Improving Supreme Court Forecasting Using Boosted Decision Trees

Aaron Russell Kaufman[1], Peter Kraft[2] and Maya Sen[3]

[1] PhD Candidate, Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA. Email: aaronkaufman@fas.harvard.edu
[2] PhD Candidate, Department of Computer Science, Stanford University, 353 Serra Mall, Stanford, CA 94305, USA. Email: kraftp@cs.stanford.edu
[3] Associate Professor, John F. Kennedy School of Government, Harvard University, 79 John F. Kennedy Street, Cambridge, MA 02138, USA. Email: maya_sen@hks.harvard.edu

Abstract

Though used frequently in machine learning, boosted decision trees are largely unused in political science, despite many useful properties. We explain how to use one variant of boosted decision trees, AdaBoosted decision trees (ADTs), for social science predictions. We illustrate their use by examining a well-known political prediction problem, predicting U.S. Supreme Court rulings. We find that our ADT approach outperforms existing predictive models. We also provide two additional examples of the approach, one predicting the onset of civil wars and the other predicting county-level vote shares in U.S. presidential elections.

Keywords: statistical analysis of texts, forecasting, Learning

1 Introduction

What predicts U.S. Supreme Court rulings? What predicts whether a country will suffer a civil war? How might we forecast U.S. presidential election outcomes at the local level? These are important questions. For example, dozens of papers and hundreds of journalists have sought to predict Supreme Court rulings (e.g., Ruger *et al.* 2004; Epstein, Landes and Posner 2010; Black *et al.* 2011), which are delivered only after months of closed-door deliberation but nonetheless involve key issues in American politics—including civil rights, voting rights, presidential powers, and national security. In the ten months that the Supreme Court was privately deliberating a prominent same-sex marriage case, for example, thousands of couples married without assurances that the federal government would recognize their marriages.[1]

In this paper, we introduce one tool that, though underused in political science, offers attractive properties for social science prediction problems: AdaBoosted decision trees (ADTs). ADTs capture gains in prediction when there are many variables, most of which add only limited predictive value. We illustrate their utility by predicting Supreme Court rulings using a novel dataset that includes case-level information alongside textual data from oral arguments. Using this approach, we predict more than 75% of all case outcomes accurately, with even higher accuracy among politically important cases. Substantively, we are able to accurately predict approximately seven more cases per year (out of around 80) compared to the baseline of predicting that the petitioner will always win, which yields 68% accuracy. To illustrate the broad applicability of ADTs, we provide two additional examples: (1) predicting whether civil war occurs in a country in a given year (which we predict with 99.0% accuracy) and (2) predicting county-level U.S. presidential election outcomes (which we predict with 96.7% accuracy, using the 2016 election as our example).

Political Analysis (2019)
vol. 27:381–387
DOI: 10.1017/pan.2018.59

Published
19 February 2019

Corresponding author
Aaron Russell Kaufman

Edited by
Jeff Gill

Authors' note: Replication materials available at the *Political Analysis* Dataverse: https://doi.org/10.7910/DVN/JJCXTH (Kaufman, Kraft and Sen 2018)

[1] See Appendix A0 for discussion of the substantive importance of Supreme Court prediction.

2 AdaBoosted Decision Trees and their Applicability to Social Science Questions

With exceptions (e.g., Green and Kern 2012; Montgomery and Olivella 2016; Muchlinski *et al.* 2016; Bansak *et al.* 2018), tree-based models are rarely used in political science, which tends to focus on substantive and/or causal interpretation of covariates.[2] Tree-based models—which are designed to incorporate flexible functional forms, avoid parametric assumptions, perform vigorous variable selection, and prevent overfitting—are common, however, in machine learning. These approaches are well suited for identifying variables important for forecasting, which could include variables that are not causal in nature *per se* but that are nonetheless predictive and for analyses involving large numbers of variables of potential (but uncertain) substantive importance.

