# What predicts the precedential value of judgments by the Court of Justice of the European Union? A Machine Learning Framework.*

David Hilpert

This version: January 26, 2024†

## 1 Motivation

The Court of Justice of the European Union (CJEU) has often been referred to as an engine of European integration (Pollack, 2003). As such it is said to drive the legal integration of Europe in times of blockade among member state governments – and sometimes even against their expressed interest (Carrubba and Gabel, 2014; Larsson and Naurin, 2016; Schmidt, 2018). Along these lines, scholars have studied how the CJEU builds up its case-law in order to construct a consistent mode of reasoning across cases. Like other international courts, the CJEU is seen to build precedent over time in service of the legal integration of Europe (Lupu and Voeten, 2012).

The focus of this project is to examine factors that predict the precedential value of CJEU cases. Which factors can explain which cases are referred to in, and serve as the base of, future judgments at the CJEU?



† For the latest version, check out davidhilpert.github.io.

# 2 Methodology

## 2.1 Data

To investigate this question, I gathered data on CJEU judgments under the preliminary reference procedure (Article 267 TFEU). I collect 6338 judgments in the timeframe between 1995 and 2023 from EUR-Lex (2023), the EU's legislative database. The aim is to stochastically predict precedential value. I capture this in an outcome variable that is an absolute count of how often a case gets cited by future judgments within 3 years after its publication. This means that only judgments until 2020 can be considered, since these are the last whose cite-count can be evaluated based on the data I have. (This is called right-truncation). The final dataset comprises 5310 judgments.

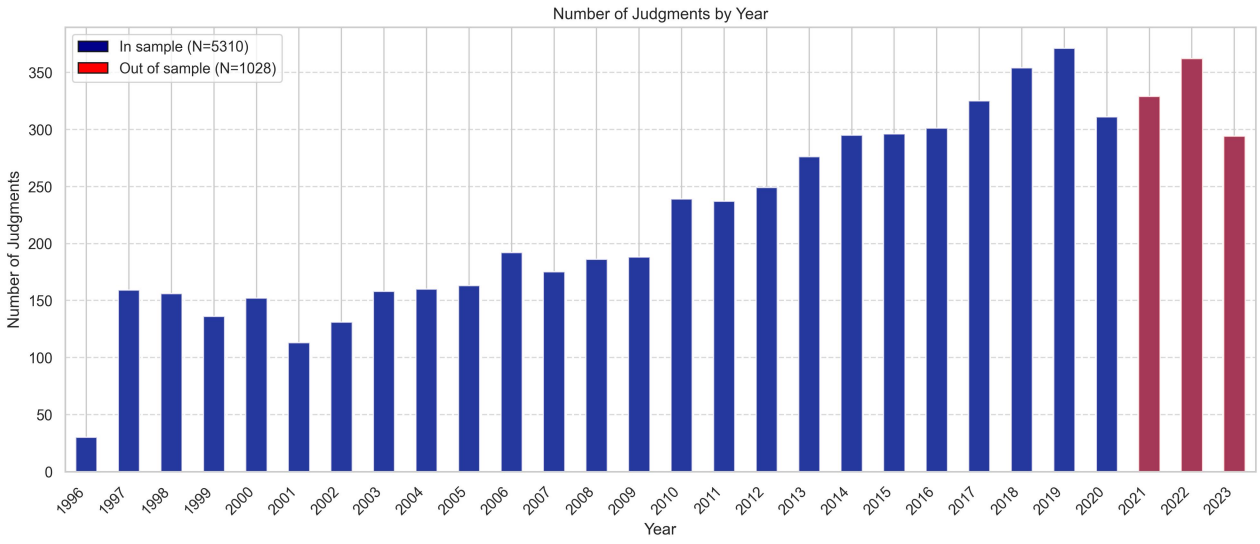


Figure 1: Overview of sample

## 2.2 Features

In explaining why some cases are being cited more often than others, one can resort to demand-side and supply-side factors. On the demand side, one can think of reasons why future judges would like to ground their reasoning in prior cases. This may include legal

as well as political considerations. However, from the present perspective, it is difficult to anticipate the future composition of the court and which considerations will drive judgments. For the purpose of this project, I will seek to predict the number of times judgments get cited exclusively focusing on supply-side factors.

