Machine Learning Tools and Techniques: Grammar of Law through Data and Evidence for Justice Reform (DE JURE)

Nikhil Supekar (ns4486), Angela Lai (ayl316),
Shwetanshu Singh (ss11404), Alexandria Salib (as12453)

May 15, 2021

1 Introduction

We study the DE JURE Motivated Reasoning project asking how district court judges appointed by Democrat presidents differ from judges appointed by Republican presidents. In particular, we look at the characteristics of cases where Democrat and Republican district court judges tend to favor the plaintiff. We are particularly interested in whether the "treatment" of a judge's political party seems to have an effect on whether they tend to rule in favor of the plaintiff (which we define as the outcome or response). First, we test whether the political party of the judge of each case seems to be randomly assigned. We then run CATE estimation using a variety of learners to estimate the treatment effect of each case. Finally, we inspect feature importances and SHAP values of a model fit to classify cases in favor of the plaintiff vs the defendant and infer the types of cases where Republican judges tend to favor the plaintiff vs. the defendant, and similarly for Democrat judges.

We worked with Daniel Chen, Henrik Sigstad, and Sandeep Bhupatiraju from the DE JURE program at the World Bank and thank them for their guidance. This project has not been attempted in the past. The Github repository associated with this project is linked here.

2 Data

Overview

District level data is used to explore the decisions made by the district court judges. Our data set consists of three broader elements:

1. Biographical Directory of Article III Federal Judges: from 1789 - present. A subset of this data, Federal Judicial Service, has been highlighted by the team.
2. 2.2 mil District Court Opinions with Judge Identifiers
3. Federal Court Cases: Federal Judicial Center Integrated Database (IDB) spanning cases from 1970 to present.[1]

Cleaning

The case opinions were provided as semi-structured zipped XML files (totaling 180+ GB). We used NYU's HPC cluster (Greene) compute to parallel process the opinions to extract structured case data. The DE JURE team provided us with sample code for parsing the XML files which we debugged and modified to accommodate data irregularities in the larger data set. The Federal Judicial Center has exposed judge data in a CSV format that contains judge covariates. First, we parsed the XML files for a subset of features and output the results in JSON files. We then merged the case data and biographical judge data using a fuzzy matcher on the names of the judges. One of the zipped XML files was corrupted, potentially accounting for the difference between the expected number of opinions and the number of observations in our data set.

The resulting data set is a zipped 8.5 MB CSV file. It has 1,997,522 cases with the following columns:
`decision_date, filing_date, Court Name, Party of Appointing President, CIRCUIT, JURIS, NOS, ORIGIN, RESIDENC, CLASSACT, DEMANDED, TERMDATE, DISP, PROCPROG, NOJ, AMTREC, JUDGMENT, TAPEYEAR, DISTRICT, OFFICE, COUNTY, TRCLACT, PROSE, ARBIT, TRANSOFF, TRMARB, IFP, STATUSCD`.

Table 1 describes the viable feature columns in more detail. Note that we only use a subset of these for the CATE estimation: columns `TRMARB, ifp, transoff, arbit,` and `RESIDENC` were excluded because they had too many null values. We also omit the columns `court_name` and `TAPEYEAR` due to a high correlation with existing features.