The simplest tree-based models partition a dataset into "leaves" according to covariates and predict the value of each leaf. For example, a decision tree predicting Supreme Court rulings might start by splitting cases by whether the government is the respondent. If so, the algorithm may predict that the government wins. If not, the algorithm may examine the provenance of the case and, if there is a circuit split, predict that the petitioner wins. If it is not a circuit split, then it may examine whether Anthony Kennedy spoke frequently at oral arguments. If he did, the algorithm may predict that the respondent wins.

Our analysis relies on boosted decision trees, discussed in Montgomery and Olivella (2016) and which are newer to political science. (For an application of boosted regression trees to refugee allocation, see Bansak *et al.* 2018.) Boosting creates trees *sequentially*, and as Montgomery and Olivella (2016, p. 11) explain, each new tree then "improves upon the predictive power of the existing ensemble." The base classifier relies on "weak learners," decision rubrics that perform only slightly better than chance.

We use one of the most widely used boosting algorithms, AdaBoost. (See Appendix G, Section 8.4 for a discussion of other boosting approaches and why we chose AdaBoost.) AdaBoosting initializes by giving each observation equal weight. In the second iteration, AdaBoost will assign more weight to those units that were incorrectly classified in the first iteration. Focusing on those units that are hard to classify makes this approach well suited to social science problems, many of which involve heterogeneity and outliers.[3]

2.1 Pros and Cons of ADTs

ADTs' properties make it attractive for social science research. First, it has desirable asymptotics in improving predictive accuracy, especially when there are many features that each only contribute a small predictive gain. In predicting Court outcomes, although baseline accuracy is high, the predictive capacity of any one variable is small, leaving little room for improvement. This is common in the social sciences. Predicting the advent of civil wars has high baseline accuracy since there are few wars, but each additional predictor adds little information (Ward, Greenhill and Bakke 2010). Changes in which party controls the U.S. Presidency are often summarized by the "bread and peace" model: the incumbent party wins when the economy is growing, except during unpopular wars (Hibbs Jr 2000). This produces high baseline accuracy, with other variables adding little (Gelman and King 1993). Second, AdaBoost provides a useful theoretical guarantee: for any given iteration, as long as that model's predictions are consistently better than random chance, the overall model's training error is guaranteed to decrease (Mukherjee, Rudin and Schapire 2011).[4] Lastly, AdaBoost is agnostic to predictor or outcome data types, be they binary, continuous, or

[2] See Appendix E for discussion of why machine learning may be underused in political science.

[3] For a more technical walk-through of the AdaBoosting algorithm, see Appendix G.

[4] Train error refers to in-sample model fit, while test error refers to out-of-sample predictive accuracy. Here we measure predictive accuracy using exponential loss. This property of AdaBoost ensures that there are no local optima and no way to overfit.

https://doi.org/10.1017/pan.2018.59 Published online by Cambridge University Press

categorical (Elith, Leathwick and Hastie 2008), simplifying its implementation in dealing with mixed datasets of many predictors.

We also note drawbacks. First, ADTs sacrifice some interpretability of estimates for flexibility of functional form. By avoiding assumptions about the relationship between Court rulings and covariates, for example, ADTs provide more robust predictive capacity. However, they preclude discussions of statistical significance or effect sizes; rather than interpreting coefficients on covariates, ADTs rely on "feature importance." (Appendix C discusses how feature importance could nonetheless provide substantively important information that models like OLS miss.) Second, ADTs are computationally expensive without being parallelizable. Third, ADTs have many tuning parameters inherited from decision trees, and a few added from AdaBoost. Fourth, ADTs tend to overfit easily, especially compared to random forests (Elith *et al.* 2008). This can be controlled by limiting the learning rate (see Appendix G) at the cost of computation time. Lastly, there exist important problems for which AdaBoost fails. With insufficient sample sizes, primarily unpredictive covariates, or unsuitable base models, AdaBoost will show no improvement over more naive methods. Despite this, AdaBoost has been shown to work well in a wide variety of experimental settings among benchmark problems in computer science (Freund and Schapire 1996).

