Leveraging Voting History to Predict Support

Preserving the legacy of retiring thought-leaders

Andy Gonzalez

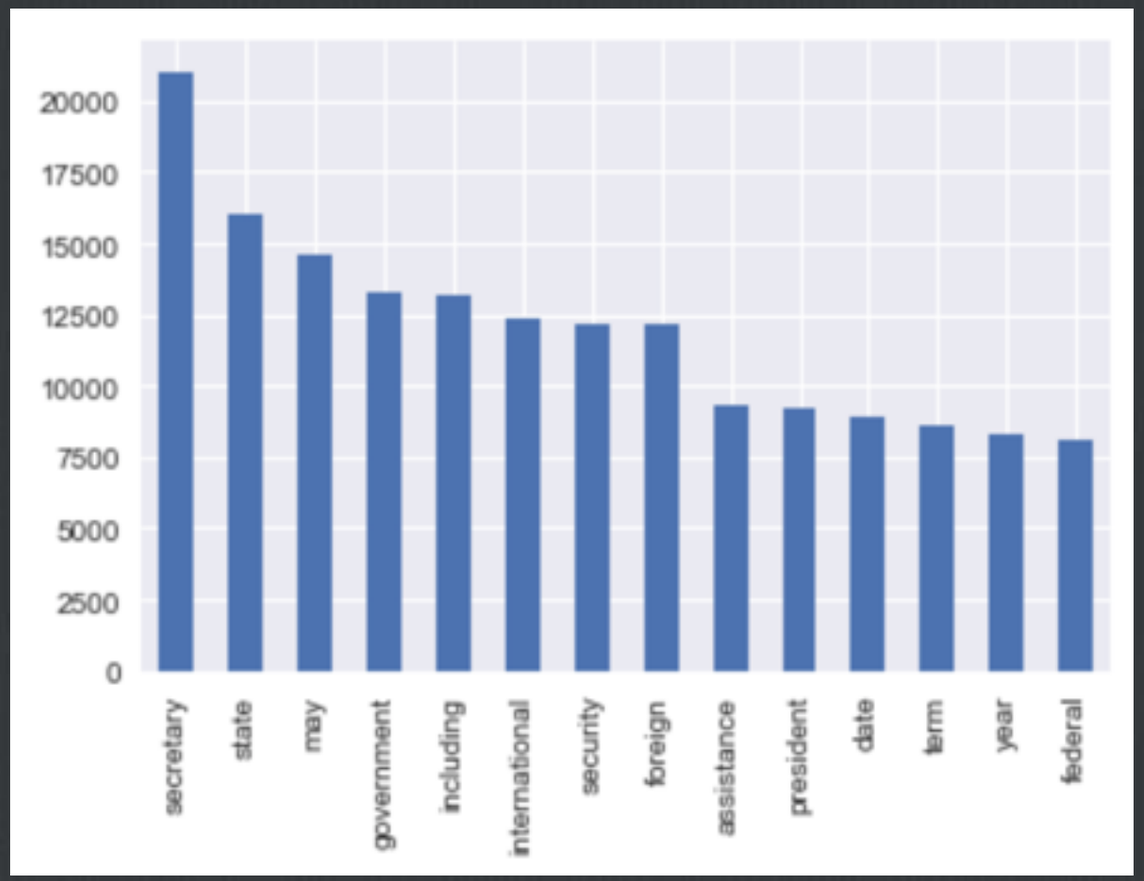
The Why:

- Original idea was to predict "good" legislation for Latin America
- More valuable and lasting to model voting behavior of the issue's thought-leaders

Rep. Ileana Ros-Lehtinen

The How:

