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

SEBASTIAN STERNBERG[*]
University of Mannheim

July 3, 2017

ABSTRACT

Ex ante forecasting approaches become increasingly used to analyze and predict judicial outcomes, reaching impressive forecasting performances. 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 these 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 variables to the forecast of court decisions. This study addresses these two points by ex ante predicting over 2,900 proceedings decided by the German Federal Constitutional Court between 1974 and 2010, using only information available prior to the respective decision. The findings show that similar methodological approaches successfully applied to the Supreme Court also work when being applied to a Kelsenian European constitutional court. The results also demonstrate that legal context alone is already a good predictor for the outcome of case. Most importantly, the predictive performance is significantly improved when information about the political context of a decision is added. These findings are important because they support the view of a multifaceted decision-making of constitutional courts which is best characterized by the ensemble of both legal and political factors.

Keywords: Quantitative Legal Prediction; Judicial Decision-Making; Machine Learning

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

I thank Thomas Gschwend, Marcel Neunhoeffer, Tilko Swalve, Henrik Litleré Bentsen, Benjamin Engst and Caroline Wittig for their helpful comments and suggestions. I acknowledge financial support from the Graduate School of Economic and Social Sciences, University of Mannheim. Replication code and data sets can be found at https://github.com/ssternbe/GFCC_forecasting.

I. Introduction

Which factors predict constitutional court decision-making: legal context or political context? Legal academics and social scientists have long scrutinized judicial decisions to understand what motivates the judges. For over half a century, both disciplines examined past cases to explain and predict court decisions. This retrospective orientation has changed over the last decade. Building upon recent efforts in theoretical and applied machine learning, several authors attempt an ex ante forecast of US Supreme Court decision-making (Ruger et al., 2004; Kastellec, 2010; Guimera and Sales-Pardo, 2011; Katz, Bommarito II and Blackman, 2017), reaching impressive forecasting accuracies.

Yet, existing work on the forecast of court decisions has two important limitations. First, these studies exclusively focus on the US Supreme Court. The US common-law system is guided by the norm of stare decisis, under which past cases in a given area of the law have precedential effect on future cases. This consistency between fact patterns and outcomes arguably facilitates machine learning based predictions. Moreover, previous studies often use prior knowledge about the past voting behavior of individual judges to obtain case outcome predictions. Unfortunately, this rich source of information cannot be leveraged for most European constitutional courts due to the non-disclosure of individual judge's votes. Both points raise concerns about the external validity of existing studies.

Second, none of the existing studies have explicitly tested the relative contribution of legal context versus political context variables to the forecast of court decisions. There is a long-standing debate about which factors influence judicial decision-making, and thus assist to its prediction. Traditional legal scholars tend to highlight the importance of legal features and questions of legal doctrine a case raises. Other scholars emphasize the role of non-legal factors, for instance judges' attitudes or public opinion, for court decision-making. Forecasting judicial outcomes cannot end this stylized debate by providing causal evidence for one side or the other. Yet, what it can do is practically testing which factors actually contribute to the prediction of court decision-making. The key argument is that if traditional legal scholars are right, then the legal context of a decision alone should be sufficient to predict constitutional court decision-making. However, if legal realists and social scientists have a point, than including political context into the forecasting model should increase its predictive performance.

The contribution of this paper is to address the two limitations outlined above. I employ an ex ante prediction of decisions of the German Federal Constitutional Court, an archetype of the Kelsenian European constitutional court type. Leveraging a popular machine learning approach (random forests) which is still relatively unexplored in legal scholarship and political science, I forecast over 2,900 proceedings decided between 1974 and 2010 using only information available prior to a respective decision. I find that similar methodological approaches successfully applied to the Supreme Court also work when being applied to a Kelsenian European constitutional court. Testing the importance of legal and political context factors for the prediction I find that legal context alone is already a good predictor for the outcome of case. Depending on the proceeding type, it is possible to predict between two out of three and three out of four decisions correctly. Most importantly, the predictive performance is significantly improved when information about the political context of a decision is added. I conclude that the ensemble of both legal and political factors is needed to characterize court decision-making. My findings are important because they enhance our understanding of judicial behavior, and support the view of a multifaceted decision-making of constitutional courts.

II. Existing Approaches to Predict Constitutional Court Decision-Making

Legal academics and social scientists have long scrutinized judicial decisions to understand what motives the judges. Although both disciplines typically disagree on how the judges arrive at their decisions, there is one thing they have in common: the retrospective focus. Those who study judicial decisions often look at past decisions and historical facts, in other words, the outcome of past decisions (e.g. “affirm”, “reverse”) or individual judges’ votes, to assess the consistency of explanatory theories[1]. While this is neither wrong nor inappropriate, a focus on the prediction of judicial could help to provide some additional evidence in supporting or refusing some general explanatory frameworks. This is already a common approach in disciplines such as economics or medicine.

With the rise of Artificial Intelligence over the last decade, quantitative legal prediction – focusing on the ex ante prediction of future legal outcomes – has finally found its way into the study of judicial politics. Here the preferred means to an end is machine learning. Machine learning in general is defined as

[1] This is also true for most of the political science model who often analyze past data to verify some motivational hypotheses. Yet, like Ruger et al. (2004) point out, although these models often claim to “predict” judicial outcomes, what they do is more technically called “postdiction” (Ruger et al., 2004, 1154).

“a subfield of computer science concerned with computer programs that are able to learn from experience and thus improve their performance over time" (Surden, 2014, 89). The main purpose of machine learning is to detect patterns and correlations in data and to derive predictions about future outcomes. 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. Not the explanatory, but the predictive power of a variable is important here. Or, put differently, “one uses the observables to build the model rather than using the model to assign causal weight to those observables” (Katz, 2013, 952).

An outcome that draws the most interest of scholars of various fields are constitutional court decisions. Although the prediction of constitutional court outcomes (or legal “prophecy”, as Holmes (1897) calls it) per se is a long-standing idea originating from very early stages of judicial research, only a few efforts have been made to establish ex ante forecasting models for constitutional courts. 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. 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 to the release of the Supreme Court’s decision. Their machine learning model only relies on observable case characteristics such as the type of respondent, the type of petitioner, or the issue area of a case. They train their model on the “Rehnquist Court" using data from 1994 to 2002, and then test its predictive performance on the October 2002 term. Known legal experts attempted to predict the same outcomes, too. 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 humans experts correctly classified 67.9%.

As a follow up of this work, Guimera and Sales-Pardo (2011) investigate whether and to what extent it is possible 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 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. They are able to predict 83% of the individual justice's votes correctly, but do not forecast the case level outcomes directly.

