Institute for New Economic Thinking

Working Papers
ineteconomics.org/workingpapers

Predicting United States Policy Outcomes with Random Forests[†]

Shawn K. McGuire[*] and Charles B. Delahunt[**]

Working Paper No. 138

October 27nd, 2020

ABSTRACT

Two decades of U.S. government legislative outcomes, as well as the policy preferences of rich people, the general population, and diverse interest groups, were captured in a detailed dataset curated and analyzed by Gilens, Page et al. (2014). They found that the preferences of the rich correlated strongly with policy outcomes, while the preferences of the general population did not, except via a linkage with rich people's preferences. Their analysis applied the tools of classical statistical inference, in particular logistic regression. In this paper we analyze the Gilens dataset using the complementary tools of Random Forest classifiers (RFs), from Machine Learning.

We present two primary findings, concerning respectively prediction and inference:
(i) Holdout test sets can be predicted with approximately 70% balanced accuracy by models that consult only the preferences of rich people and a small number of powerful interest groups, as

[†] We thank Thomas Ferguson and Benjamin Page for their valuable insights and suggestions.
[*] Independent Researcher, Seattle, WA. smcguire@uw.edu.
[**] Applied Mathematics, University of Washington, Seattle, WA. delahunt@uw.edu

well as policy area labels. These results include retrodiction, where models trained on pre-1997 cases predicted “future” (post-1997) cases. The 20% gain in accuracy over baseline (chance), in this detailed but noisy dataset, indicates the high importance of a few wealthy players in U.S. policy outcomes, and aligns with a body of research indicating that the U.S. government has significant plutocratic tendencies.
(ii) The feature selection methods of RF models identify especially salient subsets of interest groups (economic players). These can be used to further investigate the dynamics of governmental policy making, and also offer an example of the potential value of RF feature selection methods for inference on datasets such as this one.

https://doi.org/10.36687/inetwp138

Keywords: political economy, elections, political parties, political money.

JEL Classifications: H10, D72, P16

1 Introduction

In 2014, Gilens and Page presented their ground-breaking paper, "Testing Theories of American Politics: Elites, Interest Groups, and Average Citizens" (Gilens and Page, 2014), based on their research into the influence of various actors (power groups) on the outcome of important policy outcomes in the United States. The dataset spanned two decades, from 1981 to 2002, and consisted of policy cases determined to be of high importance. They identified and described an extensive array of independent variables in their paper and in their books (Gilens, 2012; Page and Gilens, 2017).

The two most important independent variables (hereafter "features") outlined in their work were:
(i) The 90^{th} income percentile's opinion (P90);
(ii) The net Interest Group Alignment ("netIGA"), a single derived variable combining the effects of 43 powerful interest groups (IGs).
They showed that public opinion at the 50^{th} percentile had little to no effect on the odds of policy adoption, challenging received notions of democracy in the United States.

Additional research over the years has further questioned notions of American democracy. Research by Thomas Ferguson and his colleagues into the effect of major investors and money on election outcomes has shown that money flows are an excellent predictor of U.S. congressional races outcomes (Ferguson et al., 2019; Ferguson, 1995). In 2016, the United States was downgraded by the UK-based Economic Intelligence Unit from"Full Democracy" to "Flawed Democracy" (Economist Intelligence Unit, 2017).

Gilens and Page applied classical methods of statistics and regression to analyze their dataset. Increasingly, Machine Learning (ML) methods are making their way into social science research (Molina and Garip, 2019), and offer a distinct, complementary approach and set of tools for dataset analysis, one focused on prediction rather than inference (Breiman, 2001b). Prediction does not require interpretability, and some ML methods are largely "black-boxes" (*e.g.* neural nets). But some ML methods (*e.g.* Random Forests) have interpretable aspects, and are thus potentially useful for inference (Domingos, 2015).

Logistic regression (used by Gilens and Page) includes an intrinsic, strong assumption of linearity (Pampel, 2000) (for details see Appendix). Random Forests (RFs) in contrast are flexible, non-linear classifiers (Breiman, 2001a) that can handle large numbers of sparsely-represented features such as the preferences of the 43 individual IGs in the Gilens dataset (Karlsson, 2014). In addition, RFs have natural metrics to assess the relative importance of each feature. Thus RFs allow us to probe the dataset in complementary ways to Gilens and Page's use of a calculated netIGA and logistic regression.

In this work we apply RF methods to the Gilens dataset with two goals, prediction and inference: We use RFs to build predictive models of U.S. policy; and we extend Gilens and Page's inferential findings as to the influence of various actors on U.S. policy outcomes. We offer two main findings: (i) Policy outcomes on holdout sets can be predicted with approximately 70% balanced accuracy (*vs* 50% chance baseline) using only a few feature categories from the Gilens dataset: rich voters' preferences, a subset (as few as 14 out of 43) of individual IGs' preferences, and policy area labels. (ii) The RF feature importance metrics enable further understanding and analysis of the salience of individual actors, and also provide an example of RFs' potential usefulness for inference on datasets of this kind.

2 Methods

2.1 Dataset

The dataset consists of 1,836 major U.S. federal government policy cases dating from 1981 to 2002. In terms of prediction, the dependent variable y is case outcome ("adopted" or "not adopted"). We reduced the large array of features in Gilens' dataset to three primary categories, described below: *(i)* voter preferences (in particular, of the wealthy) *(ii)* IG alignments, and *(iii)* policy descriptors (such as "Foreign", "Social Welfare", *etc*). Corresponding abbreviations are listed in Table 1.

Table 1. Feature abbreviations

P90	Voter preference of 90^{th} income %ile
netIGA	Net Interest Group Alignment
IGs	Individual Interest Groups (preferences)
PAs	Policy Areas
PDs	Policy Domains (a coarser version of PAs)

Voter preferences (P90): Preferences of different wealth tranches were obtained from national surveys of the general public, where participants were asked whether they favored or opposed a proposed policy change. Preferences of the different wealth tranches were then imputed at various income percentiles, viz 90^{th}, 50^{th}, and 10^{th} (hereafter P90, P50, P10). Gilens and Page noted that P90 was a (rough) indicator of the preferences of even higher income percentiles (e.g. P99). One of their key findings was that, among voter preferences, only P90 impacted case outcomes. Our models used P90 as the sole voter preference feature (use of P50 or P10 as features degraded model accuracy).

