How to Forecast Constitutional Court Decisions? Legal and Political Context in a Machine Learning Application

SEBASTIAN STERNBERG*

Ex ante forecasting approaches using machine learning become increasingly popular to analyze and predict judicial outcomes. Yet, existing work on the prediction of court decision-making has two important limitations. First, it exclusively focuses on the US Supreme Court. This raises concerns about the external validity of previous studies and their implications for courts in different law traditions. Second, none of the existing studies have explicitly tested the relative contribution of legal context versus political context factors to the forecast of court decisions. This study addresses these two points by ex ante predicting over 2,900 decisions of the German Federal Constitutional Court. I find that similar methodological approaches successfully applied to predict Supreme Court decisions also work for Kelsenian European constitutional court types. My results also show that the legal context of a decision is already a good predictor. However, the predictive performance is significantly improved when information about the political context of a decision is added. These findings therefore support the view of a multifaceted decision-making of constitutional courts which is best characterized by the ensemble of both legal and political factors

*PhD Candidate, Department of Political Science, University of Mannheim, A 5, 6, D-68131, Mannheim, Germany (sebastian.sternberg@gess.uni-mannheim.de).

1 Introduction

Is it possible to correctly predict decisions of the GFCC with an algorithm? And which factors are important for the prediction: legal context or political context factors? Algorithmic forecasting of court decisions is relatively new to the field. However, building upon recent efforts in applied machine learning, several studies already achieve impressive forecasting performances predicting US Supreme Court decision-making (Ruger et al., 2004; Guimera and Sales-Pardo, 2011; Katz et al., 2017b). Nonetheless, these studies have two important limitations. First, they exclusively focus on the US Supreme Court, which raises concerns about the applicability of these forecasts to other courts, and thus, the external validity of their findings. Second, none of the existing studies explicitly tests the relative contribution of legal context versus political context factors for the forecast of court decisions. There is a long-standing debate about which factors influence judicial decision-making. On the one hand, traditional legal scholars emphasize the importance of the legal and procedural context of a decision, while social scientists on the other hand also acknowledge the importance of the political context of a decision. Teasing out the relative importance of these factors improves our understanding of court decision-making from a predictive perspective.

The contribution of this study is to address these two limitations. First, I investigate *whether it is possible to correctly predict the decision-making of the GFCC using a machine learning algorithm*? I find that with a widely-used machine learning approach (random forests), on average it is possible to correctly predict 76.40 percent of the outcomes of over 2,900 proceedings decided by the GFCC between 1972 and 2010 using out-of-sample prediction. I also address the second limitation by explicitly teasing out the importance of variables associated with the legal and the political context of a decision. The key argument here is that if traditional legal scholars are right, then the legal and procedural context of a decision should be a sufficient predictor of court decision-making. However, if social scientists have a point, then including political context into the forecasting model should increase its predictive performance. For this reason, I also analyze *whether political context factors contribute to the prediction of court decision-making on top of legal context factors*? The results of my prediction show that the legal context alone is already

a good predictor of court outcomes. However, I find that forecasting performance can be further improved when the political context of a decision is additionally considered. I conclude that the ensemble of both legal and political factors is needed to characterize court decision-making. Furthermore, my results have important implications beyond the application to the GFCC : the value of predictive modeling for the field of social science.

2 Existing Approaches to Forecast Court Decision-Making

Forecasting the outcome of a court decision is a long-standing idea which originates from the very early stages of judicial politics research. "Legal prophecy", how Holmes (1897) termed it, has drawn considerable interest of scholars from various fields. Legal academics and political scientists have long scrutinized judicial decisions to understand what motivates courts and judges and how they arrive at a given outcome. These studies often look at past decisions and historical facts, e.g. individual judges' voting patterns, to *explain* why a certain court decided in a certain way. Most often, the goal is not to predict the outcome itself, but to use the causal connection between certain aspects of judicial decision-making to assess the consistency of some explanatory theory. Most of these studies use classic hypothesis testing, and are not interested in whether a model correctly predicts the outcome, but rather if certain estimates are statistically significant or not.

However, with the rise of artificial intelligence over the last decade, a new sub-field has emerged in judicial politics: the field of *quantitative legal prediction*[1] (Katz, 2013). In contrast to traditional causal inference approaches that make – at best – theory driven predictions about future outcomes, quantitative legal prediction focus entirely on the forecasting enterprise. Often, machine learning is the preferred method this. Machine learning in general is defined as "a subfield of computer science concerned with computer programs that are able to learn from experience and thus improve their performance over time" (Russel and Norvig, 2016, 693). The main purpose of

[1]This term was first introduced by Katz (2013). Quantitative legal prediction can be understood as an umbrella term for all different kinds of non-inferential, predictive approaches that aim at analyzing or predicting legal outcomes.

machine learning is to detect patterns and correlations in data and derive predictions about future outcomes. Not the explanatory but rather the predictive power of a variable is important here.

Over recent years, there has been a sharp increase in studies predicting the outcome of court decision-making with machine learning. In what follows, I will discuss the most prominent approaches. However, I will narrow my discussion only to approaches that actually employ an *ex-ante* forecasting approach. Ex-ante prediction is defined here as any prediction performed using information that is available prior to a judicial decision. Studies that use the texts of a decision to arrive at their predictions (e.g. Sulea et al., 2017; Medvedeva, Masha and Vols, Michel and Wieling, 2018) are excluded, since decision texts are typically not available in advance of a decision.

One of the first attempts to use machine learning to make ex-ante predictions about judicial outcomes dates back to 2004. In a seminal study, Ruger et al. (2004) held a prediction tournament in which known legal experts competed against a simple machine learning algorithm, a classification and regression tree (Breiman et al., 1984).[2] The goal of their work was straightforward: predict the votes of individual judges as well as the final decision outcome of cases referred by lower courts to the US Supreme Court in advance of the release of the Supreme Court's decision. Their machine learning model only relied on observable case characteristics such as the type of respondent, the type of petitioner, or the issue area of a case. Their model was trained on data from the "Rehnquist Court" (1994 to 2002), and then the predictive performance was tested on the October 2002 term. Known legal experts have also attempted to predict the same outcomes. The result of this prediction tournament is impressive: the simple machine learning algorithm already outperforms the legal experts by correctly forecasting 75% of all outcomes, while the human experts only forecasted 59% correctly. With respect to individual judges' votes, the model was correct in 66.7% of the cases wile human experts correctly predicted 67.9%.

As a follow up of this work, Guimera and Sales-Pardo (2011) investigate whether it is possible to make predictions of a justice's vote based on the other justices' votes in the same case by analyzing the voting behavior of each natural court between 1953 and 2004. They use the votes of

[2] A similar approach, but in a much richer setting, is currently undertaken by Katz et al. (2017a), where the authors test the predictive ability of a large crowd (a large group of humans) compared with experts and algorithms.

all judges in all previous cases, and the votes of the eight other judges in the current case to predict the vote of the ninth judge in the same case. They do not include any variables in their model, but solely rely on voting patterns. Their approach predicts 83% of the individual justice's votes correctly, but does not forecast the case level outcomes directly.

The work of Katz et al. (2017b) presents a major advance with respect to court prediction. The authors predict Supreme Court decisions over almost two centuries (1816-2015), forecasting 28,000 cases outcomes and more than 240,000 individual justice votes. Using random forests, a popular ensemble machine learning method and only relying on data available prior to the date of decision, Katz et al.'s (2017b) model correctly predicts 70.2% of the court's overall affirm/reverse decisions and correctly forecasts 71.9% at the individual justice vote level. A recent study builds on their efforts and improves the prediction to about 75%, leveraging an even more powerful algorithm (AdaBoosted decision trees) for the prediction (Kaufman et al., 2019).

3 Limitations of Existing Forecasting Approaches

All of these studies provide important insights about the predictability of court decision-making. However, I argue that existing forecasting approaches have two major limitations. First, existing ex-ante prediction models exclusively analyze and predict the US Supreme Court decision-making. This raises concerns about the external validity of previous work, and whether a similar prediction model could also be successfully applied to Kelsenian constitutional courts. In this regard, there are two issues. First, the US common-law system is guided by the norm of stare decisis, under which judges are supposed to decide cases based on similar precedents in the past. This leads to the expectation that just by how the legal system is constructed, there is supposed to be a high consistency between certain case-fact patterns. This "path-dependency" potentially facilitates the forecast, and might explain why even simple machine learning approaches (such as classification trees) already reach a high prediction accuracy (Ruger et al., 2004; Kastellec, 2010). As most European constitutional courts are under the civil-law system, there is no such thing as the norm of stare

decisis. In other words, a European constitutional court judge is formally less bound to past case outcomes when making her decision in a current case. This absence of "path-dependency" should make it potentially harder for machine learning algorithms to detect and identify patterns between certain factors and outcomes. Second, some of the previous studies use the past voting behavior of individual judges to obtain predictions (e.g. Guimera and Sales-Pardo, 2011). Unfortunately, this rich source of information cannot be leveraged for most European constitutional courts due to the non-disclosure of individual judges' votes. Both points raise concerns whether legal prediction models can also be successfully applied to European constitutional courts. The first question this study will answer is, therefore, *whether similar predictive approaches already successfully applied to the Supreme Court also work in the European court setting*?