[1] https://www.fjc.gov/research/idb/

Name	Type	Description
CIRCUIT	categorical	Circuit in which the case was filed.
FILEDATE	date	The DATE on which the case was filed in the district.
JURIS	categorical	The code which provides the basis for the U.S. district court jurisdiction in the case. This code is used in conjunction with appropriate nature of suit code.
NOS	categorical	Nature of Suit: A 3 digit statistical code representing the nature of the action filed.
ORIGIN	categorical	A single digit code describing the manner in which the case was filed in the district.
RESIDENC	categorical	Involves diversity of citizenship for the plaintiff and defendant. First position is the citizenship of the plaintiff, second position is the citizenship of the defendant.
CLASSACT	categorical	Involves an allegation by the plaintiff that the complaint meets the prerequisites of a "Class Action" as provided in Rule 23 - F.R.CV.P.
DEMANDED	number	The monetary amount sought by plaintiff (in thousands).
TERMDATE	date	The DATE the district court received the final judgment or the order disposing of the case.
DISP	categorical	The manner in which the case was disposed of.
PROCPROG	categorical	Procedural Progress: The point to which the case had progressed when it was disposed of.
NOJ	categorical	Cases disposed of by an entry of a final judgment.
AMTREC	number	Dollar amount received (in thousands) when appropriate.
JUDGMENT	categorical	Cases disposed of by entry of a final judgment in favor of plaintiff or defendant.
TAPEYEAR	number	Statistical year label on data files obtained from the Administrative Office of the United States Courts. 2099 on pending case records.
DISTRICT	categorical	District court in which the case was filed.
OFFICE	categorical	The code that designates the office within the district where the case is filed.
COUNTY	categorical	The code for the county of residence of the first listed plaintiff.
TRCLACT	categorical	Termination Class Action: A code that indicates a case involving allegations of class action.
PROSE	categorical	Pro Se field is blank in records posted before October 1995.
ARBIT	categorical	This field is used only by the courts participating in the Formal Arbitration Program. It is not used for any other purpose.
TRANSOFF	date	The office number of the district losing the case.
TRMARB	categorical	Termination arbitration code.
IFP	categorical	In forma pauperis: This field captured since October 2000.
STATUSCD	categorical	Status code to identify the type of record.

Table 1: Descriptions of features

3 Methodology

To answer the initial question with the provided data, we use CATE estimation to learn what characterizes the cases where Democrat or Republican appointed judges tend to favor the plaintiff. To do this, we first test whether it seems plausible that each case is randomly assigned a Republican or Democrat judge or if the "treatment" seems unevenly applied. We then extract features from the cases that are as far as possible predetermined and therefore plausibly not influenced by the judge (case, identity of plaintiff, etc). With this information, we aim to answer the following questions:

- How different are judges appointed by Democrat and Republican Presidents in their decisions? How has this changed over time?
- Which types of cases do Republican judges tend to favor plaintiffs over Democratic judges, and vice versa? What characteristics are exhibited by these cases?
- Are judges from one party more extreme than the other? Is there a difference in the amount of disagreement?

Testing for Randomization

We test whether the feature columns are meaningful predictors of the treatment by running a regression model where the prediction is the judge's political party and the independent variables are the features. Since this data is observational, we do not necessarily expect random treatment assignment.

We add a column for every combination of court district and year so that we can control for their fixed effects. We take the features `CLASSACT, JURIS, ORIGIN, office, NOS, district, RESIDENC, filing_year, district_year` and use a one hot encoder on the categorical features. This yields 3249 feature columns.

Next, we train an ordinary least squares linear regression model on the data and run an F-test on the parameters corresponding to the non-control and control indices — that is, the parameters corresponding to the non-district_year features versus district_year.

We initially ran this on a subset of our final data and obtained the results in 2. Based on the high p-values and relatively low F-test values, we interpreted this as meaning that the "treatment" was randomly associated with the features and proceeded with CATE estimation. However, after running the same code on the full final data, we obtained the results in 3. The degrees of freedom of these results differ from those in 2, suggesting an unexpected difference in F-test computation and that our independent variables are useful when predicting the judge's political party.

This is not necessarily surprising since one might expect that an observational study's treatment will not be randomly assigned. Features like district, year, and `RESIDENC` may also well have some association with a judge's political party. Since many, if not most, of our features are categorical and mutually exclusive, we expect that the models used in CATE estimation will be able to account for the fixed effects of case characteristics.

	F-test value	P-value	Deg. of freedom (numerator)	Deg. of freedom (denom.)
Non-district_year	1.72492091	0.18906	1	1.18e+06
District_year	1.73055249	0.18834	1	1.18e+06

Table 2: F-test results on a subset of the data.

	F-test value	P-value	Deg. of freedom (numerator)	Deg. of freedom (denom.)
Non-district_year	84.2325803	0.0	321	1.95e+06
District_year	73.93438057	0.0	3.08e+03	1.95e+06

Table 3: F-test results on the full data set.

