UNIVERSITY
OF MANNHEIM
School of Social Sciences

How to Forecast Constitutional Court Decisions?

Legal and Political Context in a Machine Learning Application

Sebastian Sternberg

June 21, 2019

University of Mannheim

Motivation I

Figure 1: A terrifying AI judge.

Motivation II

Field of quantitative legal prediction is emerging:

- Legal tech revolution in law.
- Judges become more and more "advised" by machine learning (ML) algorithms.
- Predictive modeling (mostly ML) is key here.
- Research in judicial politics knows little about predictability of law.

Natural question: can we forecast court decision outcomes?

Approach of this paper

This paper:

- Can a ML algorithm correctly predict the outcome of decisions of the German Federal Constitutional Court (GFCC)?
- Which new insights can be generated from predictive instead of inferential modeling?

Findings:

- 76.4% of over 2,900 proceeding outcomes can be correctly predicted.
- *Legal context* is a good predictor of court outcomes, but prediction can be further improved considering the *political context* of a decision.
- Predictive modeling is useful to generate new and substantial insights in judicial politics.

Existing Forecasting Approaches

Ruger et al. (2004):

- Prediction tournament of legal experts versus a simple ML algorithm predicting the October 2002 term of the "Rehnquist Court" (1994 to 2002).
- ML model was able to beat the legal experts.

Katz et al. (2017):

- Predict Supreme Court decisions over almost two centuries (1816-2015), forecasting 28,000 cases outcomes and more than 240,000 individual justice votes.
- Use over 75 different predictor variables, e.g. past voting patterns of judges.
- Correctly predict 70.2% of the case outcomes and 71.9% individual justice's votes.

Limitations of Existing Forecasting Approaches

Limitation 1: Focus on US Supreme Court:

- External validity of existing approaches questionable.
- Existing approaches use individual votes of judges for forecast; not feasible for many European courts.

Research question 1: Does a predictive approach already successfully applied to the Supreme Court also work in the European court setting?

Limitations of Existing Forecasting Approaches

Limitation 2: Evaluation of the contribution of legal context and political context variables to the prediction of court decision-making:

- Long-standing debate about which factors influence judicial decision-making.
- Some (legal scholars) emphasize the importance of jurisprudence and legal doctrine (legal context).
- Others (political scientists) argue that legal factors alone are not sufficient; political factors matter as well (political context).

Limitations of Existing Forecasting Approaches

- Many studies have evaluated the importance of legal and political context in an explanatory, but never in a predictive setting.
- **Observable implication**: if legal scholars are right, then legal context should be sufficient to forecast court outcomes. If political scientists have a point, then adding political context should improve the prediction.

Research question 2: do political context factors contribute to the prediction of court decision-making compared with legal context factor?

Case Selection: German Federal Constitutional Court

German Federal Constitutional Court (GFCC):

- GFCC is the archetype of the European constitutional court type.
- Role model for newly established court after 1990s.
- Hard case scenario: if we find evidence that political context matters for the GFCC, it presumably also matters for more political constitutional courts where the nomination procedure is more politicized (for instance, in France).

Data Set and Proceeding Types

Data set: Constitutional Court Database containing 2,910 proceedings decided between 1972-2010.

Three different proceeding types are considered:

- **Constitutional complaints:** can be filed by any person directly affected by a law or act.
- **Concrete Reviews:** can be filed by regular lower courts to review laws or statues.
- **Abstract Reviews/Organstreit:** often raise questions of fundamental political issues that are relevant for the political system.

Training a model on all these data sets at once would imply the same data generating process for them, which is unlikely.

Predictor Variables

Outcome variable:

- Binary outcome of proceeding, whether plaintiff was successful (=1) or not (=0).

Legal context variables:

- the *decision type*, the *issue area*, the *Senate*, the *legal area*, whether proceedings are *grouped together* or not.

Political context variables:

- the *ideological position* of the GFCC, the *salience* of a proceeding, the *popularity* of the opposition/government, and a measure for *public economic mood*.

Overall, I am rather over-inclusive in adding predictors to the model. ML does not have problems with correlated predictors

Method: Random Forests

Random forests (RF) (Breiman, 2001) is a popular supervised ML algorithm that combines the ensemble prediction of many (1,000) decision trees.

Why RF?

- Detecting non-linearities in the data without requiring the specification of any functional form.
- Provides built-in estimates of variable importance.
- Outperformed other learners on the prediction task.

Experimental Set-Up and Performance Evaluation

Experimental Set-up:

- For each of the three proceeding type data sets, two different random forests are developed: *legal model* only featuring legal context variables, and *combined model* featuring legal context and political context variables.