3 Application of AdaBoosting to the Supreme Court

We illustrate ADTs by predicting rulings by the U.S. Supreme Court. Because the Court decides cases of magnitude—including cases on presidential power, states' rights, and national security—even small predictive gains translate into significant policy importance. The simplest predictive algorithm for Court rulings is that the petitioner (party appealing the case) wins roughly two thirds of the time (Epstein *et al.* 2010).[5] In practice, guessing that the petitioner wins every time predicts 67.98% of cases since 2000 accurately (Appendix A1), though several studies have surpassed this baseline (Martin *et al.* 2004; Katz, Bommarito and Blackman 2014; Katz, Bommarito II and Blackman 2017). In this paper, we compare our approach to two prominent Court forecasting models, {Marshall}+ and CourtCast.[6]

We implement ADTs using the `scikit-learn` Python library.[7] We train our model (and comparison models) using two data sources from 2005 to 2015. First, we use case-level covariates from the Supreme Court Database (Spaeth *et al.* 2015). These include the procedural posture of the case, the issues involved, the parties' identities, and other case-level factors, detailed in Appendix C.[8] Second, we incorporate statements made by the Justices during oral arguments. Scholarship suggests that Justices use oral arguments to gather information and stake out positions (Johnson, Wahlbeck and Spriggs 2006). We draw on textual data from the Court's oral argument transcripts provided by the Oyez Project (Goldman 2002), which we operationalize into 55 variables, detailed in Appendix C. Finally, we optimize our model's tuning parameters using grid search (see Appendix G).

4 Results and Comparisons to Other Approaches

In Figure 1 below, we compare predictions based on (1) our model (referred to as "KKS") to (2) the "petitioner always wins" baseline rule, (3) CourtCast, (4) {Marshall}+, and (5) a generic random forest distinct from Katz *et al.* (2017). We evaluate all models using tenfold cross-validation (Efron

[5] A favorable ruling is at least a 5–4 majority. Note that we examine Court *outcomes* as opposed to the *votes* of individual Justices, in line with most papers in the literature.

[6] Source code for CourtCast is at https://github.com/nasrallah/CourtCast. See Appendix H.

[7] Complete replication materials are available on the *Political Analysis* Dataverse (Kaufman *et al.* 2018).

[8] Some of these variables are subjectively coded after the ruling is issued (for example, issue area). We see no way in which the coding would change pre- and postdecision. Appendix C provides further detail.

https://doi.org/10.1017/pan.2018.59 Published online by Cambridge University Press

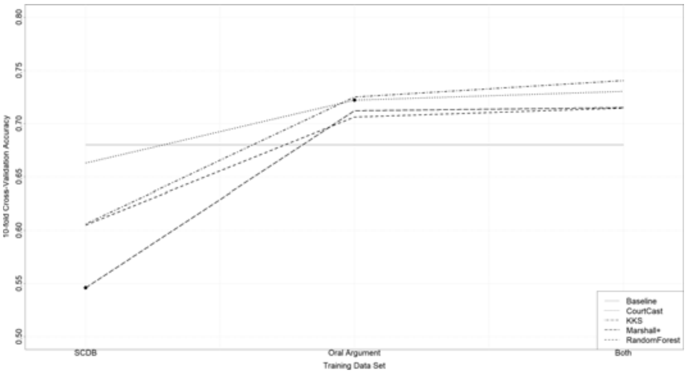

Figure 1. Cross-Validation Accuracy for KKS compared to the "petitioner always wins" baseline, CourtCast, {Marshall}+, and a generic random forest. We compare these across three datasets: Supreme Court Database ("SCDB"), oral argument data, and both datasets jointly. For {Marshall}+ and CourtCast, black dots indicate the original dataset on which those models were trained.

and Tibshirani 1997), which captures a model's ability to predict withheld samples of the observed data (see Appendix D).

In Table 1, we present each model's accuracy as reported by their authors in the original papers. For {Marshall}+, the original self-reported accuracy is much higher than we achieve (Figure 1), since it includes covariates we purposely excluded.[9] For CourtCast, self-reported accuracy is *lower* than we achieve: the original CourtCast model uses fewer training years and less accurate data than in our replications and measures accuracy using a single train-test split rather than 10-fold cross-validation.