In a recent study, Katz, Bommarito II and Blackman (2017) present a major effort in establishing a general, robust, and fully predictive forecasting model. They predict Supreme Court decisions over nearly two centuries (1816-2015), predicting 28,000 cases outcomes and more than 240,000 individual justice votes. Using random forests (the same method as used in the present study) and only data available prior to the date of decision, their model correctly identifies 70.2% of the court's overall affirm/reverse decisions and correctly forecasts 71.9% at the justice vote level.

A. Limitations of Existing Constitutional Court Forecasting Approaches

All of these studies provide important insights about the predictability of constitutional court decision-making. However, there are two major limitations in the existing forecasting approaches[2]. First, on the conceptional level, existing prediction models exclusively analyze and predict the US Supreme Court decision-making. This raises concerns about the external validity of previous work, and whether the similar prediction models can also be successfully applied to Kelsenian European constitutional courts. 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 by looking to past cases with a similar legal context. This high consistency between certain case fact patterns and outcomes facilities the forecasting enterprise. Even simple machine learning approaches (such as classification trees) already reach a relative high prediction accuracy (Kastellec, 2010; Ruger et al., 2004). 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 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, previous studies often use the past voting behavior of individual judges to obtain predictions. 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

[2] There are also several methodological concerns with regard to some studies, for instance the choice of notoriously unstable methods such as single classification trees to obtain predictions. However, these shortcomings already have been addressed by Katz, Bommarito II and Blackman (2017).

legal prediction models can also be successfully applied to European constitutional courts, and thus about the external validity of existing studies.

Second, none of the existing studies have explicitly tested the relative contribution of legal context versus political context variables to forecast court decisions. There is a long-standing debate about which factors influence judicial decision-making, and thus, assist to its prediction. On the one hand, legal scholars emphasize the role of legal features and questions raised by legal doctrine that appear in every individual case. Holmes, for instance, claims that "the prophesies of what the courts will do in fact [...], are what I mean by the law" (Holmes, 1897, 461). Traditional legalists claim that jurisprudence and legal doctrine work as a set of static, natural, apolitical rules that can be mechanically applied to decisions. Therefore, they tend to downplay the role of non-legal factors such as the political context of a decision and its contribution to the shape of the law. Rather, legalists emphasize the role of jurisprudence and legal doctrine. On the other hand, legal realists and social scientists demonstrate that often 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 (Epstein and Knight, 1995; Segal and Cover, 1989; Segal et al., 1995; Baum, 1997; Segal and Spaeth, 2002). Related, strategic accounts of judicial decision-making claim that judges are strategic actors originally pursuing policy-goals, but that the "decisions of the Court and its justice are subject to a number of internal and external constraints" (Pacelle, Curry and Marshall, 2011, 39). Such constrains can be for instance the need of public support to enforce their decisions (Vanberg, 2005; Hall, 2013), or a strategic restraint from their own policy preferences in a separation of powers framework (Epstein, Knight and Martin, 2001; Bailey and Maltzman, 2011).

Forecasting judicial outcomes cannot end this stylized debate between "legalism" and "attitudinalism" by providing causal evidence for one side or the other. Yet, what forecasting can do is a practical test of which of the factors actually contribute to the prediction of court decision-making. Or, as noted in Martin et al. (2004), "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). However, so far prediction models have only been used to test whether one can forecast constitutional court decisions per se, and not to test the predictive power of competing arguments[3]. The key argument here is that if traditional legal scholars are right, then the legal context of a decision should be enough to predict a substantial part of the decision-making. Subsequently, including the political context of a decision would not improve the prediction. However, if legal realists and social scientists have a point, than including political context into the forecasting model should increase its predictive performance.

To sum up, this section discussed the current state of the art of approaches to predict constitutional court decision-making relying on machine learning approaches. 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 constitutional court decision-making. In the next section I present a research design that addresses these two limitations.

III. An Ex-Ante Prediction of Constitutional Court Decisions

In this section I present a research design for an ex ante prediction of a Kelsenian European constitutional court that is able to assess the importance of legal context and political context for the prediction. This 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 in order to capture the legal context and political context of a case, and why random forests is a reasonable machine learning approach for the developing a prediction model.

A. Case Selection: The German Federal Constitutional Court

The purpose of this study is to develop a forecasting model to predict the decision-making of a constitutional court outside the US and to compare the predictive contribution of legal and political context variables within such a model. Here, the German Federal Constitutional Court (GFCC) is analyzed. The case selection is motivated by three reasons. First, the German Court is an archetype of the European Kelsenian constitutional court type and is considered being one of the most powerful and influential Courts world wide. It has served

[3] Only Katz, Bommarito II and Blackman (2017) devote some space to the relative contribution of their various variables to the prediction. Yet, they do not map their variables on a legal and political context, but just conclude that "much of the predictive power of our model is driven by tracking a variety of behavioral trends" (Katz, Bommarito II and Blackman, 2017, 12).

as a model for many new constitutional Courts in Eastern Europe. A prediction model for the German Court could hence be directly applied to these courts as well. Moreover, the Court operates in a civil law system, and the individual votes of judges are mostly confidential. This means one cannot simply predict individual judges' votes and aggregate them to make case outcome predictions. On these grounds, the German Constitutional Court represents a meaningful yet challenging study object from a predictive point of view. 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 (Kneip, 2008). This makes it a hard-case scenario to test the importance of political context for the prediction: If we are able to show that political context matters for the German Court, it presumably also matters for constitutional court where the nomination procedure is more politicized (for instance, in France).

B. Data

The data used in this study were compiled as part of the German Federal Constitutional Court Data Base (CCDB). The CCDB[4] features 38 years (1972-2010) of data on the German Constitutional Court's behavior. Here, I analyze 2,910 proceedings (referrals) grouped into 1801 main decisions. The Court often bundles multiple proceedings in one decision but decides on each of them individually. 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 Hönnige (2009) by treating the proceedings and their respective legal outcome as the unit of observation.

The GFCC knows over 20 different proceeding types, which differ in the actors or organizations entitled to file an application, the possible causes of action, and also in their political importance and societal relevance. In my analysis, I only focus on four proceeding types: *Constitutional Complaints*, *Abstract Judicial Reviews of Statutes*, *Specific Judicial Reviews of Statutes*, and *Organstreit Proceedings*. All the other proceeding types exhibit not enough variation, appear only rarely, or are not a decision in the classic sense[5]. Constitutional Complaints allow citizens to assert their freedoms that are guaranteed by the constitution

[4] This database is part of the research project "The German Federal Constitutional Court as a Veto Player" funded by the German Research Foundation and located at the University of Hannover (Germany) and the University of Mannheim (Germany).