Performance evaluation:

- Aggregated cross-validated scores (without hyper-parameter tuning).
- Out-of-sample prediction: split data into training and test (Out-of-sample) set. Train on training set, evaluate on test set.
- Performance metrics: Accuracy (percentage correctly predicted) and Kappa (Kappa takes into account class imbalances)

Prediction Results using Out-of-Sample Prediction

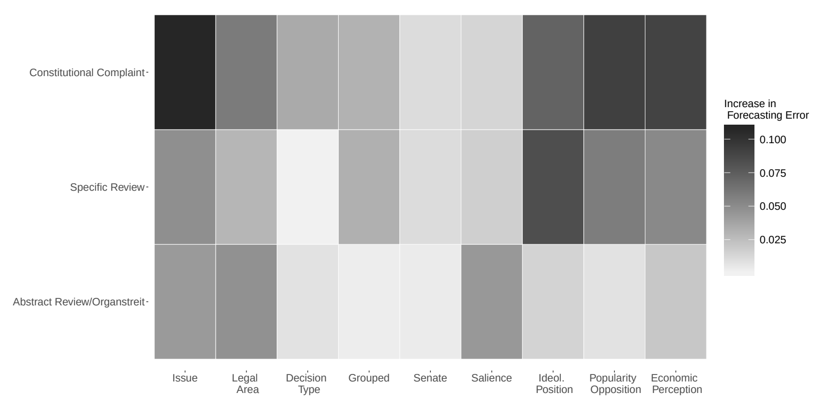
Table 1: Model Evaluation Based on Out-of-Sample Prediction

	Accuracy			Kappa	
	Legal	Combined	Baseline	Legal	Combined
Constitutional Complaint	66.67	**74.49**	52.67	0.33	**0.49**
Concrete Reviews	75.26	**81.05**	65.79	0.41	**0.57**
Abstract Reviews/Organstreit	60.38	**77.36**	58.49	0.17	**0.52**
Weighted Performance	68.47	**76.41**	56.52	0.34	**0.51**

Heatmap of Variable Importance per Proceeding Type

- Variable importance: measure for the mean increase in the prediction error if the values of a given predictor are randomly permuted.

Figure 2: Heatmap of variable importance per proceeding type



Partial Dependencies and Non-Linear Relationships in the Data

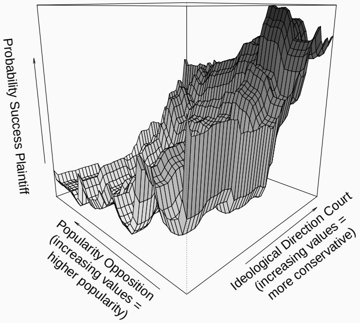


Figure 3: Partial dependence plot of ideological direction conditional on popularity opposition for concrete reviews.

Conclusion and Implications

GFCC application:

- ML algorithm can correctly forecast around three out of four proceeding outcomes.
- Similar methodological approaches used to forecast US Supreme Court decisions also work for European courts.
- Legal context is a good predictor of proceeding outcomes, but political context improves prediction even more.

Beyond the application:

- Sign of consistent judicial decision-making of the GFCC.
- Value of predictive modeling for social science: machine learning can help to identify patterns which conventional methodological approaches might overlook.

Backup Slides

Model Evaluation Based on Aggregated Cross-Validation Scores

Table 2: Model Evaluation Based on Aggregated Cross-Validation Scores

	Accuracy			Kappa	
	Legal	Combined	Baseline	Legal	Combined
Constitutional Complaints	60.14	**68.93**	53.47	0.20	**0.37**
Concrete Review	68.42	**80.18**	67.02	0.08	**0.50**
Abstract Review/Organstreit	63.54	**73.04**	60.26	0.19	**0.41**
Weighted Performance	62.55	**72.16**	57.50	0.17	**0.41**

Alternative Out-of-Sample Prediction

Splitting training and test by point in time:

- Previous split might violate iid assumption
- All observations before 2005 are assigned to the training set and all observations after 2005 are assigned to the test set.

Table 3: Model evaluation based on out-of-sample prediction using the time dimension for splitting

	Accuracy		Kappa		ROC AUC		PR AUC	
	Legal	Combined	Legal	Combined	Legal	Combined	Legal	Combined
BvR	59.33	**59.33**	**0.18**	0.13	**0.65**	0.62	0.74	**0.76**
BvL	75.26	**81.58**	0.41	**0.58**	0.74	**0.86**	0.64	**0.82**
BvE/BvF	51.51	**54.55**	-0.03	**0.06**	0.50	**0.41**	**0.50**	0.26

The Predictive Power of the Combined Model vs. White Noise

- Superior predictive power of combined only due to more variables (like increase in R^2 in regression)?
- Additional experiment where I replace the political context variables with white noise/random variables.
- RF is trained exactly in the same manner than the combined model before.

Table 4: Model Evaluation of Legal, Combined and Random Model based on aggregated cross-validation scores

	Accuracy			Kappa		
	Legal	Combined	Random	Legal	Combined	Random
Constitutional Complaints	60.14	**68.93**	62.75	0.20	**0.37**	0.24
Concrete Review	68.42	**80.18**	68.06	0.08	**0.50**	0.09
Abstract Review/Organstreit	63.54	**73.04**	66.58	0.19	**0.41**	0.26