Table 1: Accuracy and Kappa measures on test data for all the adjusted models.

model	accuracy	kappa
pred_rf	77.31%	19.43%
pred_gbm	75.00%	15.86%
pred_dnn	73.65%	13.78%
pred_lasso	73.46%	13.49%

Contents

In the previous sections, we identified that the distribution of litigations can be considered as random and identified some evidence to judicial favoritism. In this section, we evaluate if the characteristics of politicians are relevant to predict whether they will win the litigation, when compared to other variables from the litigation.

For this, we adjusted four predictive models considering the outcome of the litigation as response variable and the characteristics of the case and the political as predictors. We split 80% of the dataset for training and used the 20% to test the accuracy of each model. The adjusted models were i) Logistic regression with Lasso regularisation (Tibshirani (1996)), ii) Random Forest (Breiman (2001)), iii) Gradient Boosting (Friedman (2001)) and iv) Deep neural networks using dense layers (Goodfellow, Bengio, and Courville (2016)).

Table 1 shows accuracy and Cohen's Kappa (Landis and Koch (1977)) metrics to the ajusted models. The model with greatest performance was the random forest method, with 77% out of sample accuracy. We chose this model to proceed with our analysis.

Figure 1 shows the variable importance plot for the random forest model, based on mean decrease gini. The most important variable was the judge's tenure, followed by the case claim and the judge's salary. The candidate's election share followed the list, along with the age and role of the politician as the defendant in the lawsuit. The other variables had small contributions to predict the model results. In order to continue the analysis, we consider the four most important variables detected.

One difficulty to interpret nonlinear models such as random forests is to evaluate the direction of the effect of explanatory variables on the predicted value. One way to solve this problem is using Partial Dependence Plots, PDPs for short (Friedman (2001)), which can be used to evaluate the marginal effect of an explanatory variable x_s on the predicted value of the model integrating the others out:

$$\hat{f}_{x_s}(x_s) = \mathbb{E}_{x_c}[\hat{f}_{x_s}(x_s, x_c)] = \int \hat{f}_{x_s}(x_s, x_c) d\mathbb{P}(x_c),$$

where $\hat{f}$ is the adjusted prediction function and x_c is the set of all explanatory variables used to build $\hat{f}$. In the case of a linear model, for example, $\hat{f}_{x_s}$ is always a linear function.

Figure 2 shows the PDPs of the four most important variables detected in the random forest model. The rug on the x-axis shows the concentration of observations in the test base. The relationship between the judge's tenure and the outcome of the litigation is complex, but it is possible to identify that the probability of victory is lower in the extremes and higher in the center, with a slight upward trend in the probability of victory as the tenure grows. With respect to the second variable, we can say that the greater the value of the cause, the lower the probability of victory. The relationship with payment is also complex, because there are two regions of greater and lesser probability of victory in wages less than fifty thousand Brazilian Reais. Finally, the election share variable shows two levels of probability of victory, indicating that from a election share greater than -0.1, the politician will have a higher probability of victory, with a slight tendency of increase in the probability of victory as the election share increases, between -0.1 and 0.1.

In summary, for the purposes of our analysis, the results of this section indicate that the characteristics of the judge, the value of the cause, and the election share of the politician are the most important in determining