Interest groups (IGs and netIGA): The dataset includes alignments, on each case, for a total of 43 distinct IGs. These IGs were mostly business groups such as the American Bankers Association, and also some social IGs such as the AARP (American Association of Retired Persons). The alignment of each IG toward each case was assigned an ordinal value between 2 and -2 (Strongly Support, Somewhat Support, Neutral, Somewhat Oppose, or Strongly Oppose). Routinely, a given IG had no opinion in a particular case, so non-neutral (non-zero) values were sparse.

To aid their analysis, Gilens and Page combined the alignments of all the individual IGs to create a single feature, the "netIGA" as follows:

$$\text{netIGA} = log(F_2 + 0.5F_1 + 1) \; - \; log(O_2 + 0.5O_1 + 1), \text{ where} \tag{1}$$

F_2 = number of IGs strongly in favor, F_1 = number somewhat in favor, O_2 = number strongly opposed, and O_1 = number somewhat opposed. The $log(\ )$ accounts for diminishing effects of multiple IGs weighing in on the same side of a given case. Because RFs readily handle large numbers of features, we set aside the netIGA and instead used each IG's alignment value as a separate feature.

Policy areas (PAs): The Gilens dataset includes an in-depth policy descriptor category which we refer to as policy area (PA) labels, *e.g.* Campaign Finance, Welfare Reform, *etc.* A full list of

the 19 PAs is given in the Appendix. PAs were expressed as features via one-hot encoding (*i.e.* one feature per PA, taking values 0 or 1, according as the case was in that PA). Each case is assigned to exactly one PA.

Policy Domains (PDs): The Gilens dataset also includes six broader policy domain (PD) labels (Economic, Foreign, Social welfare, Religious, Guns, and Miscellaneous). In general, several PAs are contained in one PD, though some PAs are also PDs (*e.g.* Foreign Policy). PDs were one-hot encoded. Table 2 shows a breakdown of positive (*i.e.* adopted) and negative (*i.e.* not adopted) cases by PD over the full dataset, as well as over just the post-1997 test set (used for retrodiction). PDs were used for feature selection: We grouped cases by PD, trained a different RF model for each PD using P90 and individual IGs as features, then selected the most salient IGs for each PD based on these models.

Table 2. Positive and negative case counts. Number of positive ("adopted") and negative ("not adopted") cases, as well as percentage of cases that were positive, for each PD, over the full dataset (numbers over post-1997 cases are in parentheses). Economic policy and Foreign policy are fairly evenly-balanced, while other PDs skew negative. In some PDs (*e.g.* Social Welfare, Guns) the distribution of positive/negative cases differed pre- *vs* post-1997.

Domain	**Pos**	**Neg**	**%Pos**
Economic	160 (36)	248 (38)	39% (49%)
Foreign	244 (77)	196 (59)	55% (57%)
Social Welfare	101 (20)	310 (152)	25% (12%)
Religious	43 (18)	123 (68)	26% (21%)
Guns	18 (1)	81 (49)	18% (2%)
Misc	77 (36)	235 (95)	25% (27%)
Total	643 (188)	1193 (461)	35% (29%)

2.2 Train/test splits

We trained models in two regimes: *(i)* Random train/test splits drawn from the full dataset ($N = 25$ draws, ratio 67%, 33%). This provided robust error bars on our prediction accuracy results since each train/test split was different. *(ii)* Future prediction (more precisely, retrodiction): All pre-1997 cases in training and all post-1997 cases (including 1997) in testing (ratio 65%, 35%). Retrodiction tested stability of models over time, with the possibility that individual IGs might gain or lose relevance over the two decade duration of the dataset.

2.3 Reported metrics

We report two figures of merit: maximum Balanced Accuracy, and Area Under the Curve (AUC) of the ROC curve. Evaluated on the test set, these are standard measures of a model's classifying abilities. Maximum balanced accuracy is defined as

$$balanced\ accuracy = \frac{sensitivity + specificity}{2}, \tag{2}$$

where sensitivity is the percentage of positive cases correctly classified, and specificity is the percentage of negative cases correctly classified. The operating point (decision threshold) used to predict test cases is that which gives maximum balanced accuracy on the training set (this threshold typically gives lower test set accuracy than would be possible given oracle knowledge of test set behavior).

We report Balanced rather than Raw accuracy for two reasons. First, it has a clear, consistent baseline accuracy (50% = chance). Second, in this dataset, negative cases outnumber positive ones, by substantial margins in some policy domains (cf. Table 2). This reflects a "status quo bias" with regards to policy adoption rate that varies by policy domain. Given this class imbalance, use of raw accuracy tends to blur the difference between the different models because all models (even bad ones) can leverage the imbalance by effectively betting on system inertia (*i.e.* the fact that most legislation is not adopted).

Use of Balanced Accuracy effectively eliminates the status quo bias by giving a higher weight to the results of the minority case outcomes. This avoids the phenomena of certain policy domains receiving increases to their scores simply by the fact that they have an unbalanced set of policy outcomes. The primary result of this metric is that it produces better insight into the salience of individual IGs as their effect on prediction is less muted by the status quo bias inherent in the dataset.

2.4 Random Forest models

RFs are flexible, robust, non-linear classifiers (Breiman, 2001a) based on ensembles of decision trees. In a RF, many decision trees are generated, each of which trains on a random subset of the training data. At each node of a given tree, a randomly-chosen feature splits the data. The final prediction of a test case is an average of the trees' predictions of that case. RFs are adept at handling sparse datasets with a large number of features (Karlsson, 2014), though removing uninformative features can improve model accuracy. The accuracies of two other flexible classifiers, XGBoost (Chen and Guestrin, 2016) (a flavor of RF) and Neural Nets, were similar to standard RFs (results not reported).