CATE Estimation

We use Uber's Causal ML package for ITE/CATE estimation [1]. Compute time is nontrivial since our data has 2991 features. With the DE JURE team's approval and based on our observation that the CATE estimation results do not seem to change significantly as we increase our sample size, we run CATE estimation using a T-learner, X-learner, R-learner, and S-learner on a sample of one million cases. We further reduced the number of cases to 50k for faster development iterations and report the analysis on these.

We utilize the JUDGMENT feature (which indicates the case outcome) to represent 1 if the outcome of the case goes in favor of the plaintiff, and -1 if the outcome goes in favor of the defendant. With this setup, a CATE value greater than zero indicates that favoring the plaintiff and less than zero indicates favoring the

defendant.

The CATE estimates are consistent across learners with a mean CATE estimate consistently around 0.01. This indicates that across all cases, on average there is no difference in treatment by judges towards the plaintiff or the defendant.

Learner	Mean	Std
T	0.0114	0.201
X	0.0112	0.1057
R	0.0107	0.1547

Table 4

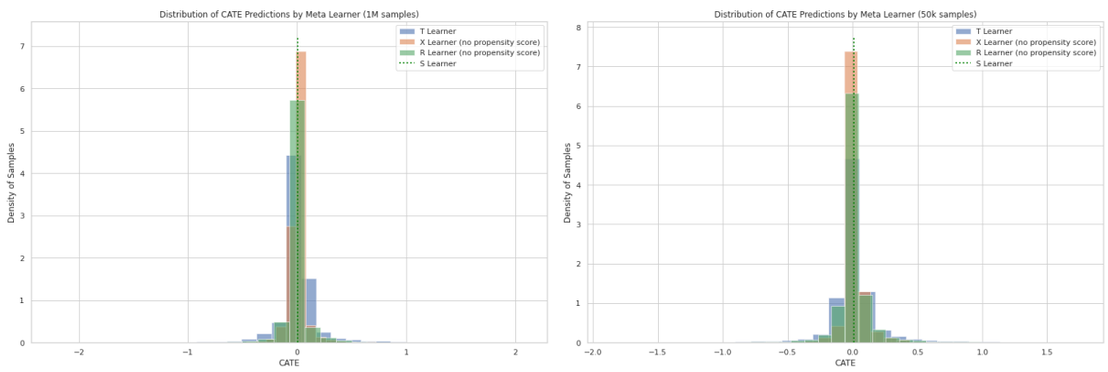


Figure 1: Results appear to stay consistent even as the sample size is reduced from one million to 50,000.

Feature Importance

With the CATE estimates handy for every case, we further wish to analyze cases where the judges strongly favor the plaintiff or the defendant. For this analysis, we look at cases where the CATE estimate is towards the higher end or lower end. We set our thresholds to look at the cases where the CATE estimate falls either top 25 percentile and bottom 25 percentile. The top 25 percentile bucket is labeled as 1 - indicating that this case is favored towards the plaintiff - and the bottom 25 percentile bucket is labeled 0 - indicating the case favored towards the defendant. This reduced the set of cases for further analysis to around 6000 cases.

We now split the data set into two parts: one each for Republican and Democrat judges and try to classify if the case favors plaintiff/defendant for each set. We fit two XGBoost models, one each for Republicans and Democrats to classify favoring plaintiff/defendant based on the buckets created above using the CATE estimates. We inspect the feature importances and SHAP values for these two models to draw inferences about various case types. NOS (Nature of Suit) and NOJ (Nature of Judgment) are two of the most interpretable features that show up in the top ranked feature importances and hence it is more interesting to see the impact of the presence of these features on favoring the plaintiff/defendant.

We report the SHAP plots for the two models for the top 20 features ranked by their mean absolute SHAP value, i.e. features ranked by their impact on the model output (either a positive or a negative impact). The way to interpret these SHAP values in our setting is: if the value of a feature X is high (depicted by red color on the plot) on the left side of zero, then judges tend to favor the plaintiff in presence of this feature and if the value of X is high on the right side of zero, then judges tend to favor the defendant. For example, the SHAP plot the XGBoost model for Republican judges suggests that for judges favor the plaintiff for cases of the type NOS = CABLE/SATLELLITE TV while favor the defendant for ENVIRONMENTAL MATTERS. We compile a list of top 10 NOS's favored by Republican/Democrat Judges in 5 and 6.