Second, none of the existing studies have explicitly evaluated the relative importance of the predictors, namely the variables used for the prediction. There is a long-standing debate about which factors influence judicial decision-making, and thus assist its prediction. Although nowadays, the traditional divide between the two "camps" of legalists on the one hand and realists on the other hand is less clear and not as stark as it has been before, there remains considerable disagreement on which factors exactly are important for legal prediction. Traditional legal scholarship still emphasizes the important role of jurisprudence and legal doctrine, and tends to downplay the role of non-legal factors. According to this notion, judges find the solution to a legal question or the case outcome by neutrally applying law through legal reasoning and interpretative methods. To exaggerate, in this regard law works as a set of static, natural, apolitical rules that can be mechanically applied to decisions. Or, as Dyevre (2008) characterizes it: "rules + facts = decision" (Dyevre, 2008, 27). This traditional legal perspective remains strong in the European constitutional court context. The German legal scholar Ossenbühl (1998) for instance states that the jurisdiction of the FCC is a decision of dispute by means and guided by methods of law not political judgment (Ossenbühl, 1998, 85).

By contrast, legal realists and political scientists argue that legal factors alone are not sufficient to fully explain and predict judicial decision-making. Attitudinalists for instance argue that

judges are single-minded political actors whose decisions reflect their unconstrained policy preferences (Segal and Cover, 1989; Segal et al., 1995; Segal and Spaeth, 2002; Baum, 2009). Related, strategic accounts of judicial decision-making claim that judges are strategic actors who originally pursue policy-goals, but must adapt their behavior to external and internal constraints from other actors from time to time. Such constraints are, for instance, following public opinion to maintain their public support (e.g. Vanberg, 2005; Hall, 2014), or a strategic restraint from their own policy preferences in a separation-of-powers framework (e.g. Epstein et al., 2001; Bailey and Maltzman, 2011).

In this study, I do not aim to enter this (sometimes still) stylized debate. Instead, I want to explicitly test and tease out which factors actually contribute to the prediction of court decision-making. This idea is already noted in Martin et al. (2004), who write that "the best test of an explanatory theory is its ability to predict future events. To the extent that social science and legal scholarship seeks to explain court behavior, they ought to test their theories not only against cases already decided, but against future outcomes as well" (Martin et al., 2004, 761). Nonetheless, none of the existing ex-ante prediction models have explicitly teased out and quantified the contribution of variables belonging to different strands of argument.[3] In this context, predictive modeling offers an excellent possibility to compare competing theories of the same outcome (Cranmer and Desmarais, 2017, 149). In particular, focusing on the prediction of a phenomenon is a simple means to verify the extent to which "theoretically informed models anticipate reality, and which among those models does a better job of it" (Cranmer and Desmarais, 2017, 149).

For this reason, I will tease out the relative contribution of both political context factors and legal context factors for the prediction of court outcomes. I conceptualize political context factors as all factors that relate to the political aspects of court decision-making. Political context factors include the ideological position of the court, the public support for the court, the public opinion towards a certain issue upon which the court will decide, or any other political factor which social scientists have carved out in their work on judicial decision-making (see above). By contrast, legal

[3] Katz et al. (2017b) use variables belonging to legal and political context, but neither map their variables to these dimensions nor do they compare the variable's contribution.

context factors describe all non-political case characteristics associated with a case. These factors include the issue area of a case, the type of legal question that is raised, or the type of plaintiff or respondent. In other words, the legal context is rather understood as the legal baseline of a case in the absence of political factors.

Evaluating the contribution of political and legal context factors for the prediction of court decision-making can thus help us to gain a better understanding of court decision-making. The key argument here is that if traditional legal scholars are right, then the legal context of a decision should already be sufficient to predict a substantial part of court decision-making. Therefore, according to the pure legalist view, adding political context to the prediction should not improve the predictive performance of a forecasting model. However, if legal realists and social scientists have a point, the observable implication is that including political context into the prediction should increase the predictive power of the forecast. The second question this study thus addresses is *whether political context factors contribute to the prediction of court decision-making compared with legal context factors*?

To sum up, in this section I have discussed several existing ex-ante prediction models that use machine learning to forecast court decision-making. I have argued that there are two limitations in prior work: *a)* the exclusive focus on the US Supreme Court which raises concerns about the external validity of previous findings; and *b)* the lack of evidence that explicitly tests the contribution of both legal context and political context variables to the prediction of court decision-making. In the next section I present a research design that addresses these two limitations.

4 An Ex-Ante Prediction Model for GFCC Decisions

In this section, I present a research design for an ex ante prediction of the decision-making of the GFCC that is able to carve out the relative contribution of legal context and political context factors for the prediction. My design addresses the two limitations outlined before. I discuss why the German Federal Constitutional Court is an appropriate study object as a European constitutional

court, which data and variables I use to capture the legal context and political context of a case, and why I use the random forests algorithm for the prediction.

4.1 Case Selection: The German Federal Constitutional Court

The purpose of this study is to develop a forecasting model that *a)* predicts the decision-making of a constitutional court outside the US and *b)* compares the predictive contribution of legal and political context factors for the prediction. Here, the GFCC is analyzed. The case selection is motivated by three reasons. First, the GFCC is the archetype of the European Kelsenian constitutional court type and is considered as being one of the most powerful and influential constitutional courts world wide. It has served as a model for many newly established constitutional courts, e.g. in Eastern Europe. A prediction model that is suitable for the GFCC could also work as a blueprint for prediction models of these other courts. Moreover, the GFCC operates in a civil law system, and the individual votes of judges are mostly confidential. This means that one cannot simply predict individual judges' votes and aggregate them to make case outcome predictions. On these grounds, the GFCC represents a meaningful yet challenging study object from a predictive perspective. Third, the institutional power of the German court provides it with a strong institutional independence of other political actors, for instance with an appointment process of judges which requires a broad inter-party agreement. This makes it a hard-case scenario to test the importance of political context for the prediction: if we find evidence that political context matters for the GFCC, it presumably also matters for constitutional courts where the nomination procedure is more politicized (for instance, in France).

4.2 Data and Analytical Approach

The data used in this study were compiled as part of the Constitutional Court Database (CCDB) (Hönnige et al., 2015). The CCDB features 38 years (1972-2010) of data on decisions of the GFCC. Here, I use 2,910 proceedings (referrals) decided in this time frame. The court often bundles multiple proceedings in one main decision but decides on each of them individually (Wittig, 2016,

27). Thus, although being reviewed in the same main decision, the proceeding of petitioner A can be successful while the proceeding of petitioner B is not. I therefore follow common practice and treat the proceedings and their respective outcomes as the level of analysis (Hönnige, 2009; Sternberg et al., 2015; Engst, 2018).

The GFCC knows over 21 different proceeding types, which differ in the actors entitled to file an appeal, the possible causes of action, and also in their political importance and societal relevance. In my analysis, I concentrate on four proceeding types: *constitutional complaints*, *concrete reviews of statutes*, *abstract judicial reviews of statutes* and *Organstreit proceedings*. These proceeding types account for 98 percent of all proceedings decided by the GFCC. The proceeding types left out appear only rarely or are not a proceeding in the classic sense. Such proceedings include e.g. the procedure to impeach the Federal President.[4]

Constitutional complaints are the most common proceeding type (1941 proceedings in my data) accounting for around two-thirds of the observations in my data. Constitutional complaints allow citizens to assert their freedoms that are guaranteed by the constitution vis-à-vis the state, and can be filed by any person directly affected by a public law or act (after all other legal remedies are used). *Concrete judicial reviews* are the second most common proceeding type (760 proceedings in my data). They can be filed by regular lower courts to review laws or statues if they are unsure whether this law is unconstitutional or not. *Abstract judicial reviews*[5] are typically filed by political actors such as the parliamentary opposition, often challenging governmental laws or statutes. Although abstract reviews are relatively rare (121 proceedings in my data), they often concern matters of political nature and hold great political and societal importance (Kranenpohl, 2010, 260). These type of proceedings are also called the "sword" of the opposition (Schneider, 1974, 222). Finally, *Organstreit proceedings* (88 proceedings in my data) may be filed if high state organs, or actors that are equivalent to such organs, disagree on their respective rights and obliga-

[4]Official annual statistics provided by the GFCC can be found at https://www.bundesverfassungsgericht.de/SharedDocs/Downloads/EN/Statistik/statistics_2018.pdf?__blob=publicationFile&v=4, accessed 12.04.2019.

[5]In line with Hönnige (2009), I also code Bund-Länder-Streits, a vertical conflict of competence between the federal and the state governments, as abstract reviews due to their equivalence as regards content.

tions under the Basic Law. Similar to abstract reviews, they often raise questions of fundamental political issues that are relevant for the political system. Because abstract reviews and Organstreit proceedings only appear relative rarely, but are both considered as rather political proceeding types, I group them together in the analysis. Therefore, the final data contains three distinct data sets for each constitutional complaints, concrete reviews and abstract reviews/Organstreit proceedings.