Here I propose to group supply-side factors into four clusters. I then derive features that can reasonably be grouped within these clusters.

1. features that relate to legal substance at stake
2. features that relate to how the judgment is crafted
3. features of the court
4. features of the (political) context

Table 1 lists the features used to operationalize each of the four clusters along with descriptive information.

The first cluster of features relates to the legal substance at stake in a given judgment. First, dummies that indicate whether the judgments affects an EU regulation (`affectsR`), directive (`affectsL`), decision (`affectsD`), earlier court judgment (`affectsCJ`) or treaty article (`affectsT`), respectively. Second, a count variable measures how often a given legal act has been touched upon before the judgment in question (`prior touches min vec`). The reasoning is that when the court gets to interpret a legal act such as a treaty article or a directive for the first time, the judgment could be more controversial than if it is the tenth court case to touch upon an article. In case a judgment affects multiple legal acts, I take the minimum count in order to capture the least-interpreted, presumably most controversial legal act affected. Finally, I include a measure of novelty for legal acts developed in my dissertation. Based on structural topic models Roberts, Stewart and Nielsen (2020), I seek to quantify the extent to which legal acts break new substantive ground, which could lead to more controversial disputes before court, and higher precedential value.

Table 1: Descriptive overview of features used in ML model

| | count | mean | std | min | max |
|---|---|---|---|---|---|
| **Legal substance** | | | | | |
| affectsR | 5310.0 | 0.3 | 0.5 | 0.0 | 1.0 |
| affectsL | 5310.0 | 0.5 | 0.5 | 0.0 | 1.0 |
| affectsD | 5310.0 | 0.0 | 0.1 | 0.0 | 1.0 |
| affectsCJ | 5310.0 | 0.0 | 0.1 | 0.0 | 1.0 |
| affectsT | 5310.0 | 0.2 | 0.4 | 0.0 | 1.0 |
| prior_touches_min_vec | 5310.0 | 33.8 | 71.4 | 0.0 | 408.0 |
| act_scores | 5310.0 | -0.1 | 0.7 | -1.0 | 1.6 |
| **Crafting of the judgment** | | | | | |
| year_lodgment | 5310.0 | 2009.0 | 7.0 | 1995.0 | 2019.0 |
| textlength | 5310.0 | 6888.6 | 3537.0 | 1.0 | 53818.0 |
| days_it_took | 5310.0 | 581.7 | 184.1 | 40.0 | 2056.0 |
| num_citations_vec | 5310.0 | 4.0 | 3.3 | 0.0 | 47.0 |
| cosine_sims_vec | 5310.0 | 0.8 | 0.1 | 0.4 | 2.0 |
| subj_cosine_sims_vec | 5310.0 | 1.0 | 0.1 | 0.1 | 2.0 |
| **Court features** | | | | | |
| grand_chamber | 5310.0 | 0.1 | 0.3 | 0.0 | 1.0 |
| chamber_size | 5310.0 | 3.8 | 2.6 | 0.0 | 22.0 |
| **(Political) context** | | | | | |
| country_of_origin_weights | 5310.0 | 0.1 | 0.1 | 0.0 | 0.2 |
| num_obs_vec | 5310.0 | 2.0 | 1.8 | 0.0 | 15.0 |
| amicus_curiae_weights | 5310.0 | 0.1 | 0.1 | 0.0 | 0.7 |
| lagged_hit_ratio | 5310.0 | 0.0 | 0.0 | 0.0 | 0.0 |