[5] Note that the proceeding types left out only account for less than two percent of the total amount of Court decisions.

vis-à-vis the state. They are by far the most common type of proceeding, accounting for roughly two third of the observations in the data. Abstract Judicial Reviews are typically launched by political actors such as the opposition, often challenging governmental laws or statutes. Although Abstract Reviews are rare, they mainly concern matters of political nature and are of strong political and societal importance. Organstreit Proceedings may be filed if high state organs, or actors that are equivalent to such organs, disagree on their respective rights and obligations under the Basic Law. Similar to Abstract Reviews, they often raise fundamental issues that are relevant for the political system. Lastly, Specific Judicial Reviews are referred by lower courts if they are convinced that a law they have to apply is unconstitutional. The greatest difference between Abstract Reviews and Specific Reviews are the petitioners: Abstract Reviews are usually filed by political actors, while Specific Reviews can formally only be filed by lower courts. In total, my data set contains 1941 Constitutional Complaints, 760 Specific Reviews, 121 Abstract Reviews[6] and 88 Organstreit Proceedings. This represents 98% of all proceedings in the period of investigation. The proceeding types are chosen to be representative for the majority of Court decisions, but also to include proceedings with a clear political and non-political context. This is important to be able to test the importance of legal and political context.

On this basis, I develop an individual prediction model for each of the four proceeding types with the same fixed set of features. This strategy is different to other Court prediction models developing just one general model (Ruger et al., 2004; Katz, Bommarito II and Blackman, 2017) for all different kind of decision, but has several advantages. First, using different proceeding types but the same fixed set of features allows for comparing the models with respect to their accuracy 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 significantly more to predicting politicized proceeding types than to predicting proceedings without such a political context. Second, analyzing all proceedings in one model requires the assumption that the data generating process is the same for all proceeding types. This would be a strong (and maybe misleading) assumption, given that the proceeding types considerably differ in their political character or societal relevance. Finally, using one general model on the whole data set would result in a heavy bias towards variables that best explain Constitutional Complaints, as they account for two thirds of the data. The final

[6] In line with Hönnige (2009), I also coded Disputes Between the Federation and the Laender as Abstract Reviews because of their equivalence as regards content.

model would hence not be a general model for all different proceeding types, but a model only predicting Constitutional Complaints.

C. Dependent Variable

The response variable is the individual outcome of each proceeding in the data set. Every proceeding is coded as to whether the GFCC ruled against ($= 0$) or in favor ($= 1$) of the petitioner. Likewise to other studies on the German Court, I consider a partial success to be a ruling in favor of the petitioner (Hönnige, 2007; Vanberg, 2005; Hönnige, 2009; Sternberg et al., 2015)[7]. This is also in line with existing studies on the US Supreme Court, and thus allows me to compare the model performances.

In order to predict the outcome of a proceeding, I employ a number of variables which represent the legal and political context of a proceeding and proved to be relevant in existing explanatory theories of constitutional Court decision-making. Note that all of these information can be obtain *a priori* and are thus exogenous to the final outcome. All information entering the model can be obtained the same day the petitioner decides to submit the case to the Court. The model thus provides a long lead time.

D. Legal Context Variables

Representing the legal and procedural context of a proceeding, I include the *decision type*, the *issue* and *issue area* of a proceeding, and the *Senate* a proceeding is supposed to be decided in. The decision type describes whether the decision was, for instance, a main decision or an emergency appeal. The issue and issue area variables describe the topic of a decision, coded according to the Comparative Agenda Coding scheme[8]. I also include the *legal area* a proceeding is concerned with and a dummy variable indicating whether a proceeding is *grouped* together with others in a main decision or appears in an individual ruling. All those variables are taken from the CCDB. I did not include information on the petitioner type or respondent type of a proceeding, because this information is already mostly covered by the proceeding type itself[9].

[7] Although being the state of the art for analyses of the GFCC, this coding practice is somewhat limited. The GFCC occasionally declares a partial success of the petitioner, for instance by stating that a law is consistent with the constitution, but that certain parts of the statute must be revised. However, incorporating a third category made it hard to compare the findings with existing studies only using a binary coding.

[8] You can find the master codebook at http://www.comparativeagendas.net/pages/master-codebook .

[9] Specific reviews for instance can only be referred by lower Courts. Including them into the model would not lead to any gain in information.

E. Political Context Variables

I also use several predictors representing the political context of a proceeding. These variables are the *ideological position* of the Court, a measure for the *salience* of a case, the *popularity* of the government at the time of a decision, and a measure for the *perceived state of the economy*. I include the ideological position of the Court because several studies find the decision-making of the GFCC to be influenced by ideological considerations (Hönnige, 2007, 2009). The ideological direction of the Court is measured by the Manifesto Common Space Scores (MCSS) (Koenig, Marbach and Osnabruegge, 2013), following the measurement strategy proposed by Hönnige (2007). For each of the eight judges in a Senate I first assign the ideological score of the party that nominated him or her at the day of the appointment, and then calculate the mean position of the Court. The MCSS are preferred over the Comparative Party Manifesto scores because the latter are increasingly criticized with respect to their spatial and temporal comparability (Lowe et al., 2011; Koenig, Marbach and Osnabruegge, 2013).

Salience is a binary variable that shows whether a case is accompanied by an oral hearing. 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). I include this variable because there is evidence that the Court is responsive to public opinion and behaves differently in salient and in non-salient cases (Vanberg, 2005). Furthermore, the popularity variable captures the ratio of the popularity of the opposition relative to the popularity of the government. I use this variable because there is evidence that popular governments win their cases more often than governments suffering from public support (Sternberg et al., 2015). The data for this variable is taken from the German Politbarometer survey (Forschungsgruppe Wahlen, 2016).

Lastly, I include the perceived state of the economy by the public because some studies suggest that the decision-making of US-Supreme Court is shaped by the economic state of the country (Brennan, Epstein and Staudt, 2009; Staudt and He, 2010). This variable is also part of the Politbarometer survey.

I did not make a specific effort to pare down the list of input variables and there is no doubt that some of them are correlated. Yet, the machine learning approach used in this paper does not suffer from challenges conventional regression analysis would face with so many (potentially correlated) variables. Therefore, I am rather over-inclusive in adding predictors to the model. All variables are once more summarized in Table 1.

Table 1: Input Features for the Forecasting Task

Legal Context	Description	Example
Petitioner Type	Person/actor/institution/organization filing a suit	Individual citizen, MPs
Respondent Type	Person/actor/institution/organization who is accused	Government, local authority
Decision Type	The type of the decision	Main decision, preliminary ruling
Issue	Issue area (Comparative Agenda Coding Scheme)	Macroeconomic Issues
Subissue	Sub issue area (Comparative Agenda Coding Scheme)	Tax policy
Senate	Senate dealing with a proceeding	Senate I or II
Law Type	If proceeding concerns a law/statute	Federal-, state- or other law
Cause of Action	The cause of action of a proceeding	Court decision, administrative act
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	If there was an oral hearing before the trial	0 = no oral hearing, 1 = oral hearing
Popularity Opposition	Ratio 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	-1 = left, 1 = conservative

F. Method

The machine learning method of choice is random forests (Breiman, 2001*a*). Random forests is a popular ensemble classifier and has proven to be a robust and accurate forecasting tool in a variety of contexts (Díaz-

Uriarte and Alvarez de Andrés, 2006; Cutler et al., 2007; Rodriguez-Galiano et al., 2012; Hayes et al., 2015; Berk, Sorenson and Barnes, 2016). Random forests are also increasingly used in the field of social sciences (Montgomery and Olivella, 2015; Muchlinski et al., 2016; Berk, Sorenson and Barnes, 2016; Cranmer and Desmarais, 2016).