Republicans favoring plaintiffs	**Republicans favoring defendants**
BANKS AND BANKING	LAND CONDEMNATION
CABLE/SATELLITE TV	BANKS AND BANKING
BANKRUPTCY APPEALS RULE 28 USC 158	STOCKHOLDER'S SUITS
CIVIL RIGHTS WELFARE	EDUCATION
CONSTITUTIONALITY OF STATE STATUTES	RAILWAY LABOR ACT
OTHER PERSONAL PROPERTY DAMAGE	TRADEMARK
OTHER PERSONAL INJURY	ENVIRONMENTAL MATTERS
PRISONER PETITIONS -HABEAS CORPUS	HEALTH CARE / PHARM
SECURITIES, COMMODITIES, EXCHANGE	MARINE CONTRACT ACTIONS
TAX SUITS	ARBITRATION

Table 5

Democrats favoring plaintiffs	**Democrats favoring defendants**
OTHER PERSONAL PROPERTY DAMAGE	LABOR/MANAGEMENT REPORT & DISCLOSURE
BANKRUPTCY APPEALS RULE 28 USC 158	LAND CONDEMNATION
CONSTITUTIONALITY OF STATE STATUTES	EDUCATION
TAX SUITS	OTHER FORFEITURE AND PENALTY SUITS
MOTOR VEHICLE PERSONAL INJURY	IRS 3RD PARTY SUITS 26 USC 7609
CIVIL RIGHTS ADA EMPLOYMENT	AGRICULTURAL ACTS - 891
HABEAS CORPUS: DEATH PENALTY	OTHER STATUTORY ACTIONS
CABLE/SATELLITE TV	STOCKHOLDER'S SUITS
CIVIL RIGHTS WELFARE	ENVIRONMENTAL MATTERS
TORTS TO LAND	DEPORTATION