Our code was written in Python (Rossum and Drake, 2009) and used the sklearn library (Pedregosa and alia, 2011). The full codebase can be found at:
https://github.com/shawn-mcguire/predicting-policy-outcomes.

2.5 Feature selection

To rank the importance of the various IGs as features, cases were divided into 6 groups according to Policy Domain. For each domain, RFs were trained on random train/test splits ($N = 21$), using P90 and the 43 IGs as features. The averaged feature importance scores were then ranked. For more details see the Appendix.

3 Results

Results are divided into two sections: Inference (feature selection by RFs); and Prediction (including retrodiction).

3.1 Inference

The goal of Inference was to identify the most salient IGs by ranking their power as predictors. The 43 IGs were ranked as predictors for each Policy Domain as described above. A typical feature ranking (for the Foreign Policy domain) is shown in Table 3. Similar tables of IG rankings for other PDs are in the Appendix.

Table 3. Feature importance scores in Foreign policy. Feature importance scores, IG stance *vs* case outcome correlations, and IG "at-bats", given as mean ± std dev over 25 random train/test splits of the full dataset, restricted to Foreign policy cases only. "at-bats" gives the number of test cases for which the IG had a non-neutral stance (out of 147 ± 7 foreign policy test cases). "(+), (-)" indicate positive or negative IG-outcome correlations.

IG	RF score	IG-outcome correlation	at-bats (out of 147)
(+) P90	41 ± 3	20 ± 3	93 ± 6
(+) Defense industry	12 ± 3	41 ± 13	39 ± 5
(-) AIPAC	7 ± 2	-18 ± 11	17 ± 4
(-) UAW union	5 ± 1	-59 ± 14	41 ± 3
(+) Auto companies	5 ± 1	48 ± 15	13 ± 2
(+) Airlines	4 ± 1	54 ± 23	8 ± 2
(-) Oil companies	4 ± 1	37 ± 33	5 ± 2

Columns 2 and 3 are scaled by 100×.

The feature ranking extracted those IGs with the most impact on accurate prediction of policy outcomes, which has direct relevance to the study of policy decision-making in the U.S. However, "saliency" is here defined as maximizing model accuracy. This has some overlap with importance in U.S. policy-making, but with various complications, for example:

(1) Saliency is not the same as effectiveness of influence, because an IG's success is also a function of the difficulty of the particular cases it lobbied for or against. By analogy, a baseball player's batting average is partly a function of the pitchers they face.

(2) Bias can be introduced into the rankings based on various predictor attributes (Strobl et al., 2007). For example, less-sparse features (that is, IGs with more "at-bats") tend to have higher rankings. This is perhaps one reason for P90's consistently high rank. In this dataset, restricting by Policy Domain mitigates differences in sparsity because IGs tend to be especially active in a particular PD, *i.e.* within a given PD the active players tend to be overall less sparse.

(3) An IG has control over its "at-bats", which includes choices about risk levels, *e.g.* whether to take a stance on many or few cases, and whether to take on difficult cases or to conservatively jump in only on low-risk winners. These choices affect its sparsity level as well as its preference-to-outcome correlation.

(4) An IG can have high predictive saliency despite a negative or neutral correlation with case outcomes. High saliency due to negative correlations are commonly seen in linear regression. In RFs, high saliency despite neutral overall correlation can arise when the RF parses the cases into subsets, such that the IG's stances are strongly correlated with outcomes within the various subsets. An example from Foreign Policy is described in section 3.1.1. The co-existence of high predictive saliency and neutral correlation with outcomes can be a diagnostic, indicating that the feature is highly correlated with outcomes on certain subsets of cases. In this manner, RF models can provide initial avenues of investigation distinct from traditional regression techniques.

To calculate correlation between an IG preferences and case outcomes, only cases where the IG was not neutral (*i.e.* was "at bat") were considered:

$$corr(IG) = \frac{0.5}{\#at\ bats}\left(\sum_{i|y_i=1} x_i \ - \sum_{i|y_i=-1} x_i\right) \tag{3}$$

where x_i = IG preference $\in${ -2, -1, +1, +2} and y_i = outcome $\in${ -1 or 1}, for the i^{th} case. The 0.5 term normalizes the IG preference values. For these calculations, P90 values were rescaled from [0,1] to [-2, 2] and then values in [-0.4, 0.4] (i.e. noncommittal) were set to 0, to allow comparison with IGs. This measure is basically the scaled inner product of IG preference and case outcome $\frac{1}{n}\langle \boldsymbol{x}, \boldsymbol{y}\rangle$, where n is the number of cases with non-zero stances. It differs from Pearson's correlation coefficient in that it is not mean-subtracted (though $\boldsymbol{x}$ and $\boldsymbol{y}$ are implicitly zero-centered) and not normalized by standard deviations.

In the Foreign policy domain (*cf* column 3 of Table 3), Defense Contractors' stance was positively correlated with outcomes (mean $\pm$ std dev: 41 $\pm$ 13), Oil Companies' stance had mixed correlation (37 $\pm$ 33), and Labor Unions' stances were negatively correlated (-59 $\pm$ 14). Note that negative correlation does not mean that the IG "lost": Its actions may have averted worse outcomes, or modified the case's legislative details in desired ways. For example, while unions may have predictive power reflected by a high RF score, this does not mean they are powerful in terms of influencing policy. More detailed work looking at the individual policies would likely need to be performed to gain more insight into their actual influence.

The most salient IGs (excluding P90) for each Policy Domain, as determined by RF feature selection, were as follows:
1. Foreign policy: defense contractors; then AIPAC, auto companies and the auto workers union.
2. Economic policy: construction and realtors; then tobacco companies and a union.
3. Social welfare policy: the AARP and investment companies; then governors, pharmaceutical companies, universities, teachers, and health insurance companies.
4. Religious policy: Christian and anti-abortion groups; then doctors, teachers, beer companies, tobacco companies, health insurance companies, and broadcasters.
5. Gun policy: the NRA (no other IG was active).
6. Miscellaneous: Chamber of Commerce; then two unions (AFL-CIO and government workers), automobile companies; then oil companies, Christian Coalition, and trial lawyers.