Random forests use an ensemble of classification trees that leverages two forms of randomness: *bagging* – short for *b*ootstrap *agg*regation – (Breiman, 1996) and *random substrates* of the predictor variables. Bagging is a resampling method that takes many repeated samples with replacement from the original data set. The typical random forest bootstrap sample contains around two-third of the observations of the original data set. The remaining one-third of the observations of the original data set not occurring in the bootstrap sample are called *out-of-bag (oob)* observations. On each of the samples, a single (unpruned) classification tree is grown. For each split in the tree, only a small number of randomly selected predictor variables (random substrates) is considered as candidates for the split. The random selection of splitting variables allows predictors that were otherwise outplayed by their competitors to enter the ensemble. This has the benefit of addressing potential collinearity issues, giving each of the correlated predictors the chance to appear in different bootstrap trees. The resulting trees are thus decorrelated, which makes the average of the respective trees more robust. When the trees are grown, each is used to predict the oob observations. The predicted class of an observation is calculated by majority voting of the oob-predictions for that observation, with ties split randomly[10].

There are strong reasons to prefer random forests to other machine learning classifiers. First, the same classifier has turned out to be a strong learner in a comparable study (Katz, Bommarito II and Blackman, 2017). Second, in a machine learning analysis of judicial decisions and legal rules, Kastellec (2010) finds that trees corresponds to the "hierarchical and dichotomous structure that often seems apparent in judicial opinions"[11] (Kastellec, 2010, 210). Third, a test using several common classification algorithms with their default parameters shows that random forests outperforms other learners[12]. Fourth, tree models are extremely effective for detecting nonlinearities and interactions in datasets with many (potentially irrelevant)

[10] A more mathematical outline of the random forest algorithm can be found in (Hastie, Tibshirani and Friedman, 2011, 588).

[11] Note that Kastellec (2010) uses single classification trees (CART) for his analysis. However, although they provide nice tree-based visualizations of the model, they are not very accurate predictors (Breiman, 2001*b*).

[12] I used CART, Random Forests, Support Vector Machines, bagged trees, k-nearest neighbors and logistic regression on the Constitutional Complaints data set that contains most of the observations. The simulation results are in Figure 3 in the Appendix.

covariates (Montgomery and Olivella, 2015, 1). This is an important property given the input variables I use. Moreover, random forests also provides instructive visualizations of forecasting performance.

IV. Results

In this section I present the results of the ex-ante prediction of decisions of the German Federal Constitutional Court. The section is divided into two parts. In the first part, I develop the random forests models for 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 is more accurate than a model using legal context factors alone. In the second part, I open the black-box of the prediction model by firstly comparing the predictive importance of the input variables across the proceeding types and then discussing interesting non-linearities in the data conventional regression analysis might have overlooked.

A. Predicting Decisions of the German Federal Constitutional Court

For each proceeding type in my data, two individual random forests models are developed: a *legal context model* only featuring legal context variable, and a *combined model* featuring legal context *and* political context variables[13]. The legal context model, including legal and procedural case features, represent the predictive approach of traditional legal scholars. Legal scholars emphasize the role of legal features and questions raised by legal doctrine that appear in every individual case. In contrast, it is argued by legal realists and social scientists that non-legal factors, for instance ideology or public opinion, are key to understand and predict judicial outcomes. If traditional legal scholars who neglect the value of political context for the prediction of legal outcomes are right, then adding political context variables to the model should not change (or increase) the predictive power of the model. Instead, the predictive capability should remain stable at the same level than before, or even drop if political context is totally irrelevant for the prediction and too much noise is added (Rogers and Gunn, 2006). 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 is more important than legal context.

[13] The random forests are estimated using the *caret* package (Kuhn, 2008), which is a wrap-up package for the original random forest implementation in R (Liaw and Wiener, 2002). Replication code can be found at `https://github.com/ssternbe/GFCC_forecasting`.

To compare the models, a robust performance assessment is crucial. I therefore apply stratified k-fold cross-validation with 5 folds per training. This is recommended when facing relatively small data sets with many categories (Kohavi, 1995). In a 5-fold cross-validation, the data is split into 5 random partitions (or "folds") of nearly the same size. A model is fit using all samples except the first fold. The held-out samples are predicted by this model and used to estimate performance measures. The first subset is then returned to the training set, and the procedure is repeated but with the second subset held out. In 5-fold cross-validation this is done five times. The obtained estimates of performance are then summarized and help to choose the best model (for more information see e.g. Hastie, Tibshirani and Friedman, 2011). Furthermore, performance measures are used diagnostically to help determine the values of tuning parameters[14].

The first column of Table 2 reports the prediction accuracy across all proceeding types for the legal model only using the legal context of a case. The accuracy of the combined model is reported in the second column. The predictive performance is assessed by the overall prediction accuracy, which is given by the sum of true positives and true negatives relative to the total number of observations. This reflects the agreement between the observed and predicted classes and allows for a straight forward interpretation. To provide an intuition about the uncertainty of the prediction, the standard deviations of the values are reported behind. The last column of the table shows the majority class of the particular proceeding type. The majority class reflects the accuracy a lazy learner could obtain by always classifying an observation according to the most frequently occurring category.

The legal model alone is already enough to predict a substantial amount of cases correctly. It is best in predicting Specific Reviews (75% correctly forecasted outcomes), followed by Organstreit Proceedings (73%). However, we have to take into account that for Organstreit Proceedings the majority class is already 74%. The legal model for Organstreits is thus worse than a lazy learner. The models for Constitutional Complaints and Abstract Reviews both reach an accuracy of 67%. On average, the legal model thus correctly predicts two out three/three out of four decisions, respectively. While the performance of the legal model is already good, predictions become even more accurate when the political context of a decision is

[14] For each random forest, 2000 trees ($ntree = 2000$) are grown because simulation studies suggest that smaller values can result in unstable estimates under certain circumstances (Strobl et al., 2007; Strobl, Hothorn and Zeileis, 2009). Furthermore, the default value of the number of random substrates ($mtry$) is $\sqrt{p}$, where p is the number of input variables. However, this is not always the optimal value (Díaz-Uriarte and Alvarez de Andrés, 2006, 4). Therefore, to find the optimal $mtry$ value for each model I simulated several random forest models with different $mtry$ values. The best model was then chosen on basis of the $Kappa$-metric.