The second cluster of features relates to the way the judgment is crafted. First, the year when the case was lodged at the CJEU is recorded (`year lodgment`). Second, the number of days that passed between the day the case was lodged and the day the final judgment is made (`days it took`). Third, the length of the judgment text and fourth, the number of distinct pieces cited in the judgment (`num citations vec`), both being a rough approximation to the care that went into crafting the judgment. Fifth, there are two measures of substantive similarity between the judgment in question and earlier cases. This measure is created drawing on the judgment texts, which are preprocessed using standard steps, including stemming, the removal of digits and stopwords, as well as several terms frequent to CJEU judgments, such as the institutions involved. The pruned judgment texts are then converted into a TF-IDF matrix. Finally, substantive similarity is captured using cosine similarity on the TF-IDF matrix. As an alternative measure, I apply the same procedure using the subject matter classification scheme available on EUR-Lex (2023).

The third cluster of features relates to characteristics of the court. Here I measure the chamber size which proxies the degree of controversy around a dispute (Larsson and Naurin, 2016). Furthermore, I explicitly control for whether a judgment is delivered by the grand chamber, as opposed to smaller chambers which consist of a subset of three or five judges and handle more routine cases.

The fourth and final clusters relates to the wider political context. This includes, first, the country of origin for the case in question. Countries are weighted by their population size (`country of origin weights`). Court judgments may raise attention particularly in countries where the dispute originates, and larger member states may wield greater political influence over the court. A second measure taps into the salience of the case among member state member state governments by measuring how many of them weigh in on the case by filing amicus curiae briefs `num obs vec`. Another version of this variable weights governmental briefs with population size as a proxy for political pressure (Carrubba, Gabel and Hankla, 2008; Larsson and Naurin, 2016).

Finally, a lagged version of the outcome variable, the cite count of substantively similar judgments weighted by the overall number of judgments in the past is included (`lagged hit ratio`). This autoregressive component takes into account that over the time observed, there are temporal trends to affect the production of precedent.

### 2.3 Model

The objective of this project is to predict the number of times a given judgment will be cited in the future, which is a regression problem. I use three different machine learning algorithms suited to this type of problem, a Random Forest regressor, extreme gradient boosting (XGBoost), and a Support Vectors Machine (SVM). For each one, hyperparameters are set through grid search. As a benchmark algorithm, I add Ordinary Least Squares regression. Prior to estimation, the data are divided into train and test sets using 80 and 20 percent of the observations, respectively. Data are normalized using a standard scaler.

## 3 Results

Table 2 compares the models in terms of their goodness of fit on the test set. The key performance metrics used are the root mean squared error (RMSE) and Mean Absolute Error (MAE). As a benchmark, I compute RMSE and MAE of the mean of the cite count (RMSE=2.43, MAE=1.30): Judging from the RMSE, estimating that a judgment receives an average number of cites will generally be off by just above 2.4 citations. This is a useful benchmark, as it captures how much we can tell just knowing the distribution of the outcome variable, and without taking into account any explanatory factors.

According to Table 2, even a linear regression model reduces the error substantially, bringing RMSE down to 2.16 (a reduction of 11.2% in terms of RMSE and 5.2% in terms of MAE). Support Vector Machines perform even worse. More sophisticated ML algorithms like Random Forest and XGBoost improve goodness-of-fit further, even though these improvements are more incremental. Random Forest arrives at a RMSE of 2.13 (an error reduction

of 12.43% RMSE) while XGBoost arrives at a RMSE of 2.1072 (error reduction of 13.41%). The race between the two is close. The error reduction of more than 13% is quite substantial, particularly taken into account that the problem of predicting precedential value has both supply- and demand-side factors, and the project here is limited to investigating features on the supply-side.