Based on this data, I develop a separate prediction model for each of the three proceeding types, but with the same fixed set of predictors. This strategy is different to other court prediction models that rely on only one general model (Ruger et al., 2004; Katz et al., 2017b) for all different kinds of decision types. However, I argue that my approach has several advantages. First, using different proceeding types but the same fixed set of predictors allows me to compare the models with respect to their predictive performance and the contribution of the same variables in a different proceeding context. It is thus possible to test whether, for instance, political context variables contribute considerably more for the prediction of political proceeding types than for the prediction of proceedings without a political context. Second, developing one prediction model for all distinct proceeding types requires the assumption that the data generating process is the same across all types. This would be a strong (and potentially incorrect) assumption, given that the proceeding types strongly differ in their character. Finally, using one general model for all proceeding types would result in a heavy bias towards predictors that best explain constitutional complaints, as this type account for the majority of the data. The final model would hence not be a general model for all different proceeding types, but a model that is good for predicting constitutional complaints. In turn, this would not be beneficial to tease out the relative contribution of legal and political context factors across different proceeding types.

4.3 Outcome Variable

The outcome variable (dependent variable) is a binary variable indicating the individual outcome of each proceeding. This variable is coded as a one if the GFCC decided in favor of the plaintiff and it is coded as zero if it decides against it. In other words, it indicates whether the plaintiff was

successful or not. Following common practice, I consider a partial success to be a ruling in favor of the petitioner (Hönnige, 2007; Vanberg, 2005; Hönnige, 2009; Sternberg et al., 2015; Krehbiel, 2016, 2019). This binary coding scheme also allows me to compare my results with the findings of existing studies later.

In order to predict the outcome of a proceeding, I employ a number of predictors which represent the legal and political context of a proceeding. All of these variables can be used for an ex-ante prediction, since they all can be obtain *a priori* to a GFCC decision, and are thus exogenous to the final outcome. In fact, all information used for the prediction are publicly available the same day the plaintiff decides to submit the proceeding to the court. The model thus provides a substantial lead time.[6]

4.4 Legal Context Variables

I conceptualize the legal context of a decision as non-political, legal case characteristics associated with a decision. In other words, these factors should represent the "legal" or "procedural" baseline of a case. This baseline can then be used to compare the predictive power of political context in the subsequent assessment. Representing the legal context of a proceeding, I include the following variables: the *decision type*, the *issue* area a proceeding, the *Senate* who is supposed to adjudicate, the *legal area* a proceeding is concerned with and whether proceedings are *grouped together* or not. The decision type describes whether the decision is, for instance, a main decision or a provisional order. The issue variable describes the topic of a decision and is coded according to the Comparative Agenda Coding scheme (e.g. macroeconomic issues, social insurance). The *legal area* of a proceeding describes the legal doctrine a decision is related to, for instance family law or asylum law. However, it does not contain information on the exact legal norms the court examines in a decision, because from an ex-ante perspective this information is not available in advance of a court decision. All of those variables are taken from the CCDB. I did not include

[6]Lead time can be defined as the amount of time between a forecast is released and the actual occurrence of the event or outcome that is predicted.

information on the petitioner type or respondent type of a proceeding, because this information is already mostly covered by the proceeding type itself.[7] Table 1 provides a more detailed description of these variables with examples.

4.5 Political Context Variables

Political context is conceptualized as all factors that relate to the political aspects of court decision-making. The following predictors are used to represent the political context of a proceeding: the *ideological position* of the GFCC, the *salience* of a proceeding, the *popularity* of the opposition at the time of a decision, and a measure for the *perceived state of the economy* by German citizens as a measure of public economic mood. These political context factors are included because prior research of political scientists have found them to be important for the decision-making of the GFCC (Hönnige, 2007, 2009; Sternberg et al., 2015).

The ideological direction of the GFCC is measured on a common left-right scale using the Manifesto Common Space Scores (MCSS) (König et al., 2013). To calculate the position of the court, I use the common measurement approach first proposed by (Hönnige, 2007, 2009) by measuring ideological distance as the absolute ideological distance between the court and the government on a common left-right scale using the ideology scores from the Comparative Manifesto Project (CMP) (Laver and Budge, 1992). The position of the government is calculated by weighting the CMP scores of the governing parties with the respective number of seats of these parties in parliament. This allows for a more nuanced measurement of the government's policy position than using the raw CMP scores without weighting. The position of each Senate of the GFCC is measured by assigning each judge the CMP score of the political party that nominated him or her on the given day this judge entered the court. Subsequently, the mean position of each Senate is calculated. In include this variable because the importance of the GFCC's ideological position is demonstrated in previous work by Hönnige (2007, 2009). The following predictors are all related

[7] These variables are not supposed to represent *all* factors that legal scholars or legal traditionalist consider as being the most important factors of court decision-making. Rather, the legal context factors should serve as a baseline to compare the political context with.

to public opinion and public support.

The salience of a proceeding, namely its importance for the public is measured by a binary variable indicating whether a proceeding is accompanied by an oral hearing or not. Vanberg (2005) uses this variable as a proxy measure for the degree of the public awareness of a case, because "cases involving oral arguments are usually cases of great significance" (Vanberg, 2005, 103). Therefore, this variable is included as a political context factor because several studies demonstrate that the decision-making of the GFCC is affected by a proceeding's salience (Vanberg, 2005; Krehbiel, 2016, 2019). The popularity of the opposition captures the difference in the opposition's popularity relative to the popularity of the governments. This variable is included as political context variable because there is evidence that popular oppositions win their cases more often than oppositions with little public support (Sternberg et al., 2015). The data for this variable is taken from the German Politbarometer survey (Forschungsgruppe Wahlen, 2019). Finally, I capture the economic mood of the German public by measuring the perceived state of the economy. Evidence from the US Supreme Court shows that its decision-making is shaped by the economic state of the country (Brennan et al., 2009; Staudt and He, 2010). This could also be the case for the decision-making of the GFCC, although this causal relationship has not yet been tested. The economic mood variable is also part of the Politbarometer survey.

As an important note, I want to stress that model building and model specification is undertaken differently in predictive modeling than compared with classical inferential modeling. In predictive modeling, the inclusion of certain variables into a model is not guided by theory or expected causal relationships between the outcome variable and the predictors. Instead, generally all available information that could be somehow relevant for the prediction is included into a model, and the predictors are only modified to obtain a better prediction (to avoid over-fitting, for instance) or reduce computational burden (this process is called feature engineering in the machine learning literature (Hastie et al., 2009)). Following this, I did not make a specific effort to pare down the list of legal or political context variables, and that there is no doubt that some of them are correlated. Nonetheless, this (some would call it "kitchen sink") approach is not problematic for my analy-

sis. The machine learning method I use does not suffer from the same problems that conventional regression analysis has with correlated predictors. Therefore, I am rather over-inclusive in adding predictors to the model. All variables are once more summarized in Table 1.

Table 1: Legal and Political Context Variables Used for the Forecast

Legal Context	**Description**	**Example**
Decision Type	The type of the decision	Main decision, preliminary ruling
Issue	Issue area (Comparative Agenda Coding Scheme)	Macroeconomic Issues
Senate	Senate dealing with a proceeding	Senate I or II
Legal Area	Legal area a proceeding is concerned with	Labor law
Grouped	Whether a proceeding is grouped with others or not	0 = not grouped, 1 = grouped
Political Context		
Salience	Whether there was an oral hearing before the proceeding	0 = no oral hearing, 1 = oral hearing
Popularity Opposition	Difference in popularity of opposition relative to government	1 = very unpopular, 11 = very popular
Economic Perception	Perceived state of the economy	1 = very good, 5 = very bad
Ideological Direction	Ideological direction of the Court (MCSS scores)	-1 = left, 1 = conservative

4.6 Method

To build my prediction models I rely on random forests (Breiman, 2001). Random forests is a popular ensemble classifier and is among the most commonly used machine learning algorithms for supervised learning (Hastie et al., 2011). Although random forests and similar tree-based methods were long neglected by the field, they become increasingly used in the social science context (e.g. Green and Kern, 2012; Beauchamp, 2017; Montgomery and Olivella, 2018; Jones and Lupu, 2018; Bonica, 2018; Kaufman et al., 2019). In what follows, I give a brief introduction to random forests. For a recent, non-technical introduction of tree-based methods for political scientists see Montgomery and Olivella (2018).

A random forest uses an ensemble of classification and regression trees (CART). CART is a supervised machine learning algorithm that iteratively divides the outcome variable observations into increasingly homogeneous groups using the predictor variables through binary splits (this is called recursive partitioning). CARTs are known to be notoriously unstable, meaning that already small changes in the data can lead to completely different splits. They also tend to be biased towards continuous covariates (Hothorn et al., 2006). A random forest overcomes these limitations by using an ensemble of many randomized trees that leverage two forms of randomness: *bagging* – short for *b*ootstrap *agg*regation – (Breiman, 1996) and *random substrates* of the predictor variables. The underlying idea is that many uncorrelated trees are constructed and then aggregated. The procedure to construct one (out of many, typically between 500 and 1,000) tree in random forest is as follows.