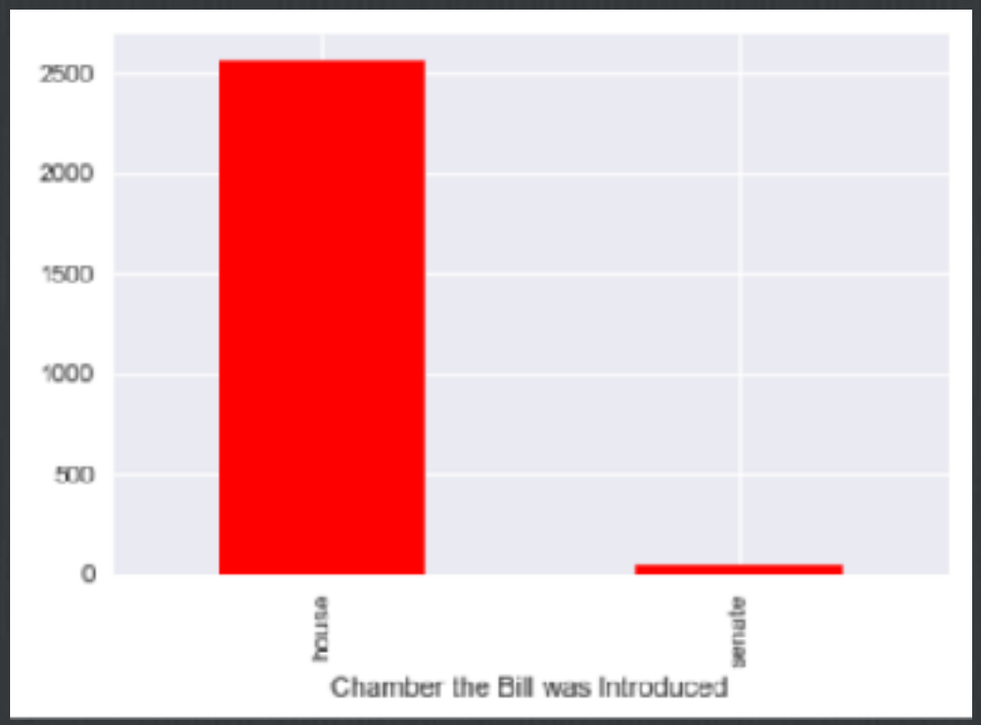
Most common terms



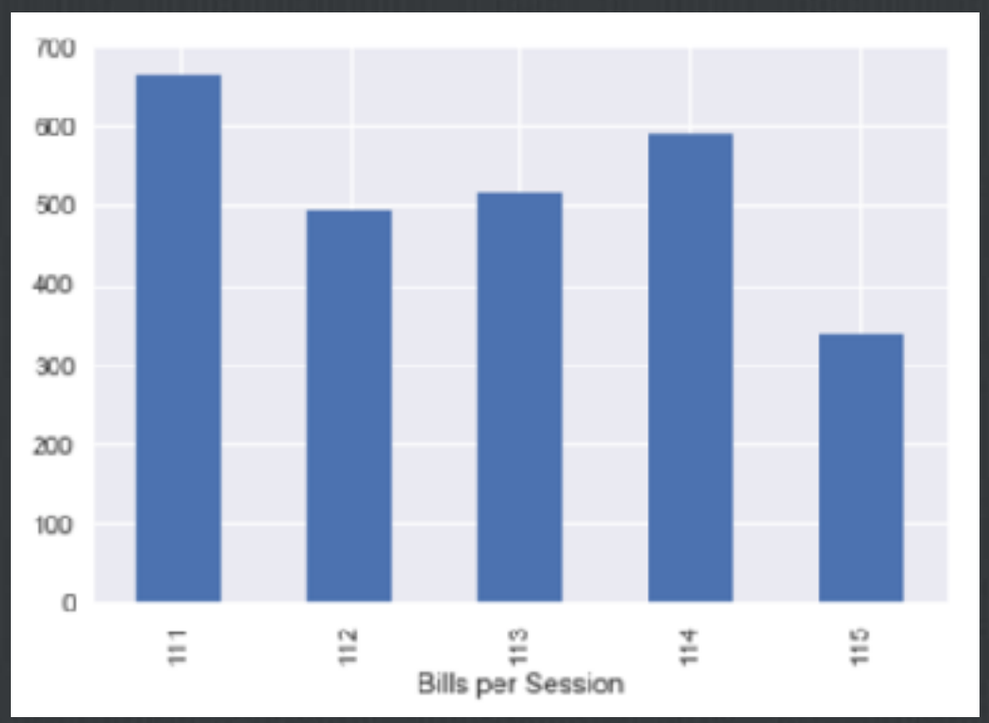
- ☐ Scraped web and APIs to get bill text and voting records (h/t ProPublica)
- ☐ Count Vectorizing IRL Sponsored legislation in the past 4 sessions of Congress
- ☐ Running different models to determine which could pick out the signal best

