Santiago Larrain
Natnaell Mammo
ML Productions

Advanced Machine Learning for Public Policy
Mid-Quarter Report

Refresh of Project Proposal Goals and Revisions

Our goal is to be able to identify how each Supreme Court justice would vote on a newly received case. To accomplish this we plan to build and train a machine learning model for each individual justice using several pieces of information about the judge and case, including:

1. Past voting history
2. What the lower court's ruling was
3. The policy issue area
4. Text features extracted from the facts of the case

The Data

As we had previously stated as our goal, we successfully scraped the Oyez website for "facts of the case" for all Supreme Court cases starting from 1956. An issue we ran into was that the Oyez website blocked html scraping using Beautiful Soup. To get around this, we found the API endpoint that the website was calling and directly called it ourselves. While scraping the data, we noticed that the majority of cases prior to 1995 were missing "facts of the case." This was one cause that led to our decision to focus on judges of the past 20 year; the second cause was related to the second source of data we obtained.

To supplement the information from the Oyez website, we searched for a dataset that contained structured basic data about each case, most importantly how each justice voted. Fortunately http://scdb.wustl.edu/data.php had exactly what we needed plus more. One concern we had about how to predict a judge's decision as "yes" or "no" was that the outcome variable very much depended on who was labeled as the plaintiff versus defendant. This was an issue because if we took a case and switched the plaintiff and defendant, the facts of the case would stay relatively the same but the predicted outcome should be switched. Fortunately, we found a variable "direction" which labeled each justice's decision as conservative or liberal, not as agreeing with the plaintiff or defendant. These conservative vs liberal labels also varied from policy issue to policy issue (criminal procedure, First Amendment rights, federalism) and was well-outlined in the documentation. This would serve as our outcome variable.

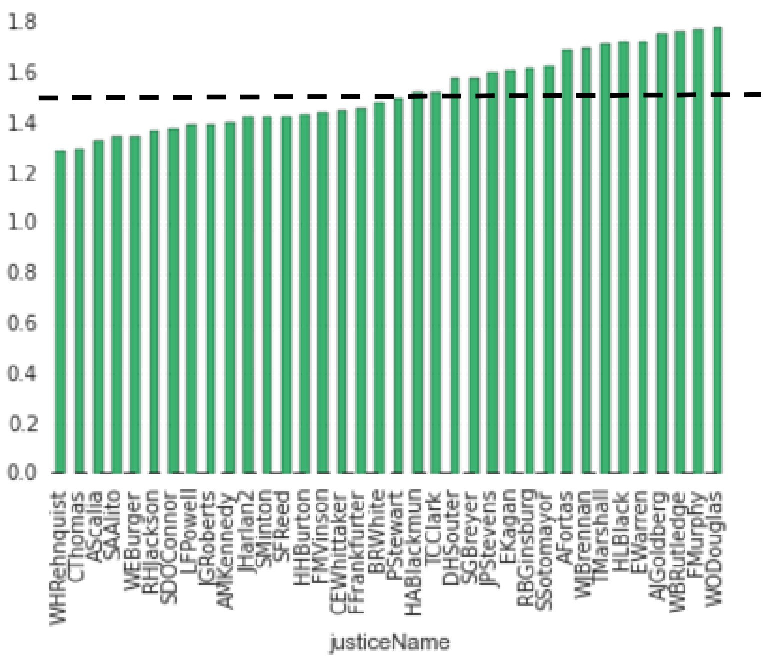
This leads to the second cause for focusing on recent justices: that the definition of a conservative vs liberal vote changes throughout time periods. So in order to maintain consistency in judge comparisons, we chose to focus on recent justices.

After scraping the Oyez website and downloading the supplementary dataset, we joined the two data sources by docket number.

Preliminary Analysis

Most people, including ourselves, already have ideas about which justices are liberal versus conservative. To get a preliminary sense of how accurate our perceptions are, we decided to create some visualizations of how justices vote. For all of the graphs below, a conservative vote = 1, liberal vote = 2.