Figure 1 indicates for each model the dataset (Supreme Court Database, oral arguments data, or both), cross-validation accuracy, and comparison to baseline accuracy. We generate these accuracy statistics by training the respective models on data from 2005 to 2015. We find that all models perform best using the joint dataset; all perform second best with the oral argument dataset. The KKS model using only case covariates performs less well, achieving an accuracy of more than 7 points below baseline. Using oral argument data, however, it exceeds baseline by more than 5 points with an accuracy of 72.5%. With joint data, it achieves an accuracy of 74.04%. Its added accuracy of 6.06 points over baseline is almost triple the added accuracy originally reported

Table 1. Accuracy (self-reported) for (1) the "petitioner always wins" baseline, (2) Katz *et al.* (2017)'s Random Forest, (3) {Marshall}+, (4) CourtCast, and (5) KKS. "Data" indicates the training dataset: case-level covariates from the Supreme Court Database ("SCDB"), transcript data from the oral arguments ("oral argument"), or both. The KKS model using all covariates almost triples the added accuracy of the next best model.

Model	Data	Self-reported Accuracy	Self-reported Accuracy – Baseline (pp)
Baseline	None	67.98%	0
Katz *et al.* 2017[10]	SCDB	70.20%	2.22
{Marshall}+	SCDB	70.20%	2.22
CourtCast	oral argument	70.00%	2.02
KKS	SCDB	60.6%	−7.4
KKS	oral argument	72.50%	4.5
KKS	Both	74.04%	6.06

[9] Specifically, the original {Marshall}+ analysis includes covariates gathered after oral argument, such as the month of the ruling. When we include all original {Marshall}+ covariates, we achieve a replicated accuracy that is comparable to their original results.

https://doi.org/10.1017/pan.2018.59 Published online by Cambridge University Press

Table 2. KKS model accuracy by decision margin.

Case Type	Baseline	{Marshall}+	CourtCast	KKS
Margin: 5–4	0.62	0.64	0.63	**0.66**
6–3	0.60	0.57	0.67	**0.73**
7–2	0.68	0.68	0.70	**0.76**
8–1	0.72	0.71	**0.82**	**0.82**
9–0	0.69	0.72	**0.78**	0.77
Issue: Criminal Procedure	0.67	0.63	0.71	**0.73**
Civil Rights	0.74	0.70	0.74	**0.77**
First Amendment	0.69	0.69	0.69	**0.74**
Economic Activity	0.65	0.67	**0.75**	0.73
Judicial Power	0.65	0.71	**0.73**	**0.73**
Federalism	0.50	0.50	0.55	**0.66**
Government is Party: Yes	0.65	0.65	0.74	**0.74**
No	0.68	0.68	0.69	**0.74**

by Katz et al. 2017. Substantively, this means our model correctly predicts about seven more cases (out of 80) per term than baseline—a meaningful improvement.

Interestingly, no model using only case covariates surpasses baseline accuracy; it is unsurprising that oral argument data, collected much closer to the decision, are more predictive than case covariates determined years prior to a ruling. We also note that by introducing the joint dataset to {Marshall}+ and CourtCast, both outperform their originally reported results, though neither perform as well as KKS on either the oral argument or the joint datasets.

5 Predictive Accuracy Conditional on Covariates

Our model enjoys an overall gain of approximately six percentage points over baseline, but this often increases when we examine subsets of cases. Close 5–4 decisions go to the petitioner 61% of the time on average, and our accuracy for 5–4 cases is 66%, five points above that baseline. We correctly predict 73% of 6–3 cases, 76% of 7–2 cases, 82% of 8–1 cases, and 77% of 9–0 cases; our model provides the biggest accuracy boost, 13 points, for 6–3 decisions.

Our model also outperforms the baseline in cases related to judicial power (nine points) and federalism (16 points) and where a state or federal government is a party (nine points). We see weaker gains in criminal procedure, civil rights, and First Amendment cases (Table 2). Our model outperforms {Marshall}+ and CourtCast in all subgroups except two: CourtCast performs one point better in unanimous cases and two points better in economic activity cases. However, both previous models often fail to exceed the baseline: {Marshall}+ in eight subgroups and CourtCast in two.