The preferences of the wealthy (P90) outranked every IG in every policy domain except Social Welfare, where the AARP dominated. Tables of IG data for each Policy Domain (feature importance, correlations with outcomes, and number of "at-bats") are given in the Appendix.

3.1.1 Advantages of RF's nonlinear flexibility in feature selection

RFs do not have the built-in linear structure of logistic regression. In the inference context, this flexibility means that RFs can give insights into feature salience which are unavailable to logistic regression.

An example from the foreign policy domain data is described here. The highest ranked features for the RF model were P90 and Defense Contractors. The logistic regression model for the foreign policy domain ranked P90 and Defense Contractors as the 2nd and 6th most salient features, respectively. We examined policy cases involving P90 and Defense Contractors to gain some comparative understanding.

When Defense Contractors strongly favor a policy change, they almost always get their way, even when the wealthy are opposed. However, when the Defense Contractors oppose a policy change, P90 has a strong positive correlation with outcomes. Logistic regression fails to parse this effect, giving many False Negatives (see Fig 1). RFs handle this readily, yielding 95% balanced accuracy for the RF *vs* 79% for logistic regression. That is, the high importance of P90 in Foreign policy cases depends on a relationship more readily encoded by the RF model.

We note that RFs selected significantly more salient features than did logistic regression. For details see the Appendix, section 5.2.

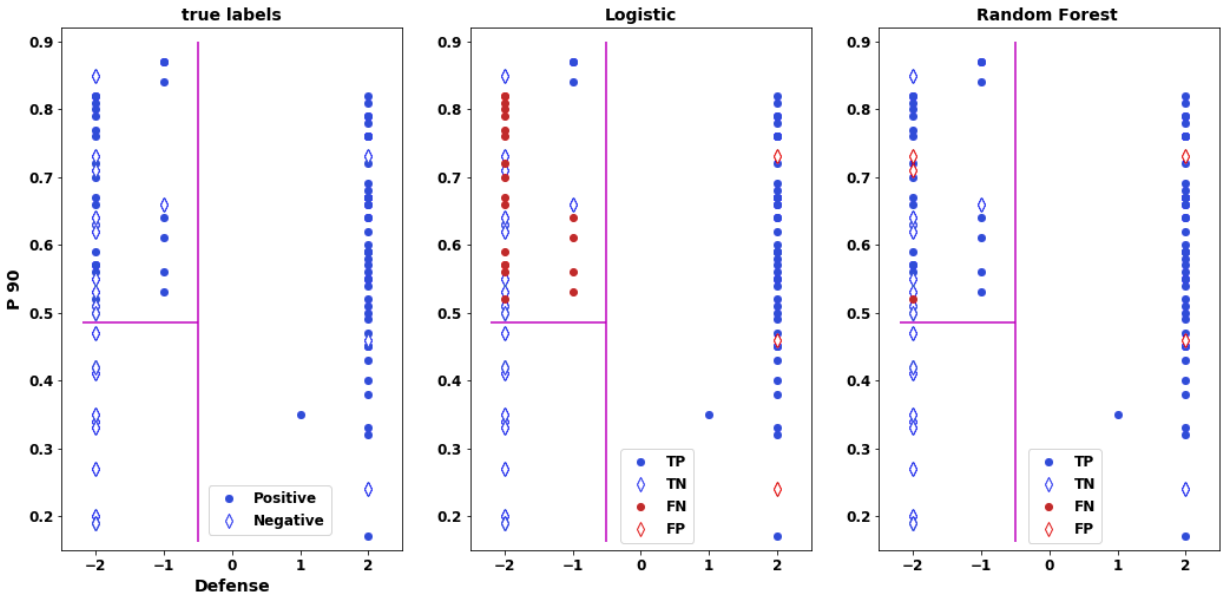

Fig 1. Benefit of nonlinear splits. Foreign policy outcomes on cases where Defense Contractors had a non-neutral stance. y-axis = P90, x-axis = Defense Contractors' preference. Solid dots are positive outcomes, open diamonds are negatives. TP = true positive, TN = true negative, FN = false negative, FP = false positive. **left:** True outcomes. There are three regions, marked by magenta lines: when Defense Contractors favored a case it almost always passed, regardless of P90 (right half of plot). When Defense Contractors opposed a case then either: it failed if P90 opposed it (lower left) or it had reasonable odds of passing if P90 favored it (upper left). **middle:** Logistic predictions (blue = correct, red = wrong), with many errors in the upper left region. **right:** RF predictions (blue = correct, red = wrong). RFs readily parse all three regions.

3.2 Prediction

The goal of Prediction was to examine to what degree U.S. legislative outcomes might be predicted from the preferences of rich voters and Interest Groups, and which actors were most informative.

Our key finding is that legislative outcomes can be accurately predicted using P90, policy descriptors (PDs or PAs), and individual IGs. Balanced accuracy on test sets of the trained RF reached ≈70%, a gain of 20% over baseline chance.

P90 was a vital feature, in the sense that excluding it always degraded accuracy. Inclusion of P50 (median wealth voter preferences) as a feature slightly degraded accuracy, consistent with the finding of P50's irrelevance in (Gilens and Page, 2014). Use of either PD or PA labels as a feature increased accuracy. Training and then combining separate models for each Policy Domain did not increase overall accuracy (results not reported).

We report results for RF models using four feature sets (listed below), in two train/test scenarios: *(i)* Multiple random train/test splits of the full dataset ("Random Draw" in Table 4); and *(ii)* Retrodiction (i.e., train on pre-1997 cases and test on post-1997 cases, a single train/test split). Feature sets used were:

Set A: P90 and netIGA (baseline, from (Gilens and Page, 2014)).

Set B: P90, PDs, and all 43 individual IGs.

Set C: P90, PDs, and 14 IGs chosen by RF feature selection (Gini impurity).