Figure 1 Average Vote Type for Each Justice



Following plots show average vote type by policy issue for a given justice.

- Value > 1.6 is Blue. Value < 1.4 is Red. Value between 1.4 and 1.6 is Green.

Figure 2 Justice Clarence Thomas

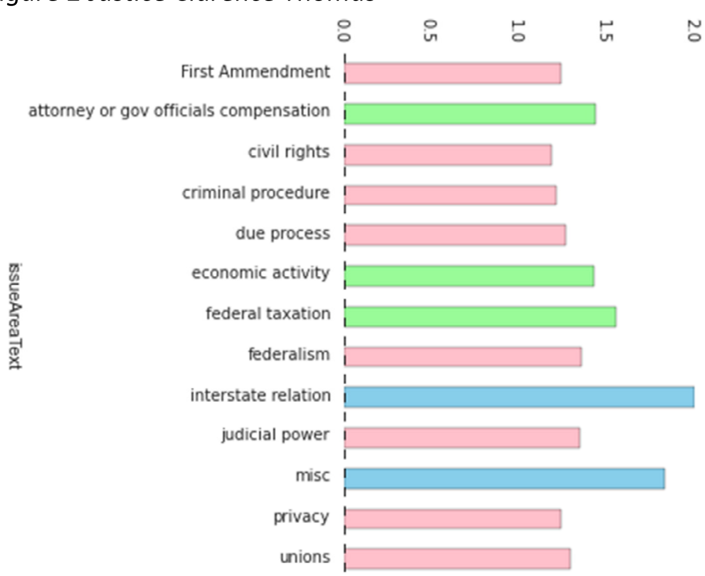


Figure 3 Sonia Sotomayor

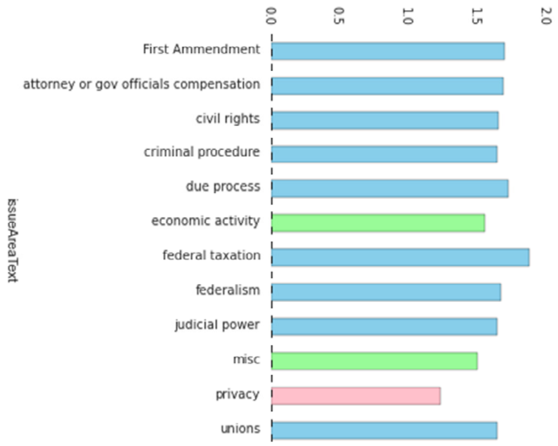


Figure 4 Antonin Scalia

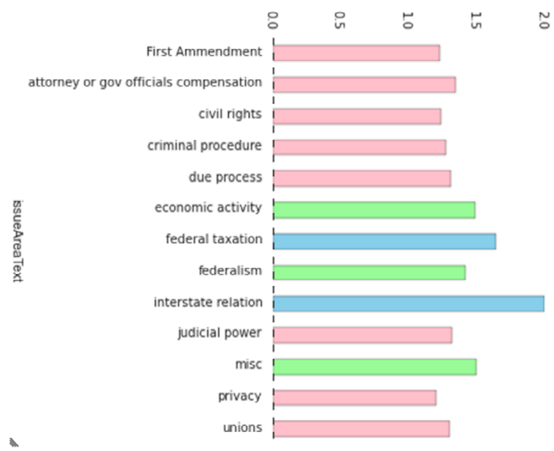


Figure 5 Elena Kagan

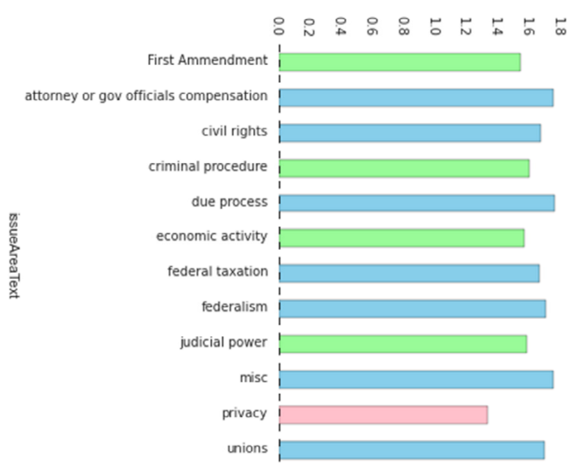


Figure 6 Heatmap to Visualize Correlations between Justices

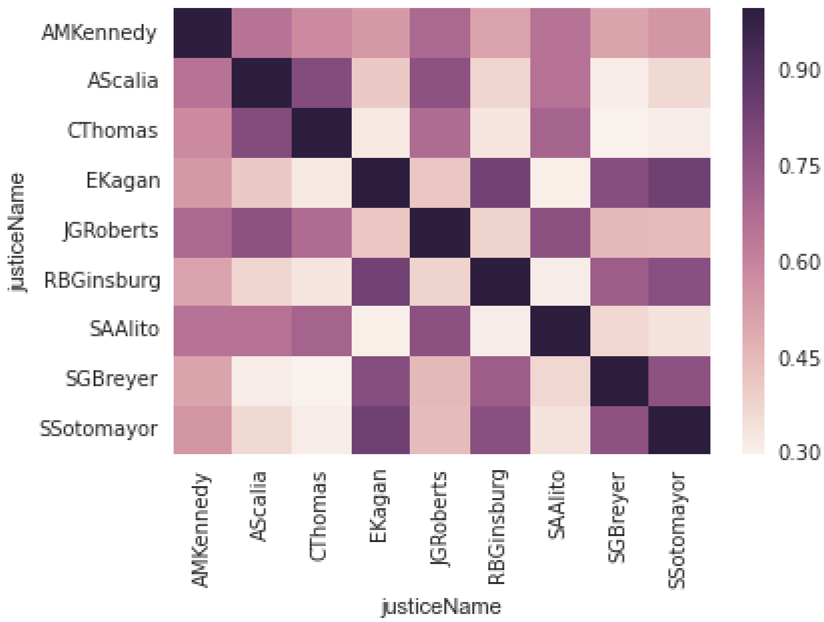


Figure 7 Correlation Table

justiceName	AMKennedy	AScalia	CThomas	EKagan	JGRoberts	RBGinsburg	SAAlito	SGBreyer	SSotomayor
justiceName									
AMKennedy	1.000000	0.654281	0.584671	0.542353	0.691095	0.514097	0.654954	0.511647	0.549968
AScalia	0.654281	1.000000	0.793745	0.410021	0.769038	0.379525	0.657727	0.311701	0.367536
CThomas	0.584671	0.793745	1.000000	0.324544	0.678583	0.335116	0.705017	0.300501	0.316785
EKagan	0.542353	0.410021	0.324544	1.000000	0.415570	0.834611	0.306481	0.788691	0.842929
JGRoberts	0.691095	0.769038	0.678583	0.415570	1.000000	0.381577	0.774492	0.454671	0.446280
RBGinsburg	0.514097	0.379525	0.335116	0.834611	0.381577	1.000000	0.312251	0.728031	0.779550
SAAlito	0.654954	0.657727	0.705017	0.306481	0.774492	0.312251	1.000000	0.371924	0.340696
SGBreyer	0.511647	0.311701	0.300501	0.788691	0.454671	0.728031	0.371924	1.000000	0.773160
SSotomayor	0.549968	0.367536	0.316785	0.842929	0.446280	0.779550	0.340696	0.773160	1.000000

Next Steps

Our next step is to create features that we will feed into our models. We plan to include indicators for policy issue type, the term, how the lower court voted, which circuit was the lower court from, and bags of words extracted from the facts of the case that exclude stop words. After creating the features, we will then evaluate how different machine learning algorithms perform using cross-validation.