Table 2: Forecasting Accuracy for the Legal Context Model and Combined Model

Proceeding Type/Models	**Legal Model**	**Combined Model**	**Majority Class**
Constitutional Complaints	0.67 ± 0.03	0.72 ± 0.02	0.51
Specific Reviews	0.75 ± 0.01	0.83 ± 0.02	0.77
Organstreit Proceedings	0.73 ± 0.07	0.84 ± 0.01	0.74
Abstract Reviews	0.67 ± 0.03	0.73 ± 0.02	0.50

Note: Forecasting accuracy for the legal model and the combined model. The legal model only uses legal context variables, while the combined models used both legal and political context variables. The accuracy is measured as the percentage of correctly classified instances. The values are obtained by stratified 5-fold cross-validation. Standard deviations of the accuracies are behind the accuracy value. As a reference point, the majority class of each proceeding type is reported in the last column.

taken into account. This holds for every proceeding type. Especially for the two rather political proceeding types, namely Organstreit Proceedings and Abstract Reviews, the addition of political context to the prediction model improves the accuracy considerably (+11 percentage points (pp) for Organstreit Proceedings, +5 pp for Abstract Reviews). However, considering political context also improves the accuracy of Constitutional Complaints (+5 pp) and Specific Reviews (+8 pp). These results stay robust when other predictive performance measures are used[15].

The results lead to several conclusions. First, the findings for the US Supreme Court – that machine learning models can successfully predict decision outcomes – can be generalized at least to the German Constitutional Court, an archetype of the Kelsenian European Constitutional Courts. Similar machine learning approaches reach similar accuracies[16], although there is no norm of stare decisis or individual judge's votes to leverage which both facilitates machine learning based predictions. Second, it is evident that political context improves the prediction of all proceeding types. While this might be intuitive for proceeding types that often deal with political matters, it is rather surprising that political context also improves the prediction of non-political proceeding types. This is a strong and surprising finding, because a part of the German Court literature still refuses to consider that the Court's decision-making might be influenced by factors other than pure legal reasoning (Böckenförde, 1976; Hesse, 1999; Ossenbühl, 1998). However, I want to

[15] Because the traditional prediction accuracy has some known limitations, I provide additional performance measures such as the $Kappa$-metric, the oob-error rate and the ROC/AUC score of each model in Appendix B.

[16] The simple, unweighted accuracy across all proceeding types for the combined model is, on average, 78% correctly classified cases. The weighted accuracy (accuracy of a proceeding type × the amount of observations of this proceeding type) is 0.75% correctly predicted cases. The accuracy is 78% correctly predicted cases for Ruger et al. (2004), and 70% for Katz, Bommarito II and Blackman (2017).

emphasize again that this does not mean that political factors are more important than legal ones. Instead, just the ensemble of legal and political variables collectively contribute to the prediction in the combined model. Nevertheless, with respect to the central research question of this paper – does the political context of a decision actually contributes to the prediction of its outcome – I conclude that knowing the political context of a decision indeed improves the prediction. To further investigate the role of legal and political context I look at each variable's importance for the forecast in the next section.

B. The Relative Contribution of Legal Context and Political Context for the Prediction

What role do political variables actually play for the prediction? And more important, is there any variation in their contribution across proceeding types? Previously it has been shown that the combined model consisting of both legal and political context variables is superior to the legal model only including legal factors alone. However, to emphasize the importance of political context it would be helpful to compare the actual contribution of both legal and political context variables for the outcome prediction.

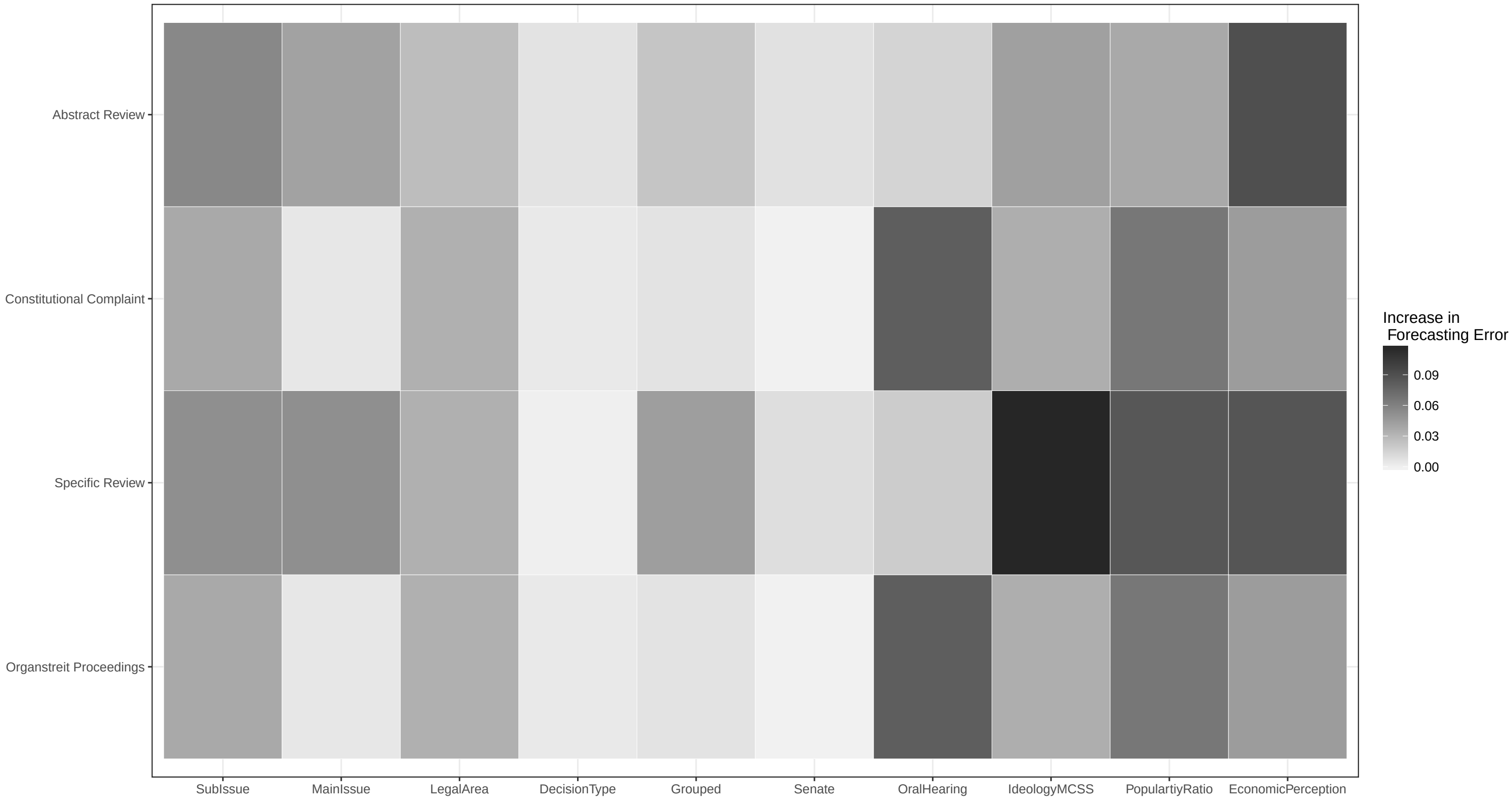
The importance of a variable for the random forests forecast can be obtained via its *variable importance*. Variable importance (also known as permutation importance) is a measure for the mean decrease in forecasting error if an input variable is randomly permuted, and thus, the relationship between this input variable and the outcome is broken. An outline of the algorithm can be found in Appendix C[17].