Table 2: Performance metrics of algorithms optimized through grid search

| Algorithm | RMSE | Error Reduction RMSE | MAE | Error Reduction MAE |
|---|---|---|---|---|
| E[y] | 2.4335 | (benchmark) | 1.3032 | (benchmark) |
| SVM | 2.2212 | 8.72% | 1.2360 | 5.16% |
| Linear Regression | 2.1615 | 11.18 % | 1.2309 | 5.55% |
| Random Forest | 2.1309 | 12.43% | 1.2250 | 6.00% |
| XGBoost | 2.1072 | 13.41% | 1.2186 | 6.49% |

Figure 2 plots the predictions of the best model[1] (y-axis) against the true cite counts in the test set (x-axis). The scatter plot shows an upward-trending cluster of data points, illustrating that the models pick up a substantial amount of signal from the data. Still, the correlation is far from perfect.

[1]XGBoost with the following hyperparameter settings, found through grid search: learning rate: 0.05, maximum depth: 3, minimum child weight: 2, number of estimators: 120

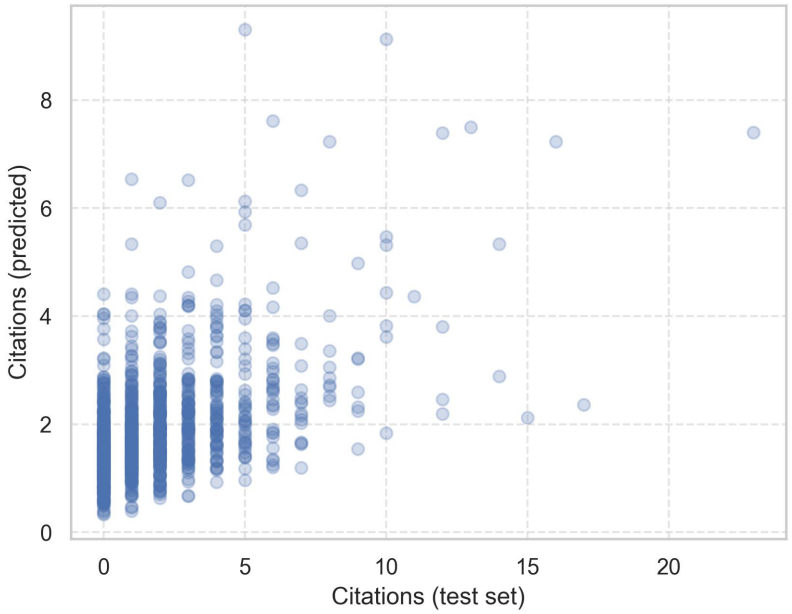


Figure 2: Comparison of true against predicted cite count

Figure 3 is more informative about the tendency of the model to over- or underestimate the true precedential value. It show that the prediction error has a long negative tail, but that the bulk of the density is just above 0. The model tends to be above, if relatively close to, the true cite counts, but large under-estimates do occur.