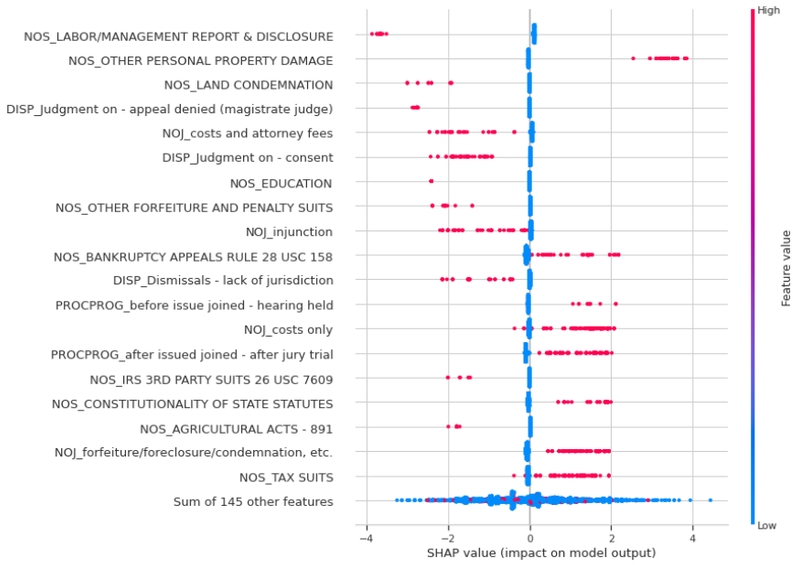
Table 6

4 Challenges and Future Work

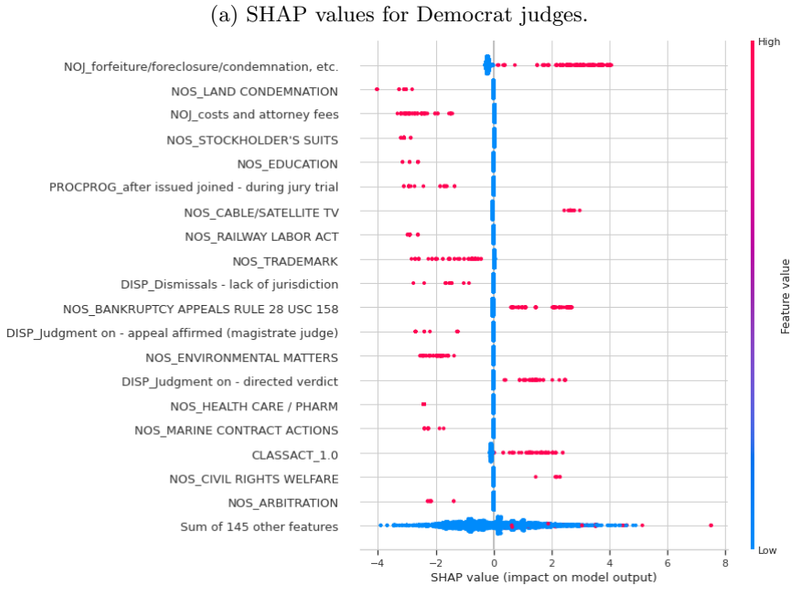
We initially faced some challenges parsing and preparing the data for analysis and omitted features like citations and opinions. We also left out a number of variables due to the high number of null values. Future groups may want to consider taking further steps to deal with data irregularities and will naturally want to include data from the corrupted zip file. They could also try different approaches for testing for randomization. For the feature analysis, they may wish to use textual features like the court opinion and citations to get a clearer picture of the types of cases where Democrats/Republicans do or do not rule in favor of the plaintiff. Rather than using our work as a baseline, perhaps they would see if a slightly different approach leads them to similar conclusions.

References

[1] Huigang Chen et al. *CausalML: Python Package for Causal Machine Learning*. 2020. arXiv: 2002.11631 [cs.CY].

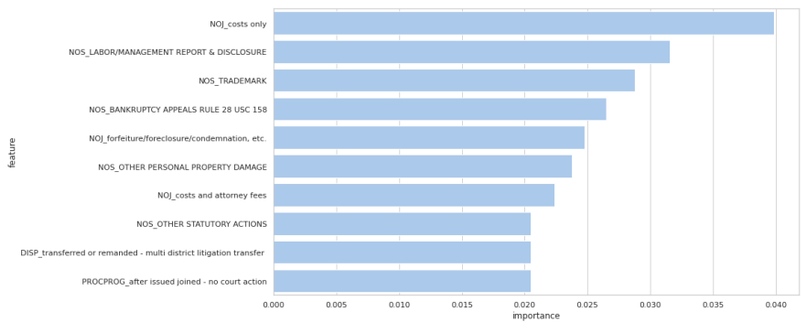


(a) SHAP values for Democrat judges.

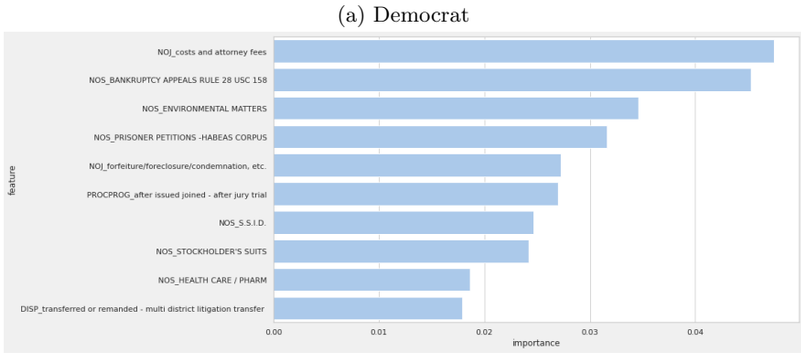


(b) SHAP values for Republican judges.

Figure 2: If a feature has red points on the right side of zero, the feature plays an important role in favoring the plaintiff. Otherwise, the feature plays a role in favoring the defendant.



(a) Democrat



(b) Republican

Figure 3: These plots show the importance each feature plays in an XGBoost model predicting the outcome of a case for either Democrat or Republican judges.