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 the heatmap, the higher the variable importance of the given predictor for the respective proceeding type. Therefore, for example, the salience of a given case has a forecasting importance of a little less than 0.10 for Organstreit Proceedings. This means that if the information about the salience of a case is blocked from the forecasting process, the prediction accuracy for Organstreit Proceedings drops by 10% (from 84% to 0.74%)[18].

[17] In short, the basic idea is to randomly permute the values of any given predictor when forecasts are being constructed. The intuition behind is that if a feature is not useful for predicting an outcome, then altering or permuting its values will not result in a significant reduction of model performance.

[18] Note that this does not measure the effect on prediction if this variable was not available, because if the model was refitted without this variable, other variables could be used as surrogates (Hastie, Tibshirani and Friedman, 2011, 593).

Figure 1: Heatmap of variable importance per proceeding type



Note: The different predictors are displayed on the horizontal axis, while the different type of proceedings are shown on the vertical axis. Darker fields indicate a higher importance of the respective variable for the respective proceeding type. The combined model was used to estimate the variable importance.

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 salience of a decision is important for the prediction of Constitutional Complaints and Organstreit Proceedings, but not so much for Abstract and Specific Reviews. Another inference is that the ideological position of the Court, the popularity of the opposition and the economic perception variable are the most important predictors for Specific Reviews. This is surprising, given that these proceedings are said to be rather unpolitical. From the legal input variables, only the two issue variables stand out. Perhaps the main inference from Figure 1 is that political context is relevant for case outcomes. The previous conclusion that political context improves the prediction of Court decisions is thus confirmed.

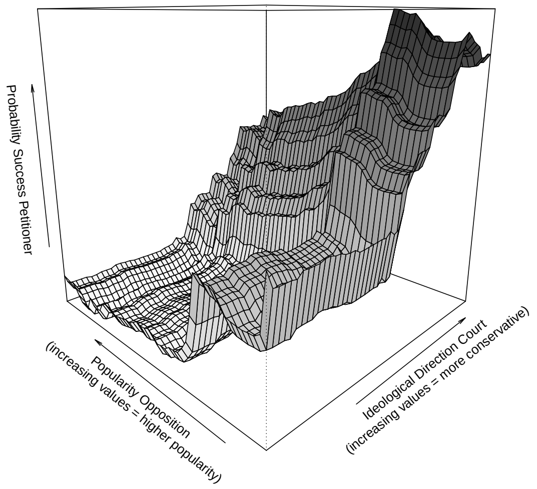
These conclusions should be considered with caution. As stated earlier, some of the predictors are strongly associated, which complicates the interpretation of the variable importance (Strobl et al., 2007, 2008; Strobl, Hothorn and Zeileis, 2009). Variable importance is also not a hypothesis test for the causal relationship between an input and the outcome. That said, however, the importance of some predictors, for instance the importance of salience and the popularity of the opposition for Organstreit Proceedings, is in line with the findings of other studies on the decision-making of the Court (Vanberg, 2005; Sternberg et al., 2015).

Variable importance only indicates that an input variable contributes to the prediction of an outcome. It does not allow to draw inferences about the effect direction between an input and the outcome. For instance, from the variable importance we know that the salience of a case is an important predictor for Organstreit Proceedings, but we do not know whether the salience of a case decreases or increases the winning chances of the petitioner. This information is contained in partial dependence plots.

C. Relationships between Inputs of the Outcomes to be Forecasted

Partial dependence plots are a method to visualize the partial relationship between predictors and the outcome. In short, such plots give a graphical representation of the marginal effect of a variable on the predicted outcome, after accounting for the average (mean) effects of the other input variables (Hastie, Tibshirani and Friedman, 2011, 369). The algorithm is again outlined in the Appendix C.

Figure 2: Partial dependence plot for ideological direction of the Court × popularity opposition/government on the petitioner's success probability



Note: Partial dependence plot for the relationship between popularity opposition/government and the ideological direction of the Court on the probability of a petitioners success. The Y-axis shows the probability of a petitioner's success, while popularity of the opposition (left) and ideological direction (right) are on the X-axis. The combined model from Table 2 was used for estimation. The graph shows a clear non-linear relationship between the outcome and the two predictors.

Figure 2 shows the partial dependence plot for the interaction between the ideological direction of the Court and the popularity of the opposition on the probability of a petitioner's success. The combined model for Specific Reviews was used. This scenario was chosen because the variable importance plots in Figure 1 shows that these variables are important predictors for the rather apolitical proceeding type of Specific Reviews.

We can draw several conclusions from the partial dependence plot. First, there is a negative association between oppositional popularity and petitioner success, indicated by the flat surface in lower left part of Figure 2. However, this effect is dependent on the ideological direction of the Court: the more conservative the ideological position of the Court, the higher the likelihood of a petitioner's success (indicated by the sharp rise in the upper right of Figure 2). In other words, the winning chances of a petitioner in this scenario are the lowest if the opposition is very unpopular and the Court is rather left, whereas the winning chances are the highest if public support for the opposition is low and the Court is rather conservative.

Second, and more important, the effect between both 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 advance). This is an important observation, because these results suggest that the rather apolitical proceeding types such as Specific Reviews and Constitutional Complaints might not be as apolitical as one thinks, but have been investigated using the wrong methods.

Partial dependence plots are not supported for each predictor for space reasons. For most of the inputs the direction of all effects are largely as one would expect. The salience of a case, for instance, strengthens the effect of other political context predictors such as the ideological position of the Court. This is in line with research that shows 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. Again, most of the input relations are non-linear.

Although these insights should be treated critically concerning the inference of causal processes from them, they at least suggest that a lot of the effects are non-linear. Basically none of the relationships between inputs and outcome displays the characteristic S-curve of a logistic regression model. Machine learning techniques could thus be a fruitful approach to identify the effects of political context in proceeding types that are said to be apolitical.