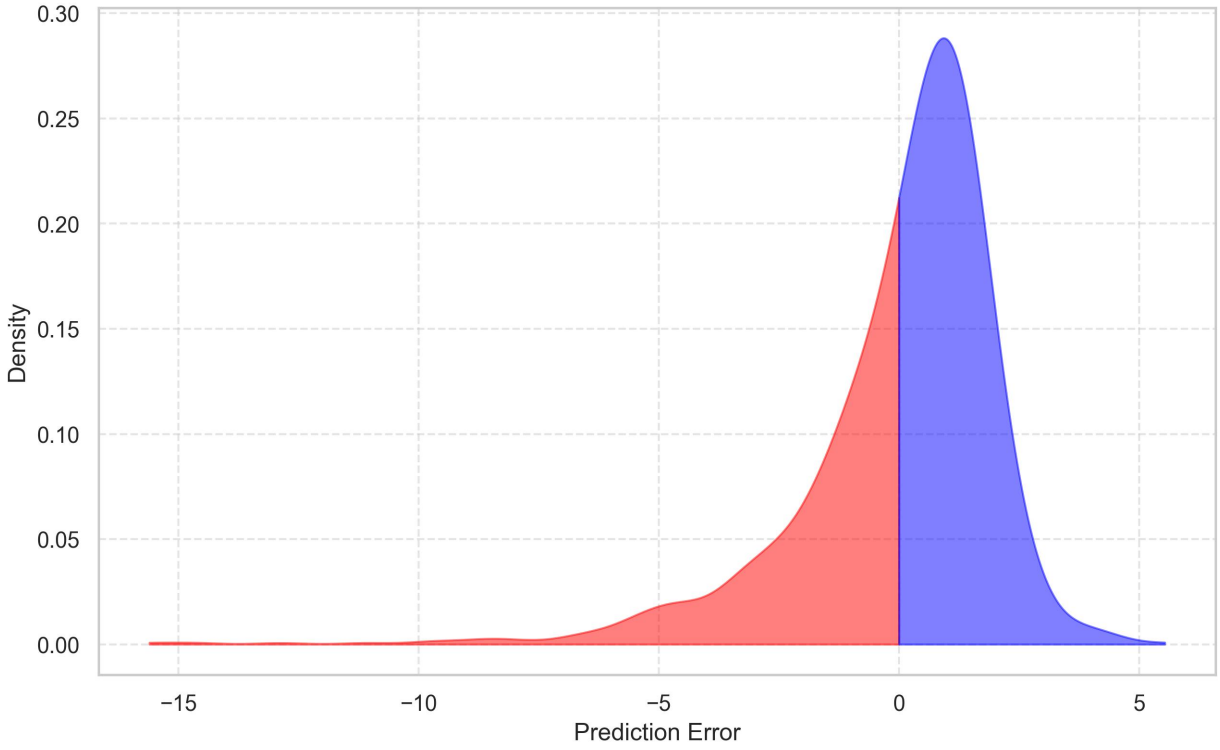


Figure 3: Density plot of prediction errors

Figure 4 shows the size of prediction error over time. Prediction error tends to be the highest at the beginning and the end of the observation period. Overall, the predictive quality seems relatively stable over time. Temporal influences on predictive quality remain to be determined.