Set D: P90, PAs, and all 43 IGs.

Balanced accuracies and AUCs for RF models using the four Feature Sets and in the two regimes are given in Table 4. We offer the following observations:

(1) All feature sets used P90, which was always the most important. The feature sets differed in how they used IGs, and how they used Policy descriptors.

(2) Sets B - D substantially outperformed Set A (*e.g.* 11% mean increase in AUC).
Use of individual IGs *vs* the simplified netIGA was the main factor behind this improvement in prediction accuracy. Figure 2 shows the mean improvement in accuracy of Set B over Set A, broken out by individual IGs ("Random Draw" regime).

(3) Set C had equivalent performance to Set B, indicating that a small subset of 14 IGs (chosen by RF feature selection) carried as much salience as the full set of 43 IGs. We note this does not imply that these 14 IGs were the only ones that mattered: Certainly they were important actors, but they likely also encoded correlated salience of other IGs.

(4) Set D posted better performance than Model B, indicating that Policy Area labels had more salience than the coarser Policy Domain labels.

(5) Set D gave the best results (70% balanced accuracy, 78% AUC). The full list of features for Model D, with their importance rankings, is given in Table 6 in the Appendix.

Table 4. RF accuracy with various feature sets. Balanced Accuracy and AUC for RFs given various feature sets. "PD" means Policy Domains. "PA" means Policy Areas. Results are given as percentages, mean ± 1 std dev ($N = 25$).

Model	Random Draw		Retrodiction	
	Bal Acc %	AUC %	Bal Acc %	AUC %
A: P90, netIGA	61.5 ± 1.7	66.2 ± 2.1	64.6 ± 0.6	70.2 ± 0.3
B: P90, 43 IGs, PDs	67.3 ± 1.6	74.9 ± 1.7	69.0 ± 0.7	75.7 ± 0.2
C: P90, 14 IGs, PDs	67.3 ± 1.6	75.1 ± 1.6	68.8 ± 0.5	75.7 ± 0.2
D: P90, 43 IGs, PAs	**70.1 ± 1.5**	**77.7 ± 1.5**	**71.3 ± 0.7**	**76.5 ± 0.4**

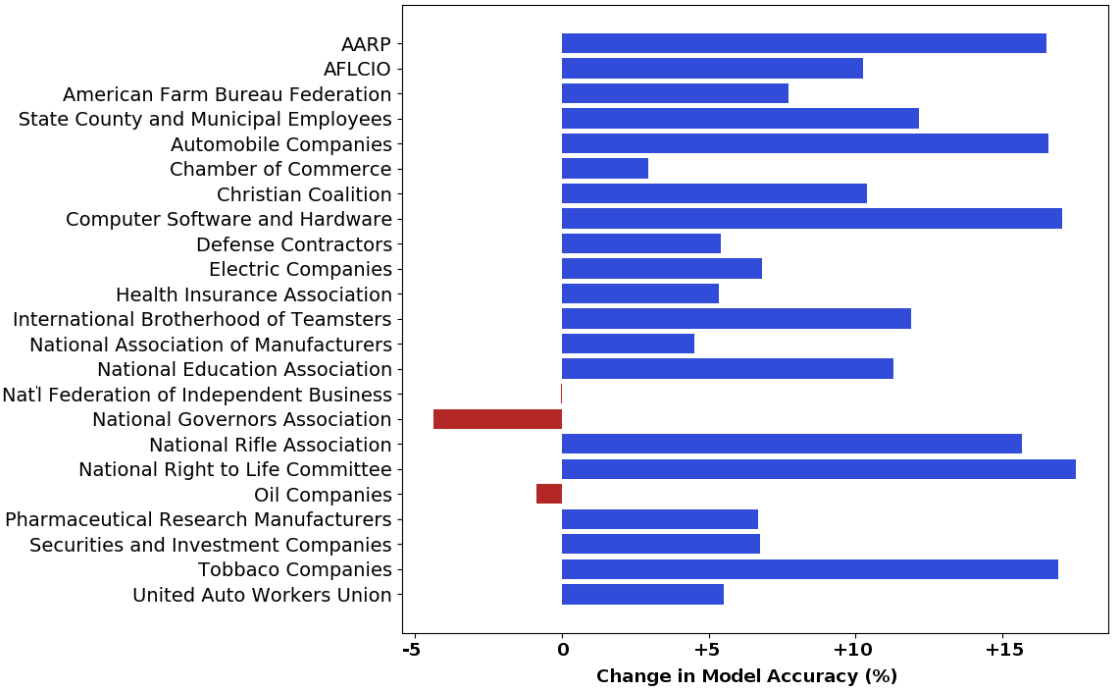

Fig 2. Gain in accuracy per IG, Set B *vs* Set A. Set B used P90, all 43 IGs and Policy Domains. Set A used P90 and netIGA only. For each IG, we measured test set accuracy on those cases where the IG was "strongly in favor" or "strongly opposed" and plotted *accuracy(Set B) - accuracy(Set A)*. Set B consistently improved accuracy on these IG-based groups of cases. Random draw regime, $N = 25$. Each IG shown had at least 20 cases in the test set.

4 Discussion

In this work we applied the tools of Machine Learning to the Gilens dataset, as an aid to inference as well as for prediction. RF algorithms readily handle sparse datasets that can be difficult for traditional regression techniques. RF methods enabled prediction of policy outcomes with surprising accuracy, especially considering the intrinsic noisiness of the dataset. They also enabled us to examine the role of individual IGs and the effects of various features of the dataset. Additionally, we have shown how RFs can provide insights into feature saliency distinct from and

arguably superior to traditional techniques such as logistic regression. There has been little research dedicated to predicting and understanding legislative action at the federal level using ML methods (Nay, 2017; Yano et al., 2012). Our effort is distinct from these few examples in that our results provide an initial lens for further research into the relations between distinct wealthy actors and policy outcomes.