V. Conclusions and Implications

This study has highlighted the ability of machine learning analysis to forecast decisions of the German Federal Constitutional Court, an archetype of the Kelsenian European constitutional court type. Specifically, I successfully conduct an ex ante forecast of the outcomes of over 2,900 proceedings decided between 1974 and 2010, using only data available prior to decision. Such a prediction model is a novelty for the German Court, and European courts in general. My findings thus confirm the external validity of similar work on the US Supreme Court. Moreover, I explicitly test the predictive contribution of legal context and political context variables to the forecast. I find that legal context alone is, on average, a relatively good predictor of for the outcome of case. Most importantly, my results show that the predictive performance is improved when the political context of a decision is leveraged. I also find that political context matters for different kind of proceeding types, not only the ones known to be politicized. Constitutional court decision-making is thus best characterized by the ensemble of legal and political context factors.

With these findings in mind, it is possible to discuss a few implications of the results. Is it a threat for the nature of law if the Court's decision-making is largely correctly forecasted by an algorithm? The answer is no. A certain degree of regularity or consistency in judicial decision-making is important to the basic functioning of the rule of law. People need to be able to anticipate the legal consequences of their actions. Therefore, for the sake of legal certainty, it is desirable that cases with the same context lead, on average, to the same legal outcomes. Moreover, no algorithm could in any way substitute for the important work judges do in their reasonings.

Furthermore, my analysis suggests that the relationship between most of the input variables, either legal or political context ones, is non-linear. This non-linearity would not be captured by conventional statistical social science methods. Therefore, when scholars analyze rather non-political decisions and conclude that there is no effect of political context variables this null-finding might be driven by a the choice of methods.

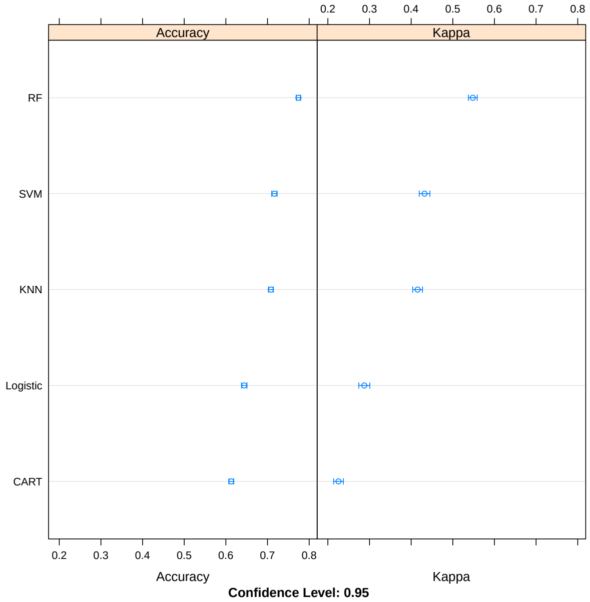
There may be another implication of this study beyond academia: the world of litigants before the German Court. No matter lawyers, politicians or citizens, the expected outcome of a case, that is, the perceived probability to win or lose, plays a crucial role in a litigant's decision to press an appeal. In this regard the results may be beneficial for practicing attorneys and their clients likewise.

My findings also open up new avenues for study. Because many other European constitutional courts have the same institutional design like the German Court, a similar methodological approach could be directly applied to them. Moreover, my findings can be used as a guide to further investigation. Why, for instance, is the salience of a proceeding so important for some proceeding types but not for others? What is the underlying causal mechanism that links the perception of the economic shape of the country over the issue area of a case to its outcome? While being beyond the scope of this study, these are some questions that future research, either quantitative or qualitative, could address.

Appendix

Appendix A: Algorithm Selection

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



Note: The table reports the predictive performance of different algorithms using the combined model. The different algorithms are on the Y-axis, while the model performance (measured in the percentage correctly classified instances and the Kappa-metric) are on the X-axis. The points represent the estimated performance measure, while the bars indicate 95% confidence intervals. All models were trained and tested on the constitutional complaints data set.

Appendix B: Additional Performance Measures

Table 3: Additional Performance Measures of Legal Context and Combined Model

	Legal Model			**Combined model**		
	AUC	Kappa	oob error	AUC	Kappa	oob error
Organstreit Proceedings	0.63 (0.03)	0.20 (0.24)	32.95%	0.82 (0.05)	0.55 (0.03)	12.05%
Constitutional Complaints	0.73 (0.04)	0.34 (0.06)	30.90%	0.79 (0.03)	0.42 (0.04)	24.47%
Abstract Reviews	0.75 (0.09)	0.34 (0.15)	38.02%	0.84 (0.02)	0.46 (0.04)	32.20%
Specific Reviews	0.71 (0.02)	0.38 (0.02)	25.79%	0.86 (0.02)	0.58 (0.03)	14.21%

Note: Forecasting accuracy for the legal model and the combined model. For each model, three performance measures are reported: the Area under the Curve (AUC) of the Receiver Operation Characteristic (ROC), the Kappa-metric, and the oob-error rate. As you can see, all performance measures report better values for the combined model.

Appendix C: Algorithms

Outline of the Random Forests Algorithm (adopted from Berk, Sorenson and Barnes (2016))

1. From a training data set with N observations, take a random sample of size N with replacement. The observations not chosen at random become the test ("out of bag", or "oob") data.
2. Take a random sample without replacement of p predictors (the default is $\sqrt{p}$)
3. Construct the first classification partition of the data based on the CART methods
4. Repeat step 2 for all subsequent partitions until any further partitioning does not improve the model fit. Do not prune.
5. Drop the out-of-bag data (i.e., data not used to grow the tree) down the tree. Store the class assigned to each observation along with each observation's predictor values.
6. Repeat steps 1-5 a large number of times
7. Using only the class assigned to each observation when that observation is not used to grow the tree, count the number of times over trees that the observation is classified in each outcome category.
8. Assign each case to an outcome category by a plurality vote over the set of trees.

Outline of the Variable Forecasting Importance Algorithm

1. Estimate the prediction error on the OOB portion of the data. Note that the OOB portion was not used to construct the trees.
2. For p predictors, repeat Step 1 p times, but each time after permuting the values of each predictor variable randomly. Permuting breaks the relation between the response and the predictors. For each of the permuted predictors, the prediction error is estimated again.
3. For each of the p predictors, take the average difference of the two accuracies (the prediction error with and without a given predictor permuted) over all trees. This is the final measure of variable importance.

Note that because my predictor variables are of different types (different scales of measurement and different numbers of categories), one can also use the so-called conditional variable importance measure as suggested by Strobl et al. (2008). This measure is more robust towards potential correlations between predictors. Using this particular variable importance measure did not made difference as compared to the original one.