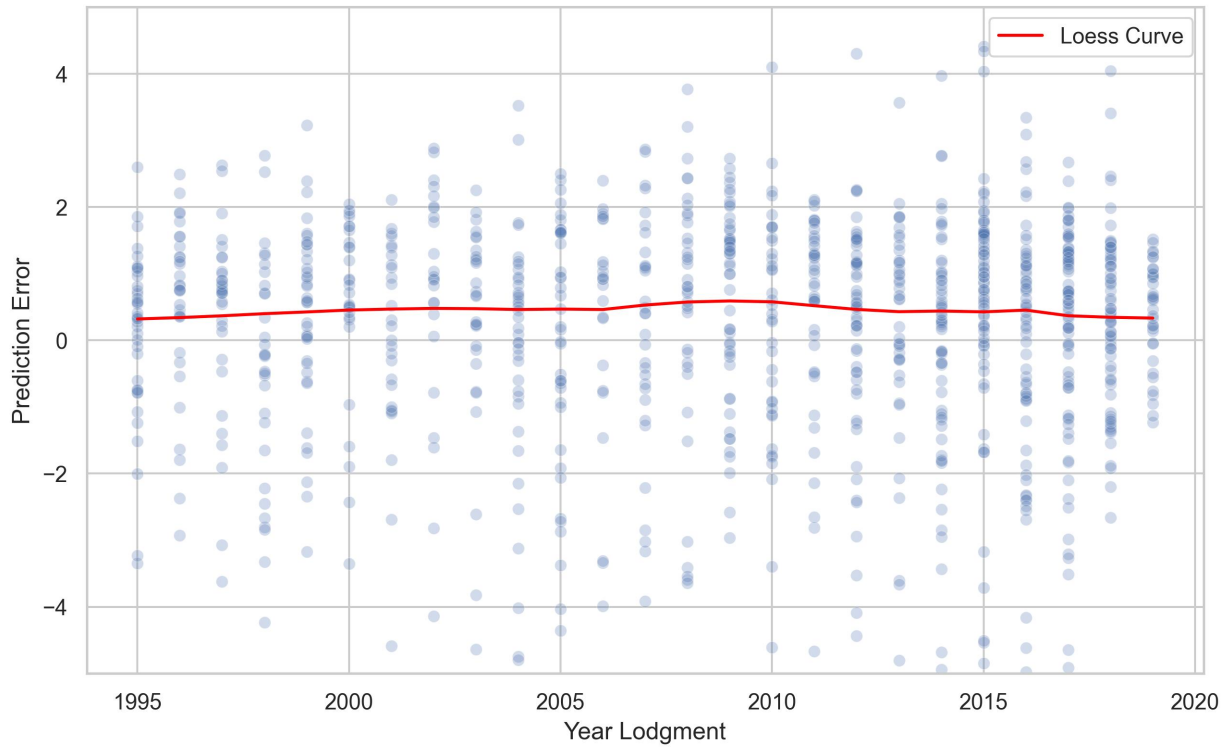


Figure 4: Temporal trend in prediction errors

Figure 5 finally plots the 25 most important features of the model. What stands out are `textlength` (cluster 2: crafting of the judgment), `grand chamber` (cluster 3: features of the court) and `amicus curiae weights` (cluster 4: political factors). Legal substance (cluster 1) also plays an important role, which is particular pronounced in the important weight of `prior touches min vec`.

