towards Blacks	Integer between 0 and 100
Feeling_big_business	Respondent's feeling towards Big Business	Integer between 0 and 100
Felling_unions	Respondent's feeling towards labor unions	Integer between 0 and 100
Feeling_military	Respondent's feeling towards the US military	Integer between 0 and 100
Feeling_congress	Respondent's feeling towards the US Congress	Integer between 0 and 100
Government_economy	Respondent's answer to: 'On a scale from one to seven, should the government see to job and standard of living?'	1. Strongly agree 2. 3. 4. 5. 6. 7. Government should leave each person alone
Government_health	Respondent's view on federal implication towards healthcare	1. Government insurance plan 2. 3. 4. 5. 6. 7. Private insurance plan
Foreign_policy	Respondent's view on US foreign involvement	US stays home / US intervenes
Adoption_homo	Respondent's view on whether gays or lesbians should be allowed to adopt	Yes / No
Nb_immigrants	Respondent's view on whether the number of immigrants permitted to stay in the US should...	Increase Stay the same Decrease
Fed_spending_welfare	Respondent's view on whether federal spending welfare programs should...	Increase Stay the same Decrease
Fed_spending_environment	Respondent's view on whether federal spending environmental programs should...	Increase Stay the same Decrease

Table 4: Items (*Continued*): questions and possible answers.

C Variable codes

Variable	1948-2012 Item	2016 Item
Age	VCF0101	V161267
Gender	VCF0104	V161342
Race	VCF0105a	V161310x
Education	VCF0110	V161270
Income	VCF0114	V161361x
Work_status	VCF0116	V161275x
Religion	VCF0128	V161247a
Church_attendance	VCF0130	V161245
Religion_important	VCF0846	V161241
Nb_children	VCF0138	-
Home_ownership	VCF0146	V161334
Marital_status	VCF0147	V161268
Party_identification	VCF0301	-
Party_simple	VCF0301*	V161157
Level_info	VCF0050b	V168112
Abortion	VCF0838	V161232
Feeling_blacks	VCF0206	V162312
Feeling_big_business	VCF0209	V162100
Felling_unions	VCF0210	V162098
Feeling_military	VCF0213	-
Feeling_congress	VCF0228	V162104
Government_economy	VCF0809	V161189
Government_health	VCF0806	V162193x
Foreign_policy	VCF0823	V161153
Adoption_homo	VCF0878	V161230
Nb_immigrants	VCF0879a	V162157
Fed_spending_welfare	VCF0894	V161209
Fed_spending_environment	VCF9047	V161212

Table 5: Variables and corresponding ANES items. For the Party_simple variable in the 1948-2012 dataset, I split the Party_identification variable into the following classes: Democrat for the first two categories, Republican for the last two and Independent for the remaining three.

D Data availability

	1986	1988	1990	1992	1994	1996	1998	2000	2002	2004	2008	2012	2016
Age	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Gender	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Race	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Education	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Income	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
Work_status	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Religion	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Church_attendance	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Religion_important	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Nb_children	✓	✓	✓	✓			✓			✓	✓	✓	✓
Home_ownership	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Marital_status	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Party_identification	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Party_simple	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Level_info	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Abortion	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
Feeling_blacks	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Feeling_big_business		✓		✓	✓	✓		✓	✓	✓	✓	✓	✓
Felling_unions	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Feeling_military		✓		✓		✓		✓	✓	✓	✓	✓	
Feeling_congress		✓		✓		✓	✓	✓	✓	✓	✓	✓	✓
Government_economy	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
Government_health		✓		✓	✓	✓		✓		✓	✓	✓	✓
Foreign_policy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Adoption_homo				✓				✓		✓	✓	✓	✓
Nb_immigrants				✓	✓	✓	✓	✓		✓	✓	✓	✓
Fed_spending_welfare			✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Fed_spending_environment	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓

Table 6: Items and their availability through time.