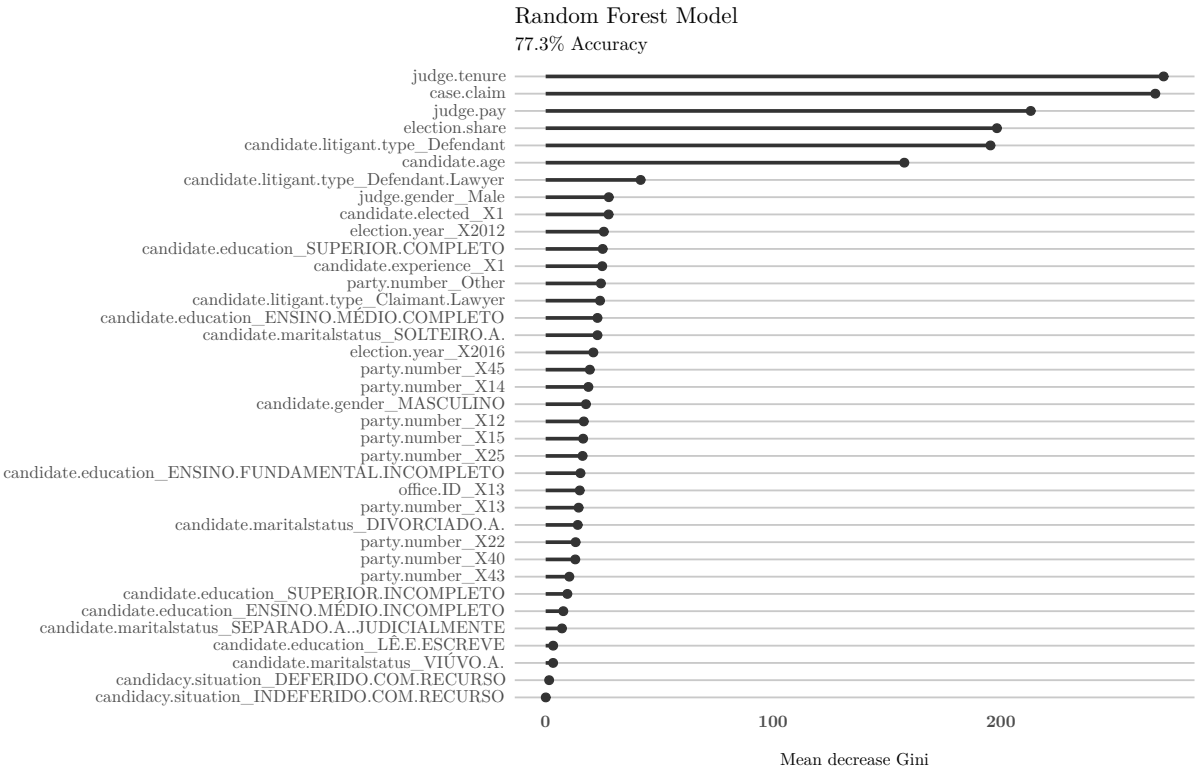

Figure 1: Variable importance plot for the random forests model.

whether the decision is favorable to the politician. In particular, the election share has two levels of probability of victory, with an approximately linear and positive effect between -0.1 and 0.1.

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

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*. JSTOR, 1189–1232.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT press.

Landis, J Richard, and Gary G Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics*. JSTOR, 159–74.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1). Wiley Online Library: 267–88.

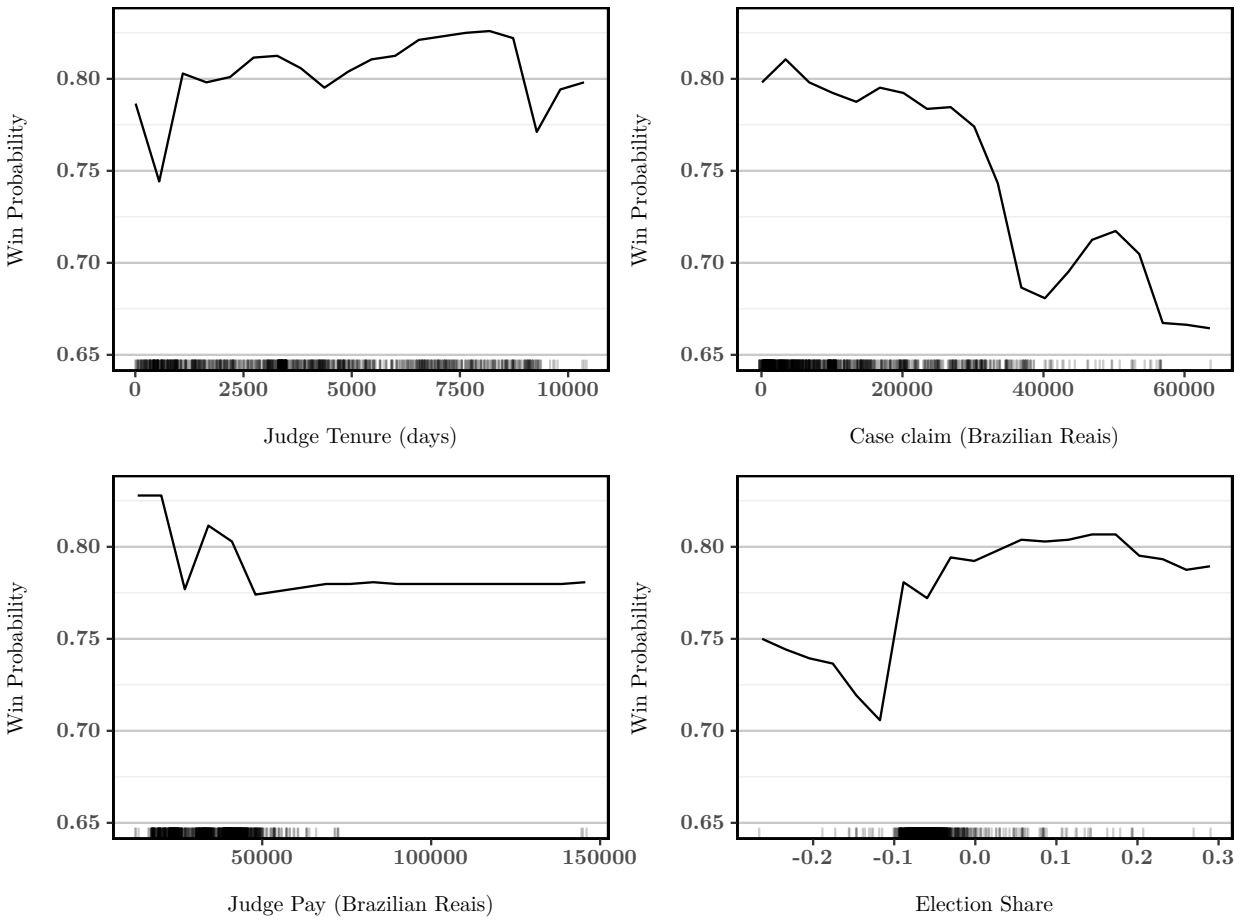

Figure 2: Partial dependence plots for the most important predictors according to the random forests model.