First take a random sample with replacement, typically containing about two-third of the observations, while the remaining (one-third) of the observations are hold "*out-of-bag*" (*oob*). On the bootstrapped sample, construct a decision tree. At each node of the tree, randomly select m out of p predictors, where m is a hyper-parameter and is typically chosen by the researcher. Out of these m randomly selected predictors (random substrates), the one that gives the best classification at this node is used to partition the data. This process is repeated at each subsequent node, such that at each node a random substrate of m predictors is chosen. The random selection of splitting

variables allows predictors that were otherwise outplayed by their competitors to enter the ensemble. This has the benefit of obtaining less correlated and thus, more robust trees. The model then averages predictions over all trees, whereby the predicted class of an observation is calculated by majority voting of the *oob*-predictions for that observation. In Appendix A I outline the random forest algorithm in further detail.

There are four reasons to use random forests and not another machine learning classifier. First, random forests has proved to be a strong learner in a comparable study (Katz et al., 2017b). Second, in an analysis of judicial decisions and legal rules using a single decision tree, Kastellec (2010) finds that the tree structure actually mirrors the "hierarchical and dichotomous structure that often seems apparent in judicial opinions" (Kastellec, 2010, 210). Third, an experiment using several popular classification algorithms shows that random forests outperforms other algorithms.[8] Fourth, random forests is very efficient in detecting non-linearities in the data without requiring the specification of any functional form and also provides built-in estimates of variable importance. All of these aspects make random forests the optimal method choice for my prediction task.

5 Results

In this section I present the results of the ex-ante prediction of proceedings decided by the GFCC. The section is divided into two parts. In the first part, I use random forests to predict the outcomes of each proceeding type in my data using the same fixed set of input variables. I show that a combined model consisting of legal and political context variables yields to a higher predictive performance than a model using legal context factors alone. Moreover, I conduct a simulation that shows that the increase in predictive power is not just an artifact of adding more variables to the model. In the second part, I open the black-box of the prediction model by comparing the

[8] I test the predictive performance of Classification and Regression Tree (CART), Random Forests, Support Vector Machines, k-nearest neighbors, extreme gradient boosted trees and regularized logistic regression on the constitutional complaints data. Predictive performance was assesses using 10-fold cross-validation without hyper-parameter tuning. Cross-validation was performed such that every algorithm received exactly the same data slices, to make the model comparison as fair as possible. This constitutional complaints data set is used because it has the largest N. The classification results are found in the Appendix C.

predictive importance of the predictors across the proceeding types. The section concludes with a discussion of the ability of random forests to detect interesting non-linearities in the data that conventional regression analysis might have overlooked.

5.1 Predicting Proceeding Outcomes of the GFCC

In order to tease out the relative importance of legal and political context for the prediction of GFCC decision-making, I run a series of experiments. For each of the three proceeding type data sets, two different random forests are developed: a *legal model* only featuring the legal context variables, and a *combined model* featuring the legal context *and* political context variables. The legal model here serves as a "legal" baseline and is used to evaluate the predictive performance one can expect by just using the legal and procedural context of a given proceeding. The combined model is used to assess whether and to what extent political context can improve the model's predictive power. To repeat, the observable implication with respect to this comparison is that if legal realists and political scientist are right by arguing that political context matters, then the inclusion of this context into the prediction model should increase its predictive capability. If political context is irrelevant for the prediction, then its inclusion should not change model performance. At this point I want to highlight again that my analysis does not seek to disentangle the causal effect of legal and political context on judicial behavior, nor to test whether political context outweighs legal context.

For a fair model comparison, a robust model performance evaluation is of crucial importance. In predictive modeling, the goal is to obtain an estimate of *true error* (also known as *generalization error*). True error is a measure of how well a model can predict outcomes of previously unseen data (Efron and Hastie, 2016; Cranmer and Desmarais, 2017). An estimate of true error is important in practice, as it allows one to check whether a model generalizes well to unseen data or just memorizes the patterns in the training data (i.e. over-fitting).

With this in mind, I provide two performance evaluations of the models. In the first performance evaluation, I report the model's performance based on their aggregated cross-validation

score *without hyper-parameter tuning*. Cross-validation, when correctly applied, can be used to obtain an almost unbiased method of true error without setting aside additional test data (see Cawley and Talbot, 2010; Efron and Hastie, 2016). However, note that combining cross-validation for model tuning and to estimate true error at the same time leads to serious misreporting of performance measures (Neunhoeffer and Sternberg, 2019).

In my experiments, on each of the three data sets[9], I perform (stratified) 10-fold cross-validation and hold only the hype-parameter of random forests fix at $m = \sqrt{p}$, where p is the number of predictors and m is the number of random substrates. This value is recommended by Hastie et al. (2011) for classification problems using random forests (Hastie et al., 2011, 592). In short, cross-validation refers to randomly dividing a data set into K about equally sized folds, where each fold contains about $\frac{N}{K}$ observations. A random forest classifier[10] is then trained K times on all but the kth fold, where k runs from 1 to K. In every iteration, a performance measure is used to evaluated the model performance on the kth fold (holdout/test fold) that was not part of the training. Finally, the average (across the K folds) of a performance measure is reported, which is the aggregated cross-validation score. However, as Cawley and Talbot (2010) show, even if cross-validation is applied correctly, the variability of such hold-out methods can lead to overfitting in a finite sample nonetheless (Cawley and Talbot, 2010, 2084-2086). This, in turn, would lead to reporting an overly optimistic model performance.

For this reason, I report the results of an *out-of-sample* prediction as a second evaluation. Out-of-sample prediction is considered as the gold standard to obtain an unbiased estimate of true error (Hastie et al., 2009, 220). In out-of-sample prediction, a model is trained on a training set and then used to predict the observations of a test set (the out-of-sample data). During the training process, hyper-parameter tuning can be performed. This is because due to the strict split between

[9] I only use the training data sets (see next paragraph) to obtain the cross-validated performance scores. This ensures that each model only has access to exactly the same amount of information. Taking the cross-validation scores of the whole data set would constitute an unfair model comparison, because then models of the cross-validation procedure would have seen more data than the models of the out-of-sample evaluation.

[10] The random forests are estimated using the *R* packages *caret* (Kuhn, 2008) and *ranger*, a fast (parallel) implementation of random forests (Wright and Ziegler, 2015). For each random forests, $1,000$ trees ($ntree = 1,000$) are grown because simulation studies suggest that smaller values can result in unstable estimates under certain circumstances (Strobl et al., 2007, 2009).

training set and test set, the final model evaluation cannot suffer from over-fitting since the test set never occurred in the model building process.[11] For each of the three proceeding data sets, I randomly divide the data into a training set, containing 75 percent of the observations, and a test (out-of-sample) set with the remaining 25 percent.[12] On each of these training data sets, I train two random forests models: one using only the legal context variables, and one using both. Tuning is performed to find the best set of hyper-parameters using five-fold cross-validation and random grid-search. These models are then used to predict the outcomes of the observations in the test set.

As performance metrics, I report the accuracy and Cohen's $Kappa$ (Cohen, 1960). Accuracy is simply defined as the sum of true positives and true negatives divided by the overall number of observations. The $Kappa$ metric takes into account the class distributions and is based on the observed accuracy (accuracy of the classifier) and the expected accuracy (expected accuracy of a random classifier). In Appendix D, I report additionally the receiver operating characteristic area under the curve (ROC AUC) and the precision recall area under the curve (PR AUC).[13] In order to calculate the accuracy, the conventional threshold of 0.5 is used for positive predictions. The majority class (baseline) is also reported to compare the performance of the random forest with respect to a naive learner. A naive learner is defined here as a classifier who always assigns the majority (most frequently occurring) category of the training set.[14]

Table 2 reports the model evaluations based on the aggregated cross-validation scores across the three different data sets. All columns labeled as "legal" report the performance of the legal model and all columns labeled as "combined" report the performance of the combined model. The corresponding confusion matrices of each model are provided in Appendix E. The best models according to the respective performance measure are highlighted in bold. We see that the legal model itself is already sufficiently good to predict a substantial part of all decisions correctly, outperforming the baseline for all proceeding types. The weighted accuracy across all proceeding

[11] Of course, a model can over-fit the training data, although the over-fit will lead to a poor out-of-sample prediction.

[12] The randomly created training and tests sets are of the following sizes (N of training set, N of test set): Constitutional complaints 1,455, 486; concrete reviews 570, 190; Abstract reviews and Organstreit proceedings 156, 53.

[13] The calculation of all performance metrics is defined in Appendix B.

[14] Note that it only makes sense to report the baseline for accuracy. This is because the *kappa* measure already takes into account the majority class in its calculation (*kappa* = 0 means that majority voting takes place).

types is 62.55 percent.[15] Using the weighted accuracy is important to obtain the overall percentage of correctly-predicted proceedings, since the proceeding data sets are of different sizes.