In the Data Set

Bills per Chamber



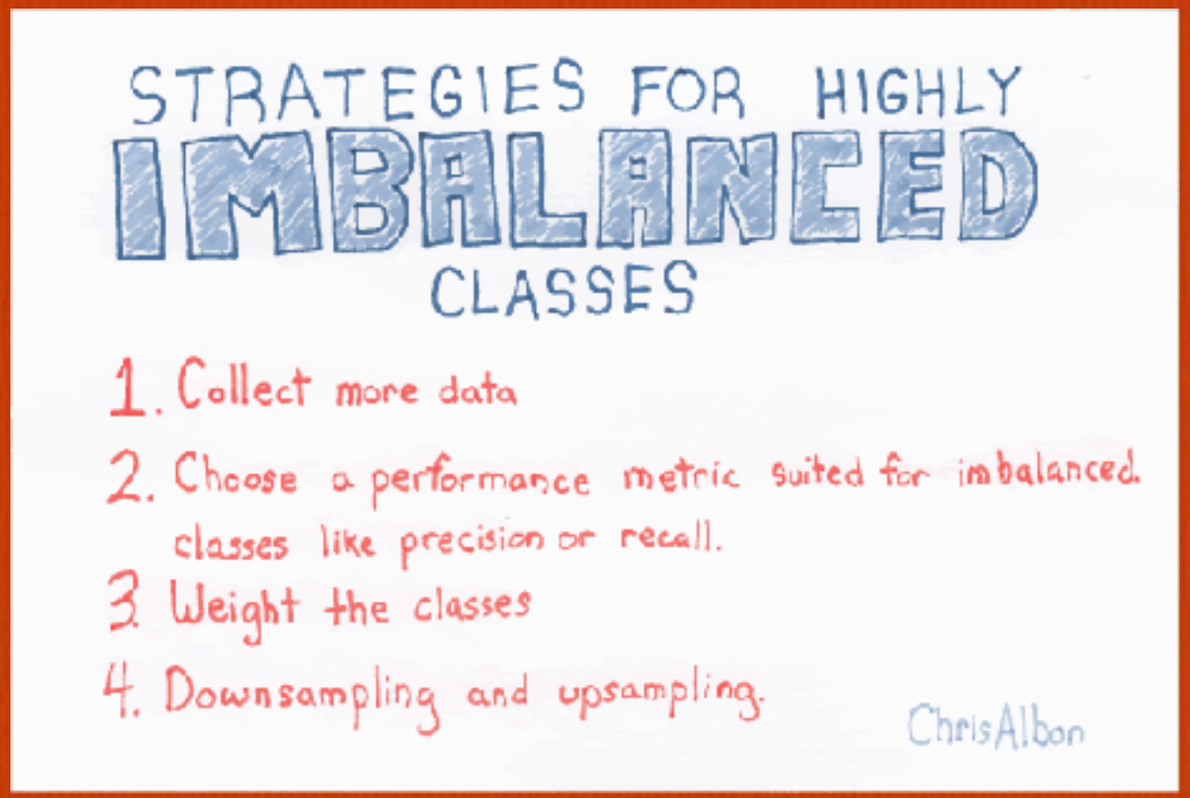
Bills per Session



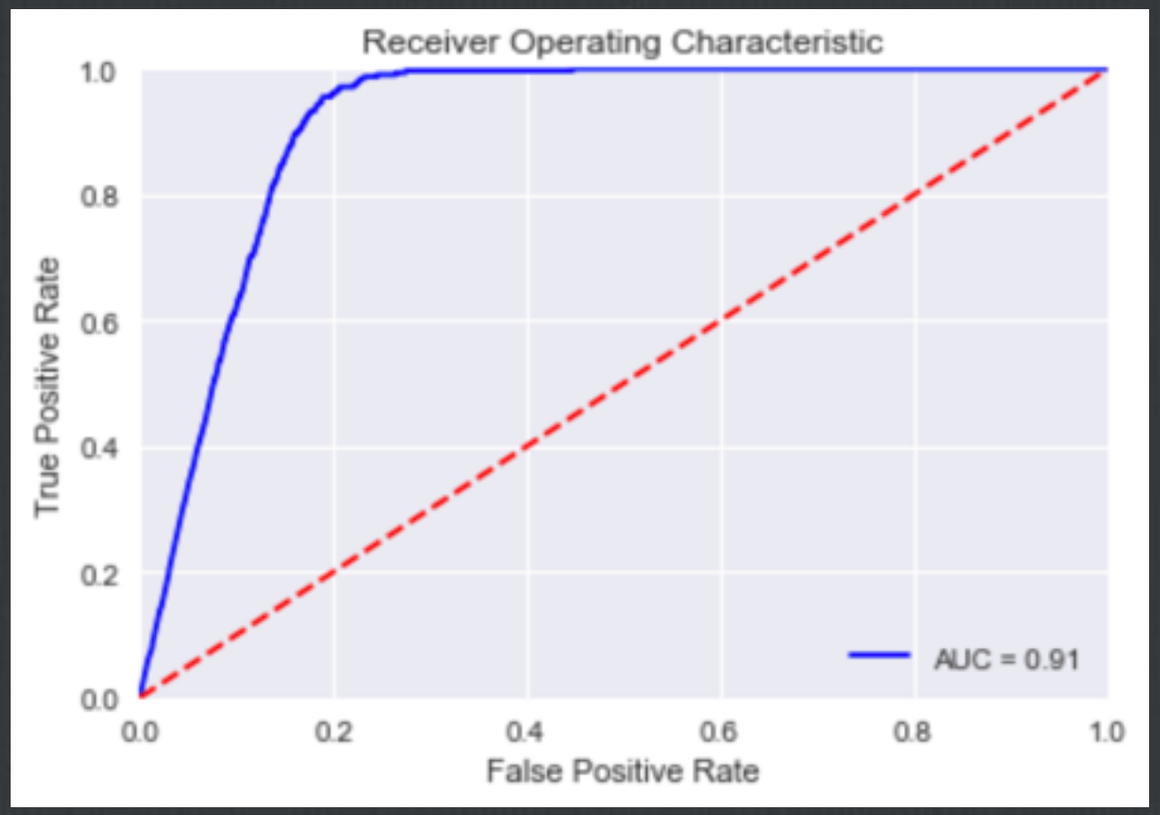
The Issues:

Imbalanced Classes

Generic Legislative Language



The Results:



- Best Model: Logistic Regression using TF-IDF vectorized text
- The final cross-validated accuracy: 82%
- AUC metric for the chosen logistic regression: .69

Feature Importance and Model Performance

Feature Importance

Feature	Importance
may used	0.010387
appropriated	0.009593
budget	0.008227
passed away	0.007729
cited	0.007429
covered	0.007324
title	0.007247
recorded	0.007196
commitment achieving	0.007177
expenditure	0.006648
state continues	0.006101

Results of Log Reg using TF-IDF

		prediction outcome	
		p	*n*
actual value	*p'*	291	21
	n'	386	1789

Key Takeaways

- ☐ Predicting voting behavior based on legislation support is difficult because often they use boiler plate language composed by someone other than the member.

- ☐ Given enough positive response data points, predicting member support for a legislation is possible to a high enough accuracy, but with mediocre recall.

- ☐ Logistic regression are a powerful and light tool to use on count vectorized text.

Questions?