We found that case outcomes on test sets were predictable with ≈70% balanced accuracy using only the preferences of the wealthy (P90), individual interest groups, and Policy Area labels as features. These results also held for a retrodiction split, training on pre-1997 cases and testing on post-1997 cases. We believe that the high predictability of policy outcomes using models based on only a few wealthy actors (P90 and certain IGs) reinforces Gilens and Page's findings about the plutocratic tendencies of U.S. government and aligns with other research examining the effects of money on policy creation (Ferguson et al., 2020; Ferguson, 1995; Igan and Prachi, 2014; Tahoun and Lent, 2016; Igan et al., 2011).

The original analysis by Gilens (Gilens and Page, 2014) showed that business interest groups as a whole had a much greater effect on policy outcomes than the mass-based interest groups. It is interesting to note that many of the business groups in our analysis had lower RF scores than would be expected (see Appendix: Table 6). We identify two contributing factors: first, the business groups are more numerous than the mass-based groups and consequently are more fragmented over the range of policy cases. Thus, each IG may have strong predictive power regarding a small subset of cases but a lower RF score because of their sparse activity overall. Consequently, if the individual business group contributions were combined to create a "net business IG", this meta-IG would have a much higher RF score. Second, many business IGs weigh in on the same case with the same alignment. It was found by (Gilens and Page, 2014) that when a business-oriented interest group took a position on a given policy, there was a less than 5% chance of another business group taking an opposing stance. The RF scoring methodology (Gini impurity) results in the dilution of the feature importance scores of IGs weighing in on the same side of a policy. These considerations suggest conducting an RF analysis utilizing features created by combining business IGs, similar to the analysis conducted in (Gilens and Page, 2014) using logistic regression.

Limitations in the independent variable categories must be mentioned and provide a gateway for thinking about future research. We start first with the interest group metrics and then the opinions of the wealthy.

Gilens and Page used the "Fortune Power 25" index to determine which interest group alignments to investigate. However, this limited index does not include the most prominent general big-business lobbies such as the Business Roundtable and the Business Council (Ferguson, 2013). IG alignments are also based on major media reports of lobby activity which may not be as accurate as desired. Additionally, the interval nature of the relatively coarse IG alignment variables lends the process to judgement calls by those assigning the values which may introduce error. We feel more detailed information regarding the relative strength of each IG's effort to affect a specific policy could lead to improved results and understanding. This should involve tracking the money flows inherent in the relationship between interest groups and policy creators. Research has already shown the strong link between industry monetary donations to Congress and voting patterns (Ferguson et al., 2020).

We suspect that the 90^{th} percentile serves as a proxy for the opinions of the very wealthy and that our results would likely be better with, say, the opinions of the 99^{th} percentile. However, chasing down the opinions of the top 1 percent via a random population survey would be a massive effort. There has been an effort in recent years to capture the policy opinions of the very wealthy as

distinct from the general population via a more targeted methodology. A unique study led by Benjamin Page effectively captured, via targeted surveys and interviews, the policy preferences of a small group of individuals in, or near, the top 1% of wealth in the United States (Page et al., 2013). The results indicated, among other things, that the very wealthy held policy preferences that were much more conservative in such important domains as social welfare programs, taxation, and regulation of the economic system. Their research also showed that the very wealthy were more politically active, with higher rates of financial contributions to, and contact with, public officials compared to the general population. Within this elite cohort, roughly two-thirds contributed an average of approximately $4,600 to political organizations and campaigns. While this survey method yielded very important understanding, we must also consider that the sharpest increase in correlation between policy outcomes and wealthy opinions may occur somewhere in the high end of the top 1 percent. Trying to increasingly hone in on the stated opinions of the uber wealthy in, say, the 99.9^{th} percentile has a few potential issues. First, it would likely involve a large, possibly prohibitive, amount of effort. Second, the results may improve correlation with - but not definitively address - the likely lodestar variable affecting policy outcomes, which is the transfer of large amounts of money to policy makers from the wealthiest sources focused intensely on particular policies. For example, the biggest corporations who can influence in their own right or members of the billionaire class throwing around their financial weight in the political sphere (Page et al., 2018; Meyer, 2017; Ferguson, 1995).

In short, we feel the best results in terms of inference from, and prediction of, individual policy outcomes would flow from data tracking the primary payments to congressional members from the most concentrated sources with the most to lose or gain from a particular policy change. Previous analyses such as the one conducted by (Ferguson et al., 2020) provide a potential blueprint for tracking such political money. Utilizing national surveys of the general public was necessary for Gilens and Page to demonstrate that the real-world data had no regard for previous claims asserting the median voter effect on policy. However, we think that our Random Forest analysis would yield even better results with independent variables more directly linked to the electronic delivery of suitcases of cash to policy makers.

5 Appendix

5.1 Logistic regression

Given a feature-label pair $\{\boldsymbol{x}, y\}$, logistic regression gives an estimate $\hat{y}$ of the class y, using features $\boldsymbol{x}$, by passing a linear combination of feature values $\boldsymbol{\beta x}$ through a (non-linear) sigmoid function $\sigma(s) = \frac{1}{1-e^{-s}}$. The coefficients $\boldsymbol{\beta}$ are the fitted parameters.

$$\hat{y} = \frac{1}{1 - e^{-\boldsymbol{\beta x}}} \tag{4}$$

5.2 Feature selection details

In order to identify the most important IGs, we used RF's standard Gini impurity method (Breiman and Friedman, 1984), indicated for this dataset because the features were sparse,

non-categorical, and similarly scaled (in range [-2, 2]) (Karlsson, 2014). Because IGs were ranked for each Policy Domain separately, the PDs were not features.

For feature ranking purposes, P90 scores were rescaled from [0,1] to [-2, 2] to match the IG preference range. Unlike Gini impurity, RF's Permutation method had unstable results (IG rankings changed if P90 was excluded). In general, the RFs selected different subsets of important IGs than those selected via the β coefficients of logistic regression.

Features selected by RF (Gini impurity) significantly out-performed features selected by logistic regression, in the sense that using RF-Gini features gave a model (either RF or logistic regression) much higher accuracy than using features selected by logistic regression. Features chosen by FR's Permutation method landed between the two. Table 5 shows the gain in accuracy due to RF (Gini)-chosen features vs logistic-chosen features.