However, and this is the important observation, we also see that for all proceeding types, the model performance is improved when the political context variables are added (combined model). Across all proceeding types, the weighted accuracy improves to 72.16 percent. This means that using the combined model, it is possible to correctly forecast approximately three out of four outcomes. On average, across all proceeding types, adding the political variables to the classifier increases the predictive performance by about 9.61 percentage points in terms of weighted accuracy and 0.24 in terms of $Kappa$. The higher $Kappa$ values of also indicate that the better performance of the combined model is robust when considering the class distributions. The largest performance increase is for concrete reviews, where the addition of political context improves the predictive performance by +11.76 percentage points in accuracy. We can also see that the performance is considerably increased for the political proceeding types (abstract review/Organstreit proceedings): here, the addition of the political context variables improves the prediction from approximately two out of three to correctly predicting around three out of four outcomes (+9.5 percentage points). This finding makes intuitively sense from a political science perspective: these proceeding types often deal with political matters, so that the potential influence of political context is expected to be strong here.

These results are also confirmed when looking at model evaluation using out-of-sample prediction in Table 3. We, again, observe that for all different proceeding types, the combined model has a higher predictive power than the legal model. Across all proceeding types, the weighted accuracy of the combined model is 76.41 percent, and thus about +7.94 percentage points better than compared with the legal model (68.47 percent weighted accuracy). Here, the performance improvement is the highest for the political proceeding types (+16.98 in accuracy). Note that both model performances (the legal and the combined model) have higher scores when using the out-

[15] Calculated by weighting the accuracy of the respective proceeding type with the number of observations of this type.

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**

Note: Model performances of the legal model and the combined model based on the aggregated 10-fold cross-validation scores. The random forests were built with a fixed m. The legal model only uses legal context variables, while the combined models used both legal and political context variables. The baseline category for accuracy is a naive classifier that always votes the majority category of the training set. The best performances are highlighted in bold.

of-sample prediction evaluation. This is the case because although I split the data randomly into training set and test set for the out-of-sample prediction, due to random chance we observe differences between the cross-validation scores and the out-of-sample scores (different splits of training and test set might result in different scores). These differences are so strong because the overall N of the data sets is not very large (the abstract review/Organstreit proceedings data set only contains 209 observations overall). At this point, I want to emphasize that one should not over-interpret the exact performance scores, but that my findings rather demonstrate a general tendency independent of the performance evaluation approach: on average, adding information about the political context of a proceeding improves the prediction.

5.2 The Predictive Power of the Combined Model Versus White Noise

A critical reader might wonder whether the improvement of predictive performance that we observe when adding the political context variables to the legal model is due the predictive power of these variables or due to simply adding more variables (like the expected increase in R^2 in the regression context). In order to convince such critical voices and demonstrate that the political context variables actually improve the predictions because they are related to the outcome of a

Table 3: 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**

Note: Model performances of the legal model and the combined model based on out-of-sample prediction. The legal model only uses legal context variables, while the combined models used both legal and political context variables. The baseline category for accuracy is a naive classifier who always votes the majority category of the training set. The best performances are highlighted in bold.

decision, I conduct an additional experiment. In this experiment, I substitute the four political context variables with randomly drawn variables unrelated to the outcome. I refer to these randomly drawn variables as "*noise features*" in the following.

The noise features are constructed by randomly sampling from a multivariate normal distribution with means equal to the means of the original variables and the corresponding variance-covariance matrix to capture the structure of the variables to each other. Accordingly, these randomly-sampled variables mirror the distribution and correlation structure of the original variables, but are not correlated with the other features or the outcome. For each proceeding type data set, I remove the original political context variables and replace them by noise features. The final data sets thus only include the legal context variables and the four noise features. On each of these data sets, I then run random forests models using the 10-fold cross-validation procedure without hyper-parameter tuning which has already been used to obtain the aggregated cross-validation scores reported in Table 2. I call these models the "*random models*", due to the four randomly created noise features in it. The observable implication is that if the performance of the combined models in the main analysis just improves because more variables are added, then we should also observe an increase in the predictive performance of the random models, although the noise features should have no predictive power by construction. However, if the political context variables

actually contribute to the prediction, we should observe the combined model to perform better than the random model.

Table 4 reports the result of this experiment. We can observe that the combined model still performs better than the other two models. In fact, the performance scores of the legal model and the random model are about the same for constitutional complaints and concrete reviews. Interestingly, for abstract reviews and Organstreit proceedings, the random model is about three percentage points better than the legal model. For these proceedings, adding "white noise" improves the prediction, although not as much as the original political context variables. A possible explanation for this is provided by Bishop (1995), who shows that adding noise to data can have a similar effect like l_2 regularization if the predictive method is over-fitting. However, interestingly Bishop (1995) describes that the random noise is added by "adding a random vector onto each input pattern" (Bishop, 1995, 109). In simple terms, this means that for each individual data point of some features X_1, X_2, X_3, random noise is added like $X_1 + z, X_2 + z, X_3 + z$, where X represents the original predictors and z is a random noise vector. By contrast, what I do is adding extra noise features, so that the data used for the prediction then is like $X_1, X_2, X_3, Z_1, Z_2, Z_3$, where each Z represents a randomly created noise feature.

In fact, the improvement of prediction by adding white noise for the political proceedings data might be a hint that the legal model in Table 2 over-fits, and therefore the addition of random noise makes it harder for the random forests to over-fit the data. I obtain similar findings when using out-of-sample evaluation instead of the aggregated cross-validation scores (Appendix Table F) and when replacing the draws from a multivariate normal distribution with draws from a standard normal distribution (such that the four added randomly sampled noise features are not related to each other at all).

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

Note: Model performances of the legal, combined and random model based on aggregated cross-validation scores. The legal and combined models are the same as in Table 2. The "random model" only contains the legal context variables plus four randomly created noise features. The best performances are highlighted in bold.

5.3 An Alternative Out-of-Sample Prediction

In order to further demonstrate the robustness of my findings, I provide an additional out-of-sample prediction in Appendix G where I take into account the time dimension of the data. Randomly dividing the data into training set and test set requires assuming that the data is *iid* (independent and identically distributed). The iid assumption might be violated using data with a clear time dimension (the data set covers 1972 to 2010). For this reason, I split the data into a training set and a test set where all observations before 2005 are assigned to the training set and all observations after 2005 are assigned to the test set. This test set is then used for the out-of-sample prediction. I did not use this split approach in the main analysis because splitting by an (arbitrary) point in time results in different train/test set size ratios. To illustrate, due to the split in 2005, the test set of abstract reviews/Organstreit proceedings contains around 19 percent of all observations (33 observations of 209), while the test set of the constitutional complaints contains only 8 percent allow of the observations (150 observations out of 1,941). This is because the number of proceedings decided by the GFCC is not equally distributed over time. Accordingly, a fair model comparison is difficult because the information each classifier has access to differs in terms of percentage of the overall data. Nonetheless, using the additional out-of-sample prediction the patterns of the main analysis are confirmed: adding political context improves the prediction of GFCC decision-making.

The results of this section lead to several conclusions. First, the findings for the US Supreme

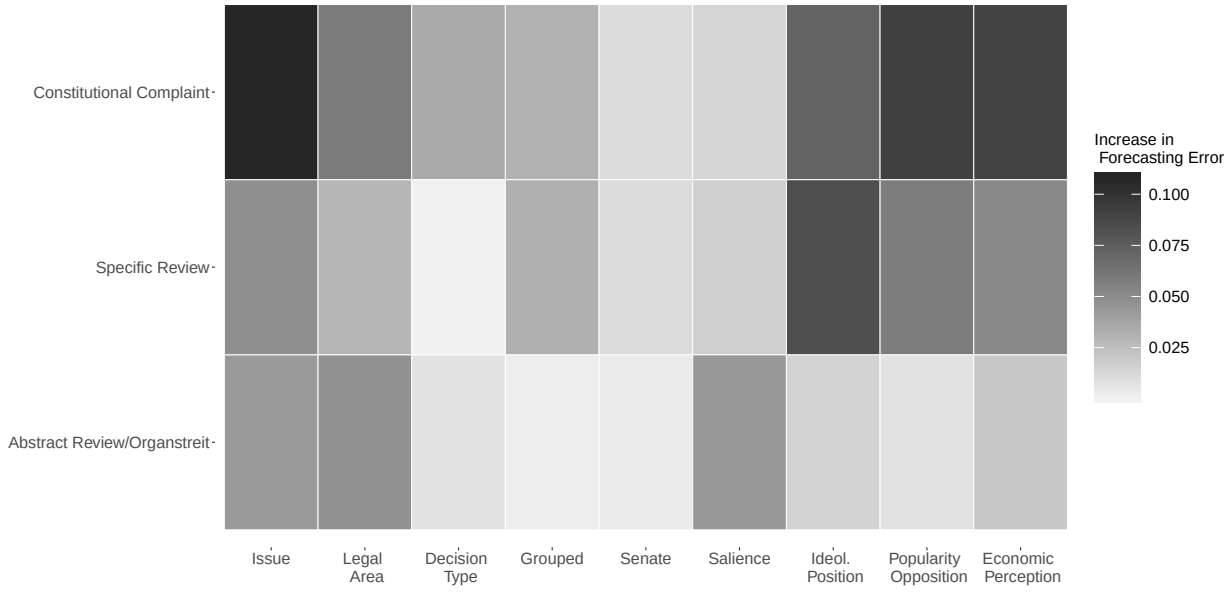
Court – that a machine learning model can successfully predict judicial decision outcomes – can be generalized at least to the German Constitutional Court, an archetype of the Kelsenian European Constitutional Courts. Similar machine learning approaches can reach similar accuracies. Across all proceeding types, the weighted accuracy of the combined model is 76.41 percent (out-of-sample prediction) and 72.16 percent (aggregated cross-validation scores). This is very close to the achieved performances of Ruger et al. (2004) with 78% and better than the achieved 70% of Katz et al. (2017b), who use over 95 predictors and heavy feature engineering. The first research question of this study – whether a machine learning classifier can correctly predict GFCC decision outcomes – is thus to be answered with a clear yes.