5.1 Additional applications: county-level U.S. presidential vote share & civil wars

ADTs are promising for other political science applications and may outperform even other tree-based methods. To demonstrate, we examine two applications. First, we look at U.S. presidential elections. For this, we analyze data from the 2010 U.S. Census that includes county-level age, income, education, and gender. The outcome variable is whether the Democratic Party's two-party county-level vote share in the 2016 presidential election is greater than 50%. The baseline is calculated by predicting that the Republican Party's two-party county-level vote share is greater than 50%. To assess accuracy, we use 10-fold cross-validation for the proportion of counties correctly predicted.

Second, we look at civil war incidence, examining a dataset indicating which country-years were engaged in civil wars, alongside country-level covariates derived from Collier and Hoeffler

https://doi.org/10.1017/pan.2018.59 Published online by Cambridge University Press

Table 3. ADTs outperform other methods in predicting both county-level vote share in the 2016 U.S. Presidential Election and civil war incidence.

Method	Elections Accuracy	Civil Wars Accuracy
ADTs	0.967	0.990
Random Forest	0.957	0.989
Support Vector Machines	0.954	0.983
Extremely Random Trees	0.948	0.990
LASSO	0.948	0.862
Logistic Regression	0.944	0.987
Baseline	0.876	0.861

(2002) and Fearon and Laitin (2003), including population, GDP, Polity score, ethnolinguistic fractionalization, and oil reserves. The baseline accuracy is 86.1%, achieved by predicting “no civil war” in all cases. To assess accuracy, we use 10-fold cross-validation on the proportion of country-years correctly predicted as having a civil war or not.

Table 3 presents these results. ADTs outperform competing linear, nonlinear, and tree-based methods. These improvements, even when small, are substantively meaningful. As the example of 2016 shows, presidential elections are consequential and hard to predict. In our dataset of 3,082 counties, being able to predict the likely vote of 308 more counties than baseline (and 31 more counties than the next best model), may impact how campaigns distribute resources. Predicting civil wars is likewise hugely important; accurately forecasting them holds great promise for allocating scarce peacekeeping resources. Across 6,610 country-years since 1945, our model correctly predicts 853 more cases than baseline (and seven more cases than the next best model), corresponding to 11.8 additional countries each year; it also predicts around 20 more cases than a logistic regression (0.36 more per year). Both are substantively meaningful differences that would be useful for policy experts and analysts.

6 Discussion and Conclusion

Our contributions are twofold. First, we provide an overview of ADTs, a technique frequently used in machine learning but one more novel within the social sciences. The approach is promising for many social science questions owing to its robustness to small sample sizes and its treatment of weakly predictive (though not unpredictive) covariates. As our examples show, this approach performs favorably compared to other commonly used methods across several applications. We include technical overviews and best practices guides in the Appendix.

Second, we contribute to a growing literature on Supreme Court prediction. The Court is the most reclusive branch of the U.S. government, yet it rules on some of the most important and contentious policy issues of the day. Increasing the predictive accuracy of forecasting models not only improves our understanding of how this important branch of government operates, but also, we believe, allows researchers to more credibly assess which way these influential rulings may go.

Acknowledgements

Many thanks to Matthew Blackwell, Peter Dilworth, Finale Doshi-Velez, Phillipa Gill, Gary King, Brian Libgober, Chris Lucas, Luke Miratrix, Kevin Quinn, Jeff Segal, Robert Ward, and participants at the Computational Social Science Institute Seminar at University of Massachusetts, Amherst for helpful conversations and valuable feedback. We also thank Josh Blackman, Michael Bommarito, and Dan Katz for comments during early stages of this project.

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2018.59.

References

Bansak, K., J. Ferwerda, J. Hainmueller, A. Dillon, D. Hangartner, D. Lawrence, and J. Weinstein. 2018. “Improving Refugee Integration through Data-Driven Algorithmic Assignment.” *Science* 359(6373):325–329.