We note that for a fixed set of features, RF and logistic regression models gave similar accuracies. In general, RF accuracy was slightly but not significantly better, consistent with (Couronné et al., 2018). The key advantage of RFs lay not in prediction *per se*, but rather in selecting much more salient features (*i.e.* inference).

Table 5. Difference in salience of IGs chosen by RF *vs* logistic regression. Balanced accuracy (balAcc) and AUC of RF and logistic models on test sets, using either RF-chosen IGs or logistic-chosen IGs, given as mean ± std dev. The gains in accuracy (RF-chosen minus Logistic-chosen) are shown in bold. (N= 21 train/test splits).

Setup		**Full data**		**Retrodiction**	
Model	IG type	balAcc	AUC	balAcc	AUC
RF	RF-chosen (Gini)	63.5 ± 1.6	68.2 ± 2.2	66.4 ± 2.4	69.7 ± 1.0
	Logistic-chosen	56.7 ± 1.7	61.1 ± 1.9	53.4 ± 2.7	53.9 ± 2.7
	Gain	**6.8 ± 1.9**	**7.0 ± 1.6**	**13.0 ± 4.4**	**15.8 ± 3.3**
Logistic	RF-chosen (Gini)	59.5 ± 1.8	61.4 ± 2.1	59.3	62.6
	Logistic-chosen	55.0 ± 1.6	56.6 ± 1.6	52.6	51.6
	Gain	**4.5 ± 1.9**	**4.7 ± 1.7**	**6.7**	**10.9**

"Gain" is the mean of differences (not difference of means). Logistic regression is deterministic on non-separable data, so its prediction of the post-1997 data has 0 std dev.

5.3 Set D feature importances.

Feature Set D gave the most accurate predictions. Features consisted of: P90 (continuous values ∈ [0,1]); 43 individual interest groups (values ∈ {-2, -1, 0, 1, 2}); and 19 Policy Areas (Budget, Campaign Finance, Civil Rights, Defense, Economics and Labor, Education, Environment, Foreign Policy, Government Reform, Guns, Health, Immigration, Miscellaneous, Race, Religion, Social Welfare, Taxation, Terrorism, Welfare Reform, all one-hot encoded). Table 6 shows feature rankings found by RFs (Gini impurity) for Set D.

Note that because this feature set combines one-hot and ordinal value encodings, the caveats in (Karlsson, 2014) and (Strobl et al., 2007) about distortion of feature rankings may apply. As noted previously, the RF scores for the individual PAs have the balanced accuracy target encoded into their RF scores, which effectively avoids the status quo bias intrinsic to the different policy areas.

Table 6. Set D feature importances. Mean scores ($N = 50$ runs) of the top 25 most salient features. Policy Area features indicated by "(PA)" before the feature. RF importance scores are scaled by 100×.

Feature	Mean RF score
P90	27.3
AARP	8.9
(PA) Foreign Policy	7.9
Defense Contractors	4.9
Chamber of Commerce	2.8
(PA) Social Welfare	2.1
National Association of Realtors	2
National Right to Life Committee	2
Health Insurance Association	1.8
Christian Coalition	1.8
National Rifle Association	1.7
(PA) Health	1.7
AFLCIO	1.5
(PA) Campaign Finance	1.5
(PA) Defense	1.4
American Farm Bureau Federation	1.3
American Israel Public Affairs Committee	1.3
(PA) Guns	1.2
Securities and Investment Companies	1.2
National Federation of Independent Business	1.2
National Association of Manufacturers	1.1
Automobile Companies	1.1
United Auto Workers Union	1
Tobacco Companies	1

5.4 IG rankings by Policy Domain

Importance rankings, correlations with case outcomes, and number of "at-bats" for IGs, for various Policy Domains, are given in Tables 7 - 10. For details and Foreign policy results, see section 3.1.

Table 7. Economic policy. Correlations for the most salient IGs, given as mean ± std dev. Each test set had 130 ± 6 Economics cases. "(+), (-)" indicate positive or negative IG-outcome correlations.

IG	RF score	IG-outcome correlation	at-bats (out of 130)
(+) P90	21 ± 2	12 ± 4	94 ± 5
(+) Nat'l Assoc Homebuilders	6 ± 2	23 ± 7	41 ± 5
(+) Nat'l Assoc Realtors	6 ± 2	46 ± 11	22 ± 4
(+) Teamsters union	5 ± 1	40 ± 11	19 ± 3
(+) Tobacco Companies	5 ± 1	18 ± 11	31 ± 6

Columns 2 and 3 are scaled by 100×.

Table 8. Social Welfare policy. Correlations for the most salient IGs, given as mean ± std dev. The AARP had strong positive correlation with outcomes, while 90th %ile was mixed and investment interests had strongly negative correlations. Each test set had 137 ± 7 Social Welfare cases. "(+), (-)" indicate positive or negative IG-outcome correlations.

IG	RF score	IG-outcome correlation	at-bats (out of 137)
(+) AARP	23 ± 2	60 ± 8	71 ± 7
(-) P90	18 ± 2	-3 ± 5	88 ± 9
(-) Invest & Securities Assoc	13 ± 2	-94 ± 8	15 ± 3
(+) Nat'l Governors Assoc	6 ± 1	4 ± 20	22 ± 4
(+) Pharmaceuticals	5 ± 1	48 ± 22	13 ± 4
(+) Universities	5 ± 2	55 ± 50	3 ± 1
(+) Nat'l Education Assoc	4 ± 1	27 ± 25	16 ± 3
(+) Health Insurance Assoc	4 ± 1	36 ± 22	21 ± 3

Columns 2 and 3 are scaled by 100×.

Table 9. Religious policy. Correlations for the most salient IGs, given as mean ± std dev. Only 12 IGs had non-zero alignments on Religious cases, and only 2 IGs were often non-zero. Most features, including P90, had mixed correlations with outcomes. Anti-abortion groups had strongly negative correlations. Each test set had 56 ± 5 Religious cases. "(+), (-)" indicate positive or negative IG-outcome correlations.