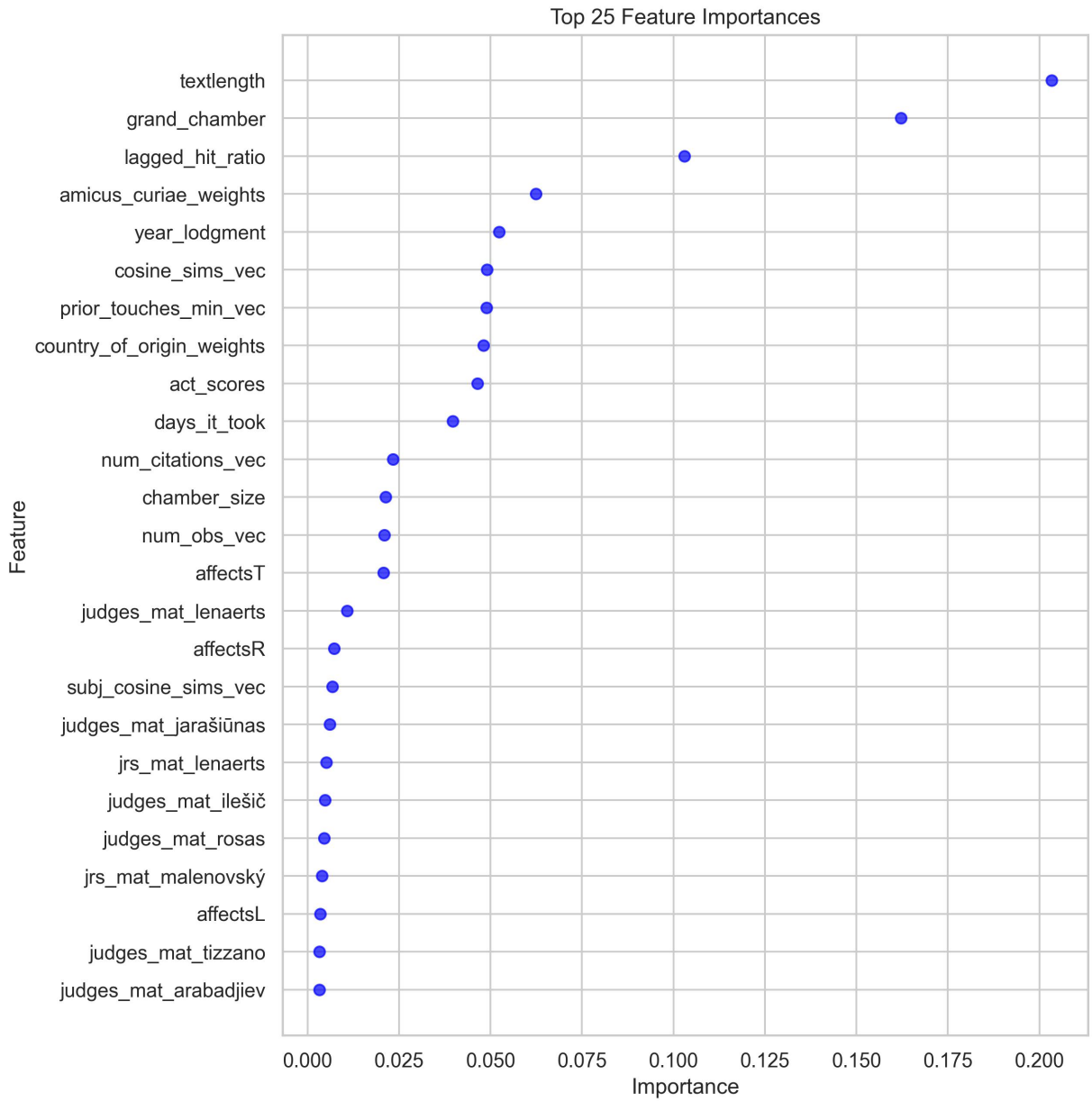


Figure 5: Feature importance: Top 25

## 4 Discussion and Conclusion

In this paper I present results from an ongoing project applying machine learning algorithms to the investigation of case-law development in the EU. I investigate factors to determine which judgments of the CJEU get cited again and again, shaping the path of the legal

integration of Europe. Using Random Forest and XGBoost algorithms, I find that different clusters of feature predict the cite count of judgments. The most important features relate to the way judgments are crafted (cluster 2), characteristics of the court (cluster 3) and the (political) context (cluster 4). I also find legal substance matter (cluster 1) to play an important role.

However, predicting the precedential value of judgments remains a problem that is not easy to track. This project remains necessarily limited to supply-side factors, that is characteristics inherent or related to the judgment in question What is omitted is the demand side, including questions such as whether judges with certain legal priors or political preferences differ in their willingness to take precedent into account. As long as data availability does not allow for a comprehensive examination of this subject, any model will offer predictions far from the optimum.

## Bibliography


Carrubba, Clifford J, Matthew Gabel and Charles Hankla. 2008. "Judicial Behavior under Political Constraints: Evidence from the European Court of Justice." *American Political Science Review* 102(4):435–452.

Carrubba, Clifford J and Matthew J Gabel. 2014. *International Courts and the Performance of International Agreements: a General Theory with Evidence from the European Union*. Cambridge: Cambridge University Press.

EUR-Lex. 2023. "Access to European Union law." Available online: http://eur-lex.europa.eu/homepage.html?locale=en, last retrieved on November 30, 2023.

Larsson, Olof and Daniel Naurin. 2016. "Judicial Independence and Political Uncertainty: how the Risk of Override Affects the Court of Justice of the EU." *International Organization* 70(2):377–408.

Lupu, Yonatan and Erik Voeten. 2012. "Precedent in International Courts: A Network

Analysis of Case Citations by the European Court of Human Rights." *British Journal of Political Science* 42(2):413–439.

Pollack, Mark A. 2003. *The Engines of European Integration: Delegation, Agency, and Agenda Setting in the EU*. Oxford University Press.

Roberts, Margaret E, Brandon M Stewart and Richard A Nielsen. 2020. "Adjusting for Confounding with Text Matching." *American Journal of Political Science* 64(4):887–903.

Schmidt, Susanne K. 2018. *The European Court of Justice and the Policy Process: The Shadow of Case Law*. Oxford University Press.