Black, R. C., S. A. Treul, T. R. Johnson, and J. Goldman. 2011. “Emotions, Oral Arguments, and Supreme Court Decision Making.” *The Journal of Politics* 73(2):572–581.

Collier, P., and A. Hoeffler. 2002. “On the Incidence of Civil War in Africa.” *Journal of Conflict Resolution* 46(1):13–28.

Efron, B., and R. Tibshirani. 1997. “Improvements on Cross-Validation: the 632+ Bootstrap Method.” *Journal of the American Statistical Association* 92(438):548–560.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. “A Working Guide to Boosted Regression Trees.” *Journal of Animal Ecology* 77(4):802–813.

Epstein, L., W. M. Landes, and R. A. Posner. 2010. “Inferring the Winning Party in the Supreme Court from the Pattern of Questioning at Oral Argument.” *The Journal of Legal Studies* 39(2):433–467.

Fearon, J. D., and D. D. Laitin. 2003. “Ethnicity, Insurgency, and Civil War.” *American Political Science Review* 97(1):75–90.

Freund, Y., and R. E. Schapire. 1996. “Experiments with a New Boosting Algorithm.” In *Proceedings of the Thirteenth International Conference on Machine Learning,* vol. 96, 148–156. San Francisco, CA: Morgan Kaufmann Publishers.

Gelman, A., and G. King. 1993. “Why are American Presidential Election Campaign Polls so Variable when Votes are so Predictable?” *British Journal of Political Science* 23(4):409–451.

Goldman, J. 2002. The OYEZ Project [On-line].

Green, D. P., and H. L. Kern. 2012. “Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees.” *Public Opinion Quarterly* 76(3):491–511.

Hibbs, D. A. Jr. 2000. “Bread and Peace Voting in US Presidential Elections.” *Public Choice* 104(1–2):149–180.

Johnson, T. R., P. J. Wahlbeck, and J. F. Spriggs. 2006. “The Influence of Oral Arguments on the US Supreme Court.” *American Political Science Review* 100(1):99–113.

Katz, D. M., M. J. Bommarito, and J. Blackman. 2014. “Predicting the Behavior of the Supreme Court of the United States: A General Approach.” Available at SSRN: http://dx.doi.org/10.2139/ssrn.2463244.

Katz, D. M., M. J. Bommarito II, and J. Blackman. 2017. “A General Approach for Predicting the Behavior of the Supreme Court of the United States.” *PloS one* 12(4): e0174698.

Kaufman, A., P. Kraft, and M. Sen. 2018. “Replication Data for: Improving Supreme Court Forecasting Using Boosted Decision Trees.” https://doi.org/10.7910/DVN/JJCXTH, Harvard Dataverse, V1.

Martin, A. D., K. M. Quinn, T. W. Ruger, and P. T. Kim. 2004. “Competing Approaches to Predicting Supreme Court Decision Making.” *Perspectives on Politics* 2(4):761–767.

Montgomery, J. M., and S. Olivella. 2016. “Tree-Based Models for Political Science Data.” *American Journal of Political Science* 62(3):729–744.

Muchlinski, D., D. Siroky, J. He, and M. Kocher. 2016. “Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data.” *Political Analysis* 24(1):87–103.

Mukherjee, I., C. Rudin, and R. E. Schapire. 2011. “The Rate of Convergence of Adaboost.” In *Proceedings of the 24th Annual Conference on Learning Theory*, 537–558. Association for Computational Learning.

Ruger, T. W., P. T. Kim, A. D. Martin, and K. M. Quinn. 2004. “The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking.” *Columbia Law Review* 104:1150–1210.

Spaeth, H. J., L. Epstein, A. D. Martin, J. A. Segal, T. J. Ruger, and S. C. Benesh. 2015. *The Supreme Court database*. Center for Empirical Research in the Law at Washington University.

Ward, M. D., B. D. Greenhill, and K. M. Bakke. 2010. “The Perils of Policy by p-value: Predicting Civil Conflicts.” *Journal of Peace Research* 47(4):363–375.

https://doi.org/10.1017/pan.2018.59 Published online by Cambridge University Press