Second, I also find evidence that political context (including public opinion) improves the prediction of all proceeding types, and thus support for the second research question – whether political context factors contribute to the prediction of court decision-making compared with legal context factors. This is a strong and interesting finding, because a part of the German legal scholarship still considers the GFCC's decision-making as totally apolitical. (Böckenförde, 1976; Ossenbühl, 1998). I want to emphasize again that this does not mean that political context outweighs the importance procedural characteristics or other legal aspects of a proceeding. Instead, just the ensemble of legal and political variables collectively contributes to the prediction in the combined model. To further investigate the role of legal and political context, I look at each variable's importance for the forecast in the next section.

5.4 The Importance of Legal Context and Political Context

Which of the variables contribute to the prediction? Is there any variation in their importance across proceeding types? The importance of a variable in random forests can be obtained via its *variable importance*. Variable importance (also known as permutation importance) is a measure for the mean increase in the oob error if the values of a given predictor are randomly permuted. The idea behind this is straight forward: If the values of a predictor are randomly permuted and the oob error remains constant, the predictor is regarded as unimportant. By contrast, the larger

Figure 1: Heatmap of variable importance per proceeding type



Note: The different predictors are displayed on the horizontal axis. The different types of proceedings are shown on the vertical axis. Darker fields indicate a higher importance of the respective predictor for the respective proceeding type. The variable importance is obtained from the combined models from Table 3 to enable a comparison of legal context and political context predictors.

the increase in oob error when a predictor has been permuted, the more important this predictor is for the forecast (Hastie et al., 2009, 593). Figure 1 shows the variable importance of all predictor variables on the horizontal axis with the respective proceeding type on the vertical axis of the heatmap. The darker a cell in a heatmap, the higher the variable importance of the given predictor for the respective proceeding type. The forecasting error of constitutional complaints increases, for instance, by about six percent if the values of the issue variable are randomly permuted, and thus withheld from the prediction.[16]

Figure 1 shows a considerable variation in the predictor's importance across the proceeding types. There is not a single predictor that is of equally strong importance for all proceeding types. The issue of a decision is a important predictor for constitutional complaints and concrete reviews, but not so much for abstract reviews/Organstreit proceedings. Some issues seem to be

[16]For the sake of terminology, it is important to note that oob variable importance does not measure the increase in forecasting error if a certain predictor is excluded from the model. This is because if the model was rebuild without this predictor, the model could put more emphasis on other predictors, which then became surrogates (Hastie et al., 2009, 593).

especially important in this regard. Not knowing whether the issue "education" or "law and crime" is present in a constitutional complaint proceeding, for instance, increases the forecasting error by about 1.8% and 2.1%, respectively (not shown in the graph). Interestingly, the ideological position of the GFCC is the most important predictor for concrete reviews, but not so important for the political proceeding types. In line with what we would expect theoretically, we also observe that the political context variables contribute more than the legal context factors to the prediction of these political proceeding types. Again, I want to highlight that variable importance is not equivalent to a causal relationship between a predictor and the outcome variable.[17] Nonetheless, it can help us to gain a deeper understanding of the factors which drive the prediction, and can hint towards interesting relationships. In the next section, I will look at how certain predictors increase the winning chances of the plaintiff, which is something we cannot infer from variable importance plots. This information is contained in partial dependence plots.

5.5 Partial Dependencies and Non-Linear Relationships in the Data

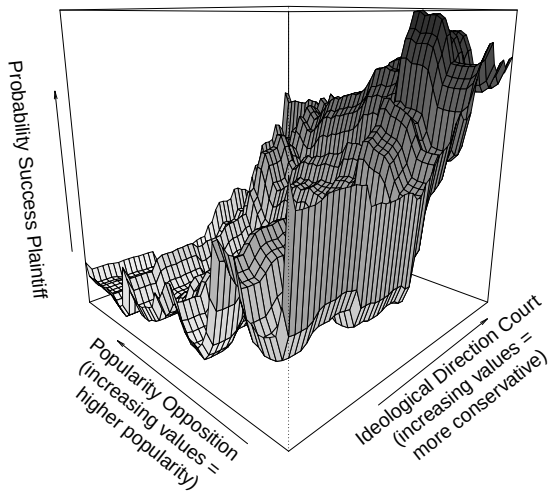
Partial dependence plots are a method to visualize the partial relationship between predictors and the outcome in forecasting models. In short, such plots give a graphical representation of the marginal effect of a variable on the predicted outcome, after accounting for the average effects of the other predictor variables (Hastie et al., 2011, 369).

Figure 2 shows the partial dependence plot for the interaction between the ideological direction of the GFCC and the popularity of the opposition on the probability of a petitioner success. I focus at this interaction because the variable importance plot in Figure 1 shows that these variables are important predictors of concrete reviews. Moreover, these are political context variables that hold the most importance for predicting a rather apolitical proceeding type. Thus, it is a surprising finding that warrants further investigation.

We can draw several conclusions from the partial dependence plot. First, there is a negative

[17] In addition, some of the predictors are correlated which can complicate the interpretation of the variable importance (Strobl et al., 2007, 2008).

Figure 2: Partial dependence plot for ideological direction of the GFCC conditional on the popularity opposition/government on the plaintiff's success probability for concrete reviews



Note: Partial dependence plot for the interaction between the popularity of opposition/government and the ideological direction of the GFCC on the probability of plaintiff success in concrete reviews. The combined model for concrete reviews from Table 2 was used for the calculation. The graph shows a clear non-linear relationship between the outcome and the two predictors.

association between oppositional popularity and petitioner success, indicated by the flat surface in lower left part of the figure. However, this effect is conditional on the ideological direction of the GFCC: the more conservative the GFCC, the higher the likelihood of a petitioner's success (indicated by the sharp rise in the upper right). In other words, the winning chances of a petitioner in this scenario are the lowest if the opposition is very unpopular and the GFCC is rather left, whereas the winning chances are the highest if public support for the opposition is low and the court is rather conservative. This is an important observation, because these results suggest that the rather apolitical proceeding types such as concrete reviews might not be per se as apolitical as one thinks. Second, and more important, the effect between the two predictors on the outcome is clearly non-linear. This non-linearity would not be captured by conventional approaches such as logistic regression, at least not without specifically specifying the functional form of this relationship in the systematic component. Machine learning approaches such as random forests learn these non-linearities in the data without the that they have to be pre-specified by the researcher.

Partial dependence plots of other predictors show that the directions of how these variables are related to other variables or the outcome are largely as one would expect. The salience of a proceeding, for instance, strengthens the effect of other political context predictors such as the ideological position of the GFCC. This is in line with existing political science research showing that judges behave differently in salient than in non-salient cases (Vanberg, 2005). Furthermore, the perception of the current state of the economy by the public plays a greater role if the main issue and sub issue of a case is an economic one. This is a relationship that makes intuitively sense. One of the important lessons of this study is that predictive modeling can help researchers to find (non-linear) relationships which conventional methodological approaches might have overlooked. In fact, most of the relationships between inputs and outcome do not display the typical S-shaped curve of e.g. logistic regression models, the model which is most often used to analyze binary outcomes. Machine learning approaches are, therefore, a fruitful approach to identify interaction effects or other non-linearities in the data.

6 Conclusions and Implications

In this study, I highlighted the ability of machine learning to ex-ante forecast decisions of the GFCC. I demonstrated that it is possible to correctly predict 76.40 percent of all outcomes of over 2,900 GFCC proceedings decided between 1972 and 2010 using only data that is available prior to a proceeding. In particular, I did not use any information which stems from decision texts, court statements or press releases or any other source that only becomes available after the actual decision outcome is released. Such a forecasting model is a novelty in European court research, and does not yet exist for the GFCC or any other European constitutional court.

I make two contributions. First, I confirm the external validity of similar work on the US Supreme Court and show that the decision-making of a European Kelsenian Court type can also be correctly forecasted by means of an algorithm. This is an important result, because the predictive setting for most of the European courts is more challenging since no individual voting records of

justices are available. Second, and this is unique to my analysis, I explicitly test the predictive contribution of legal context and political context variables to the forecast. I find that legal context is, on average, a relatively good predictor proceeding outcomes. Moreover, I find that the predictive performance is improved when the political context of a decision is leveraged. Constitutional court decision-making is thus best characterized by the ensemble of legal and political context factors.

Beyond the application to the GFCC, my findings have other important implications with respect to legal philosophy and the value of machine learning approaches for the field of judicial politics and political science in general. What does it mean for our understanding of law and judicial decision-making if a relatively simple machine learning algorithm can correctly predict a substantial number of judicial outcomes? While this might appear alarming first, I argue that in fact, this is a sign of consistent judicial decision-making of the GFCC. If an algorithm can correctly predict outcomes, it means that on average, similar proceedings with similar case characteristics are decided in a similar way. This consistency in judicial decision-making is important for the basic functioning of the rule of law. Therefore, for the sake of legal certainty, it is desirable that cases with the same context lead to the same judicial outcomes on average. Moreover, no algorithm could in any way substitute for the important work that judges do in their reasonings.