IG	RF score	IG-outcome correlation	at-bats (out of 56)
(+) P90	36 ± 3	6 ± 5	30 ± 4
(-) Nat'l Right to Life	23 ± 2	-55 ± 19	16 ± 4
(-) Christian Coalition	10 ± 2	-14 ± 14	43 ± 4
(-) Am. Medical Assoc	7 ± 2	-21 ± 38	4 ± 1
(+) Nat'l Education Assoc	6 ± 2	65 ± 26	5 ± 1
(-) Tobacco Companies	4 ± 1	-11 ± 55	4 ± 1
(-) Beer Companies	4 ± 1	-34 ± 72	2 ± 1

Columns 2 and 3 are scaled by 100×.

Table 10. Gun policy. Correlations for the most salient IGs, given as mean ± std dev. The NRA was the only active IG, and it had strong positive correlation with outcomes. P90 (and P50) had low overall correlation with outcomes, but paradoxically had much higher importance scores. This was likely due to its non-linear splitting ability even while not "getting what it wanted" (similar to the Foreign Policy example described in section 3.1.1). Each test set had 33 ± 6 Gun cases. "(+), (-)" indicate positive or negative IG-outcome correlations.

IG	RF score	IG-outcome correlation	at-bats (out of 33)
(-) P90	96 ± 1	-11 ± 11	28 ± 5
(+) Nat'l Rifle Assoc	4 ± 1	51 ± 16	32 ± 5

Columns 2 and 3 are scaled by 100×.

Table 11. Miscellaneous policy. Correlations for the most salient IGs, given as mean ± std dev. Each test set had 103 ± 8 Miscellaneous cases. "(+), (-)" indicate positive or negative IG-outcome correlations.

IG	**RF score**	**IG-outcome correlation**	**at-bats (out of 103)**
(-) P90	22 ± 2	-2 ± 4	69 ± 8
(+) Chamber of Commerce	8 ± 3	10 ± 10	26 ± 4
(-) AFL-CIO (union)	5 ± 2	-40 ± 10	21 ± 4
(-) State-County-Munic Employees (union)	4 ± 1	-41 ± 10	19 ± 3
(+) Automobile Companies	4 ± 1	12 ± 11	23 ± 3
(+) Christian Coalition	4 ± 1	42 ± 18	18 ± 3
(+) Trial Lawyers	4 ± 1	48 ± 8	17 ± 3

Columns 2 and 3 are scaled by 100×.

References

Breiman, L. (2001a). Random Forests. *Machine Learning*.

Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Stat Sci*.

Breiman, L., & Friedman, J. (1984). *Classification and regression trees*. Taylor & Francis.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proc 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining*.

Couronné, R., Probst, P., & Boulesteix, A. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*.

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books.

Economist Intelligence Unit. (2017). Democracy Index 2016: Revenge of the 'deplorables'. *The Economist*.

Ferguson, T. (1995). *Golden Rule: The Investment Theory of Party Competition and the Logic of Money-Driven Political Systems*. U of Chicago Press.

Ferguson, T. (2013). Reviewed Work(s): Affluence and Influence: Economic Inequality and Political Power in America by Martin Gilens. *Perspectives on Politics, Vol. 11, No. 1*.

Ferguson, T., Jorgensen, P., & Chen, J. (2019). How Money Drives US Congressional Elections: Linear Models of Money and Outcomes. *Structural Change and Economic Dynamics*.

Ferguson, T., Jorgensen, P., & Chen, J. (2020). How Much Can the U.S. Congress Resist Political Money? A Quantitative Assessment. *Institute for New Economic Thinking. Working Paper, No. 109*.

Gilens, M. (2012). *Affluence and Influence: Economic Inequality and Political Power in America*. Princeton University Press.

Gilens, M., & Page, B. (2014). Testing Theories of American Politics: Elites, Interest Groups, and Average Citizens. *Perspectives on Politics*.

Igan, D., & Prachi, M. (2014). Wall Street, Capitol Hill, and K Street: Political Influence and Financial Regulation. *Law and Economics*.

Igan, D., Prachi, M., & Tressel, T. (2011). A Fistful of Dollars: Lobbying and the Financial Crisis. *NBER Macroeconomics Annual 2011, edited by D. Acemoglu and M. Woodford*.

Karlsson, B. (2014). Handling Sparsity with Random Forests When Predicting Adverse Drug Events from Electronic Health Records. *IEEE Int'l Conf on Healthcare Informatics, Verona*.

Meyer, J. (2017). *Dark Money: The Hidden History of the Billionaires Behind the Rise of the Radical Right*. Anchor Books.

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Ann Rev of Sociology*.

Nay, J. (2017). Predicting and Understanding Law Making with Word Vectors and an Ensemble Model. *PLOS One*.

Page, B., Bartels, L., & Seawright, J. (2013). Democracy and the Policy Preferences of Wealthy Americans. *Perspectives on Politics*.

Page, B., Seawright, J., & Lacombe, M. (2018). *Billionaires and stealth politics*. University of Chicago Press.

Page, B., & Gilens, M. (2017). *Democracy in America?: What Has Gone Wrong and What We Can Do About It*. University of Chicago Press.

Pampel, F. (2000). *Logistic Regression*. Sage Publishing.

Pedregosa, F., & alia. (2011). Scikit-learn: Machine Learning in Python. *JMLR*.

Rossum, G. V., & Drake, F. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Strobl, C., Boulesteix, A., Zeileis, A., & alia. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*.

Tahoun, A., & Lent, L. (2016). The Personal Wealth Interests of Politicians and the Stabilization of Financial Markets. *Institute for New Economic Thinking, Working Paper No. 52*.

Yano, T., Smith, N., & Wilkerson, J. (2012). Textual Predictors of Bill Survival in Congressional Committees. *Proc 2012 Conf N Amer Chapter Assoc Comp Linguistics, Human Language Technologies*.