Outline of Partial Dependence Plot Algorithm

From the R package: “For each tree, the prediction accuracy on the out-of-bag portion of the data is recorded. Then the same is done after permuting each predictor variable. The difference between the two accuracies are then averaged over all trees, and normalized by the standard error. For regression, the MSE is computed on the out-of-bag data for each tree, and then the same computed after permuting a variable. The differences are averaged and normalized by the standard error. If the standard error is equal to 0 for a variable, the division is not done (Liaw and Wiener, 2002).

References

Bailey, Michael A. and Forrest Maltzman. 2011. *The Constrained Court: Law, Politics, and the Decisions Justices Make*. Princeton University Press.

Baum, Lawrence. 1997. *The puzzle of judicial behavior*. University of Michigan Press.

Berk, Richard A., Susan B. Sorenson and Geoffrey Barnes. 2016. "Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions." *Journal of Empirical Legal Studies* 13(1):94–115.

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

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24(421):123–140.

Breiman, Leo. 2001*a*. "Random Forests." *Machine Learning* 45(1):5–32.

Breiman, Leo. 2001*b*. "Statistical Modeling: The Two Cultures." *Statistical Science* 16(3):199–231.

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

Cranmer, Skyler J. and Bruce A. Desmarais. 2016. "What can we Learn from Predictive Modeling?" *Political Analysis* 25:145–166.

Cutler, D. Richard, Thomas C Edwards, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson and Joshua J Lawler. 2007. "Random Forests for Classification in Ecology." *Ecology* 88(11):2783–2792.

Díaz-Uriarte, Ramón and Sara Alvarez de Andrés. 2006. "Gene selection and classification of microarray data using random forest." *BMC bioinformatics* 7(1):3.

Epstein, Lee and Jack Knight. 1995. "Documenting Strategic Interaction on the US Supreme Court.".

Epstein, Lee, Jack Knight and Andrew D. Martin. 2001. "The Supreme Court as a Strategic National Policymaker." *Emory Law Journal* 50:583–611.

Forschungsgruppe Wahlen, Mannheim. 2016. "Partial Cumulation of Politbarometers 1977-201.".

Guimera, Roger and Marta Sales-Pardo. 2011. "Justice Blocks and Predictability of U.S. Supreme Court Votes." *Current Science* 6(11):1–8.

Hall, Matthew E K. 2013. "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, Trevor, Robert Tibshirani and Jerome Friedman. 2011. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* 2 ed. New York: Springer.

Hayes, Timothy, Satoshi Usami, Ross Jacobucci and John J McArdle. 2015. "Using Classification and Regression Trees (CART) and random forests to analyze attrition: Results from two simulations." *Psychology and Aging* 30(4):911–929.

Hesse, Konrad. 1999. *Grundzüge des Verfassungsrechts der Bundesrepublik Deutschland*. Heidelberg: Mohr Siebeck.

Holmes, Oliver Wendell. 1897. "The Path of Law." *10 Harvard Law Review* 457.

Hönnige, Christoph. 2007. *Verfassungsgericht, Regierung und Opposition*. 1 ed. Wiesbaden: VS Verlag für Sozialwissenschaften.

Hönnige, Christoph. 2009. "The Electoral Connection : How the Pivotal Judge Affects Oppositional Success at European Constitutional Courts." (March 2012):37–41.

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

Katz, Daniel Martin. 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 Law Journal* 62:909–966.

Katz, Daniel Martin, Michael J Bommarito II and Josh Blackman. 2017. "A general approach for predicting the behavior of the Supreme Court of the United States." *Plos-One* 12(4).

Kneip, Sascha. 2008. "Verfassungsgerichtsbarkeit im Vergleich." *Die EU-Staaten im Vergleich* pp. 631–655.

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

Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." *International Joint Conference on Artificial Intelligence* 14(12):1137–1143.

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

Liaw, a and M Wiener. 2002. "Classification and Regression by randomForest." *R news* 2(December):18–22.

Lowe, Will, Kenneth Benoit, Slava Mikhaylov and Michael Laver. 2011. "Scaling Policy Preferences from Coded Political Texts." *Legislative Studies Quarterly* 36(1):123–155.

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

Montgomery, Jacob M and Santiago Olivella. 2015. "Tree-based models for political science data." *Forthcoming at American Journal of Political Scienc* .

Muchlinski, David, David Siroky, Jingrui He, Matthew Kocher and R Michael Alvarez. 2016. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data." *Political Analysis* 24:87–103.

Ossenbühl, Fritz. 1998. Verfassungsgerichtsbarkeit und Gesetzgebung. In *Verfassungsgerichtsbarkeit und Gesetzgebung. Symposion aus Anlaß des 70. Geburtstags von Peter Lerche*, ed. Peter Badura and Rupert Scholz. München: C.H.Beck pp. 22–40.

Pacelle, Richard L, Brett W Curry and Bryan W Marshall. 2011. *Decision making by the modern Supreme Court*. Cambridge: Cambridge University Press.

Rodriguez-Galiano, V. F., B. Ghimire, J. Rogan, M. Chica-Olmo and J. P. Rigol-Sanchez. 2012. "An assessment of the effectiveness of a random forest classifier for land-cover classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 67(1):93–104.

Rogers, Jeremy and Steve Gunn. 2006. Identifying Feature Relevance Using a Random Forest. In *Subspace, Latent Structure and Feature Selection. Lecture Notes in Computer Science*, ed. Craig Saunders, Marko Grobelnik, Steve Gunn and John Shawe-Taylor. Berlin, Heidelberg: Springer pp. 173–184.

Ruger, Theodore W, Pauline T Kim, Andrew D Martin, Kevin M Quinn, Source Columbia, Law Review and No May. 2004. "The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking." *Columbia Law Review* 104(4):1150–1210.

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

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

Segal, Jeffrey A, Lee Epstein, Charles M Cameron and Harold J Spaeth. 1995. "Ideological Values and the Votes of U.S. Supreme Court Justices Revisited." *Source: The Journal of Politics* 57(3):812–823.

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

Sternberg, Sebastian, Thomas Gschwend, Caroline Wittig and Benjamin G. Engst. 2015. "Zum Einfluss der öffentlichen Meinung auf Entscheidungen des Bundesverfassungsgerichts. Eine Analyse von abstrakten Normenkontrollen sowie Bund-Länder-Streitigkeiten 1974 - 2010.".

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis and Torsten Hothorn. 2007. "Bias in random forest variable importance measures: illustrations, sources and a solution." *BMC Bioinformatics* 8(1):25.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin and Achim Zeileis. 2008. "Conditional variable importance for random forests." *BMC bioinformatics* 9(23):307.

Strobl, Carolin, Torsten Hothorn and Achim Zeileis. 2009. "Party on!" *R Journal* 1(2):14–17.

Surden, Harry. 2014. "Machine Learning and Law." *Washington Law Review* 89.

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