My findings have another implication for an important group beyond academia: the world of plaintiffs before the GFCC. For lawyers, politicians or ordinary citizens, the expected outcome of a case, namely the (perceived) probability of winning or losing, plays a crucial role in a plaintiff's decision to appeal or not. Given that a predictive model of GFCC decision-making can be improved over time and with more and possibly richer data, my results are beneficial for practicing attorneys and their clients likewise. In fact, such a model would also have consequences for the political system: for instance, the opposition would not only consider political factors in their decision to appeal to the GFCC or not, but would also be able to refrain from appealing cases where the success probability is low.

Finally, my analyses demonstrate the value of predictive modeling for social science: machine learning can help to identify patterns which conventional methodological approaches might

overlook. This is especially important with respect to non-linearities in the data. Thus, even when the goal is causal inference, such forecasting approaches can help to identify undiscovered patterns in the data and therefore, can lead to new research questions. What is the causal mechanism that links the perception of the economic shape in Germany to its outcome? Why is the ideological position of the GFCC the most important predictor for concrete reviews, a proceeding type most often only dealing indirectly with political matters. While I do not argue that machine learning will replace conventional statistical social science methods, algorithmic procedures will become increasingly common as a supplementary tool in the tool box of quantitative social scientists.

References

Bailey, M. A. and F. Maltzman (2011). *The Constrained Court: Law, Politics, and the Decisions Justices Make*. Princeton University Press.

Baum, L. (2009). *The puzzle of judicial behavior*. University of Michigan Press.

Beauchamp, N. (2017). Predicting and Interpolating State-Level Polls Using Twitter Textual Data. *American Journal of Political Science 61*(2), 490–503.

Bishop, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Computation 7*(1), 108–116.

Böckenförde, E.-W. (1976). Die Methoden der Verfassungsinterpretation: Bestandsaufnahme und Kritik. *Neue Juristische Wochenschrift 29*(46), 2089–2144.

Bonica, A. (2018). Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning. *American Journal of Political Science 62*(4), 830–848.

Breiman, L. (1996, aug). Bagging Predictors. *Machine Learning 24*(421), 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning 45*(1), 5–32.

Breiman, L., J. H. F. Friedman, R. A. Olshen, and C. L. S. Stone (1984). *Classification and Regression Trees*. Belmont: CA: Wadsworth International Group.

Brennan, T., L. Epstein, and N. Staudt (2009). Economic Trends and Judicial Outcomes: A Macrotheory of the Court. *Duke Law Journal 58*(7), 1191–1230.

Cawley, G. C. and N. L. C. Talbot (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research 11*, 2079–2107.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement 20*(1), 37–46.

Cranmer, S. J. and B. A. Desmarais (2017). What Can We Learn from Predictive Modeling? *Political Analysis 25*(2), 145–166.

Dyevre, A. (2008). Making sense of judicial lawmaking: A theory of theories of adjudication. *EUI Working Papers* (9), 1–57.

Efron, B. and T. Hastie (2016). *Computer age statistical inference*, Volume 5. Cambridge University Press.

Engst, B. G. (2018). *The Two Faces of Judicial Power. The Dynamics of Judicial-Political Bargaining*. Ph. D. thesis.

Epstein, L., J. Knight, and A. D. Martin (2001). The Supreme Court as a strategic national policymaker. *Emory Law Journal 50*, 583–611.

Forschungsgruppe Wahlen, M. (2019). Partial Cumulation of Politbarometers 1977-2017. GESIS Data Archive, Cologne. ZA2391 Data file Version 10.0.0.

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.

Guimera, R. and M. Sales-Pardo (2011). Justice blocks and predictability of us supreme court votes. *PLoS ONE 6*(11), e27188.

Hall, M. E. K. (2014, apr). The Semiconstrained Court: Public Opinion, the Separation of Powers, and the U.S. Supreme Court's Fear of Nonimplementation. *American Journal of Political Science 58*(2), 352–366.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. New York: Springer series in statistics.

Hastie, T., R. Tibshirani, and J. Friedman (2011). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* (2 ed.). New York: Springer.

Holmes, O. W. (1897). The Path of Law. *10 Harvard Law Review 457*.

Hönnige, C. (2007). *Verfassungsgericht, Regierung und Opposition: Die vergleichende Analyse eines Spannungsdreiecks*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Hönnige, C. (2009, sep). The Electoral Connection: How the Pivotal Judge Affects Oppositional Success at European Constitutional Courts. *West European Politics 32*(5), 963–984.

Hönnige, C., T. Gschwend, C. Wittig, and B. Engst (2015). Constitutional Court Database (CCDB), V17.01 [Mar.].

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics 15*(3), 651–674.

Jones, Z. M. and Y. Lupu (2018). Is There More Violence in the Middle? *American Journal of Political Science 62*(3), 652–667.

Kastellec, J. P. (2010). The Statistical Analysis of Judicial Decisions and Legal Rules with Classification Trees. *Journal of Empirical Legal Studies 7*(2), 202–230.

Katz, D. M. (2013). Quantitative Legal Prediction – or – How I Learned to Stop Worrying and Start Preparing for the Data Driven Future of the Legal Services Industry. *Emory L. J. 62*(July 2011), 909–966.

Katz, D. M., M. J. Bommarito, and J. Blackman (2017a). Crowdsourcing Accurately and Robustly Predicts Supreme Court Decisions. *Available at SSRN 3085710*, 1–11.

Katz, M. D., M. J. Bommarito, and J. Blackman (2017b, apr). A general approach for predicting the behavior of the Supreme Court of the United States. *PLoS ONE 12*(4), e0174698.

Kaufman, A. R., P. Kraft, and M. Sen (2019). Improving Supreme Court Forecasting Using Boosted Decision Trees. *Political Analysis Online fir*, 1–7.

König, T., M. Marbach, and M. Osnabrügge (2013, may). Estimating Party Positions across Countries and Time–A Dynamic Latent Variable Model for Manifesto Data. *Political Analysis 21*(4), 468–491.

Kranenpohl, U. (2010). *Hinter dem Schleier des Beratungsgeheimnisses. Der Willensbildungs- und Entscheidungsprozess des Bundesverfassungsgerichts* (1 ed.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Krehbiel, J. N. (2016, oct). The Politics of Judicial Procedures: The Role of Public Oral Hearings in the German Constitutional Court. *American Journal of Political Science 60*(4), 990–1005.

Krehbiel, J. N. (2019). Elections, Public Awareness and the Efficacy of Constitutional Review. *Journal of Law and Courts 7*(1), 53–79.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software 28*(5), 1–26.

Laver, M. and I. Budge (1992). Measuring Policy Distances and Modelling Coalition Formation. In *Party policy and government coalitions*, pp. 15–40. New York: St. Martin's Press.

Martin, A. D., K. M. Quinn, T. W. Ruger, and P. T. Kim (2004). Competing Approaches to Predicting Supreme Court Decision Making. *Symposium: Forecasting U.S. Supreme Court Decisions 2*(4), 761–767.

Medvedeva, Masha and Vols, Michel and Wieling, M. (2018). Judicial decisions of the European Court of Human Rights: looking into the crystall ball. *Proceedings of the Conference on Empirical Legal Studies in Europe 2018.*, 1–24.

Montgomery, J. M. and S. Olivella (2018). Tree-Based Models for Political Science Data. *American Journal of Political Science 62*(3), 729–744.

Neunhoeffer, M. and S. Sternberg (2019). How cross-validation can go wrong and what to do about it. *Political Analysis 27*(1), 101–106.

Ossenbühl, F. (1998). Verfassungsgerichtsbarkeit und Gesetzgebung. In P. Badura, P. Scholz, and R. Scholz (Eds.), *Verfassungsgerichtsbarkeit und Gesetzgebung. Symposion aus Anlass des 70. Geburtstages von Peter Lerche*, pp. 75–99. München: Beck.

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*(4), 1150–1210.

Russel, S. J. and P. Norvig (2016). *Artificial intelligence: a modern approach*. Malaysia: Pearson Education Limited.

Schneider, H.-P. (1974). *Die parlamentarische Opposition im Verfassungsrecht der Bundesrepublik Deutschland*. Vittorio Klostermann.

Segal, J. A. and A. D. Cover (1989). Ideological Values and the Votes of U.S. Supreme Court Justices. *American Political Science Review 83*(2), 557–565.

Segal, J. A., L. Epstein, C. M. Cameron, and H. J. Spaeth (1995). Ideological Values and the Votes of US Supreme-Court-Justices Revisited. *Journal of Politics 57*(3), 812–823.

Segal, J. A. and H. J. Spaeth (2002). *The Supreme Court and the attitudinal model revisited*. Cambridge: Cambridge University Press.

Staudt, N. and Y. He (2010). The Macroeconomic Court: Rhetoric and Implications of New Deal Decision-Making. *Northwestern Journal of Law and Social Policy 5*(5).

