Princeton Gavel, Inc.

6 Lawrence Drive

Suite 102

Princeton, New Jersey, USA

info@princetongavel.com

September 7, 2019

The Law Firm Name and Address

Dear Law Firm name,

Thank you for the opportunity to review a selection of your employment law cases. We have concluded our preliminary review, the results of which we relay in this letter and in the attached analyses. In short, these cases appear promising for large scale predictive work. There is a genuinely discrete set of features, and with a larger data set, clear trends should be established and accurate outcome predictions made.

**What We Did**

The data set is “messy,” in the sense that is comprised of scanned-in pdfs, not all of which are text documents. In addition, individual case files seem not always to be exhaustive (there appear to be missing pleadings and other documents). Thus, our initial task was focused on natural language processing (NLP): we wrote code that could automatically organize the materials and extract key features of the scanned cases and populate a data frame with these features. With the data properly organized, we then conducted initial analyses both for the purposes of gathering information and to determine whether we would be able to run more advanced analyses, ranging from regressions to decision trees to more complex machine learning approaches, with the present data set.

**What We Found**

To review our preliminary analyses, please see the attached.

We were sent 525 cases. Of these, 87 appeared to contain final decisions (decision values). Other files were incomplete for a variety of reasons: some by clerical error, some by time (the case was ongoing and had not yet concluded), and so on. Of these 87 cases with decisions, there were generally two outcomes: dismissal by summary judgment, in which costs were imposed on the plaintiff, or resolution by final judgment. Summary judgment outcomes were coded as R$0 results. For the cases that went all the way to final judgment, judgments ranged from roughly R$1.000 to R$60.000, not including costs imposed by the court.

Plaintiffs’ demands were distributed about a median of R$20.000, and decision values were roughly distributed at 50.4% of the demand value (i.e., R$10.000), although the mode was at 0% (i.e., the case was dismissed). We mapped these results according to the plaintiffs’ attorneys (see p. 8 and 11 of the attached analyses) and judges (p. 9 and 12) handling the cases, of which our NLP identified 63 different plaintiff attorneys and 53 different judges. Our NLP identified a limited set of defense attorneys: 8, although some case files were missing this information. We mapped decision values in terms of defense attorneys as well (p. 7 and 10).

We then explored correlations between key features and decision values, finding, as predicted, highly linear relationships between decision values and evidentiary counts and also demand values. We reran some of these analyses with a subset that did not include cases resolved by summary judgment (see Appendix), and the relationship to evidentiary counts and demand values remained.

Our initial analyses showed that we were underpowered even for regression analysis of a full model: for the number of features we had extracted, we would have needed at least 500 cases (with outcomes) for such analysis to be meaningful. Thus, we concluded our preliminary review at this point and did not progress to prediction.

**What We Recommend**

Perhaps the largest challenge in working with this dataset was its messiness. We wrote a fair amount of code so that we could process these cases by machine. We also had to adjust the existing code from which we drew upon so that the NLP worked with the Portuguese language, rather than our accustomed language of English. However, this extra work has a clear benefit: it leaves us well-prepared and ready to parse additional cases.

Our largest limitation was in number of cases. With less than 90 decision values, we had a very small sample with which to work. This was not a significant problem, as it did enable us to orient to employment law decision making in Brazil and adjust our methods appropriately. However, it did prevent us from doing our core work: prediction. Thus, we come to our first recommendation: provide a larger dataset, one with at least 500 decision values, but preferably one with 5.000 decision values, so that we can progress to regression and on to machine learning prediction. Without a large dataset, such analyses are not possible.

One of the features we identified—sheer quantity of evidentiary inputs—showed good promise for predicting case outcomes (correlation of .50 with all cases; correlation of .21 for subset that did not include cases resolved by summary judgment). However, this was just a crude initial assay. To provide a brief glimpse at how we would turn this assay into a valid method, consider the following. First, we would add more nuance to the approach, classifying evidence into subcategories. To do so, we would need to consult with an attorney in your firm who understands cases like these (his or her time commitment would be minimal), so that we can write code that encapsulates such understanding. Once the evidence is subcategorized, we would use NLP to find indicators in initial pleadings that correlate with subsequent production of key evidentiary subcategories. In this way, we would build backwards with the predictive weights until we are able to attach them to initial pleadings. In short, more and more predictive power would be located in initial pleadings, which would be most useful for your firm (as soon as a complaint is filed, you could have an indicator of likely case decision value).

In summary, we have two recommendations:

1. Provide a larger dataset, one that approaches 5.000 decision values.
2. Work closer with us, in order to field questions about aspects of the data and casefiles as we build out our predictive models.

We believe these labor disputes show good promise for our modeling capabilities, and we envision being able to provide you with early case predictors, so that you can properly triage and handle new disputes as they arise.

Thank you, and we look forward to continuing our collaboration with you.

Kind regards,

The Princeton Gavel Team