Sternberg, S., T. Gschwend, C. E. Wittig, and B. G. Engst (2015). Zum Einfluss der öffentlichen Meinung auf Entscheidungen des Bundesverfassungsgerichts: Eine Analyse von abstrakten

Normenkontrollen sowie Bund-Länder-Streitigkeiten 1974 - 2010. *Politische Vierteljahresschrift 56*(4), 570–598.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC bioinformatics 9*(23), 307.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics 8*(1), 25.

Strobl, C., T. Hothorn, and A. Zeileis (2009). Party on! *R Journal 1*(2), 14–17.

Sulea, O. M., M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu, and J. Van Genabith (2017). Exploring the use of text classification in the legal domain. *arXiv preprint arXiv:1710.09306*.

Vanberg, G. (2005). *The Politics of Constitutional Review in Germany*. Cambridge: Cambridge University Press.

Wittig, C. E. (2016). *The Occurrence of Separate Opinions at the Federal Constitutional Court. An Analysis with a Novel Database*. Ph. D. thesis, University of Mannheim, Berlin.

Wright, M. N. and A. Ziegler (2015). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software 77*(i01).

Appendix

A Outline of the Random Forest Algorithm

Algorithm outline of random forest, directly adopted from (Hastie et al., 2009, 558):

1. For b = 1 to B:

   (a) Draw a bootstrap sample **Z*** of size N from the training data.

   (b) Grow a random forest tree T_b to the bootstrap data, by repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

      i. Select m variables at random from the p variables.

      ii. Pick the best variable/split-point among the m.

      iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$.

   The final prediction of a new data point x is then in the classification case:

$$\hat{C}_{rf}^B(x) = majorityvote\{\hat{C}_b(x)\}_1^B$$

   where $\hat{C}_b(x)$ is the class prediction of the bth random forest tree.

B Definition of performance measures

Generally, the results of a binary classifier (and any other classifier) can be summarized by a confusion matrix. In the case of binary classification this is a 2×2 table of the four possible classification outcomes of a model. The used can all be explained with the help of confusion matrices. To get class predictions from predicted probabilities of belonging to the positive class, one has to set a threshold for positive prediction. Usually, the default value of this threshold for positive prediction is 0.5. However, any other value between 0 and 1 could be a sensible threshold for positive prediction.

Confusion Matrix

		Observed	
		Positive	*Negative*
Predicted	*Positive*	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- **Accuracy**: $\frac{TP+TN}{TP+FP+TN+FN}$
- **Precision**: Precision is defined as : Precision $= \frac{TP}{TP+FP}$, that is the ratio of correctly classified positives and all predicted positives.

- **Recall**: Recall (also called True Positive Rate (TPR)), is defined as Recall $= \frac{TP}{TP+FN}$. It measures the fraction of positive examples that are correctly labeled.
- F_1 **score**: The F_1 score is the harmonic mean of precision and recall, and defined as $F_1 = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}$.
- **Receiver operating characteristic area under the curve:** Sensitivity (recall) plotted against 1- specificity ($\frac{TN}{TN+FP}$) at various threshold settings.
- **Kappa** = $\frac{po - p_e}{1 - p_e}$, where p_o is the observed agreement (analog to accuracy), and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category.

C Comparison of predictive performance of different classifiers

Figure 3: Performance of different algorithms on the Constitutional Complaints Data, Combined Model

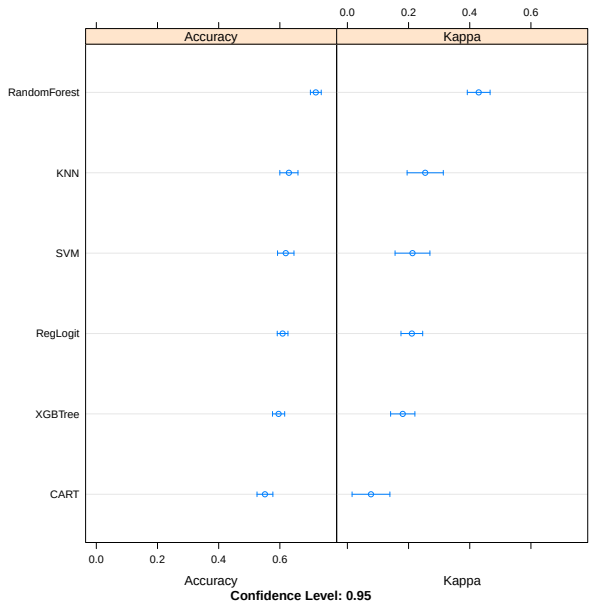


Figure 3 shows the predictive performance of multiple machine learning algorithms using 10-fold cross-validation (without hyper-parameter tuning) and the constitutional complaints data set. Classification and Regression Trees (CART), extremely boosted trees (XGBTree), regularized regression, support vector machines (SVM), k-nearest neighbors and random forests. Accuracy and Kappa are reported. Confidence intervals are just for visualization purposes and are calculated using the standard error of the respective mean (across the 10-folds).

D Additional Model Performance Metrics

Table 5: Model Evaluation Based on Aggregated Cross-Validation Scores, Additional Performance Metrics

	Accuracy			Kappa		ROC AUC		PR AUC	
	Legal	Combined	Baseline	Legal	Combined	Legal	Combined	Legal	Combined
Constitutional Complaints	60.14	**68.93**	53.47	0.20	**0.37**	0.66	**0.76**	0.70	**0.76**
Concrete Review	68.42	**80.18**	67.02	0.08	**0.50**	0.66	**0.83**	**0.85**	0.83
Abstract Review/Organstreit	63.54	**73.04**	60.26	0.19	**0.41**	0.68	**0.77**	0.73	**0.77**

Note: Model performances of the legal model and the combined model based on the aggregated 10-fold cross-validation scores. The random forests were build with a fixed m. The legal model only uses legal context variables, while the combined models used both legal and political context variables. The baseline category for accuracy is a naive classifier who always votes the majority category of the training set. The best performances are highlighted in bold.

Table 6: Model Evaluation Based on Out-of-Sample Prediction, Additional Performance Metrics

	Accuracy			Kappa		ROC AUC		PR AUC	
	Legal	Combined	Baseline	Legal	Combined	Legal	Combined	Legal	Combined
Constitutional Complaint	66.67	**74.49**	52.67	0.33	**0.49**	0.73	**0.83**	0.75	**0.83**
Concrete Reviews	75.26	**81.05**	65.79	0.41	**0.57**	0.75	**0.86**	0.65	**0.82**
Abstract Reviews/Organstreit	60.38	**77.36**	58.49	0.17	**0.52**	0.66	**0.79**	0.67	**0.82**

Note: Model performances of the legal model and the combined model based on out-of-sample prediction. The legal model only uses legal context variables, while the combined models used both legal and political context variables. The baseline category for accuracy is a naive classifier who always votes the majority category of the training set. The best performances are highlighted in bold.

E Confusion Matrices of the different classifiers

Table 1: Constit. Complaints, Legal Model

Predicted/Reference	against	in favor
against	147	79
in favor	83	177

Table 2: Concrete Reviews, Legal Model

Predicted/Reference	against	in favor
against	110	32
in favor	15	33

Table 3: Abstract Reviews/Organstreit Proceedings, Legal Model

Predicted/Reference	against	in favor
against	23	12
in favor	8	10

Table 4: Constit. Complaints, Combined Model

Predicted/Reference	against	in favor
against	169	52
in favor	61	204

Table 5: Concrete Reviews, Combined Model

Predicted/Reference	against	in favor
against	111	22
in favor	14	43

Table 6: Abstract Reviews/Organstreit Proceedings, Combined Model

Predicted/Reference	against	in favor
against	27	8
in favor	4	14

F Model Evaluation of Legal, Combined and Random Model based on Out-of-Sample Prediction

Table 7: Model Evaluation of Legal, Combined and Random Model based on out-of-sample prediction

	Accuracy			Kappa		
	Legal	Combined	Random	Legal	Combined	Random
Constitutional Complaints	64.81	**76.34**	63.58	0.29	**0.53**	0.27
Concrete Review	75.26	**81.05**	73.68	0.41	**0.57**	0.36
Abstract Reviews/Organstreit	69.93	**73.58**	72.73	0.39	**0.44**	0.42

Note: Model performances of the legal, combined and random model based on out-of-sample prediction. The best performances are highlighted in bold.

Table 7 reports the model performance of the legal, the combined and the random model using the same out-of-sample data set than used in the main analysis. Again, the combined model performs best across all metrics. We can also see that again, although less stark than in the main analysis, the addition of noise features to the model improves the predictive performance compared to the legal model for abstract reviews and Organstreit proceedings.

G Model evaluation based on out-of-sample prediction using the time dimension for splitting

Table 8: 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

Table 8 reports the performance measures for the legal and combined model using out-of-sample prediction. The test data was created by splitting the data set on each proceeding such that all observation after 2005 were assigned to the test set and all observations after where assigned to the training set. Note that this, however, results in unequal train/test splits, such that not all test sets contain the same percentage of observations. Again, the combined model achieves the best classification performance across most of the performance metrics.