Predicting Political Ideology Using Campaign Finance Data

Keri A. McKiernan and Joe A. Napoli

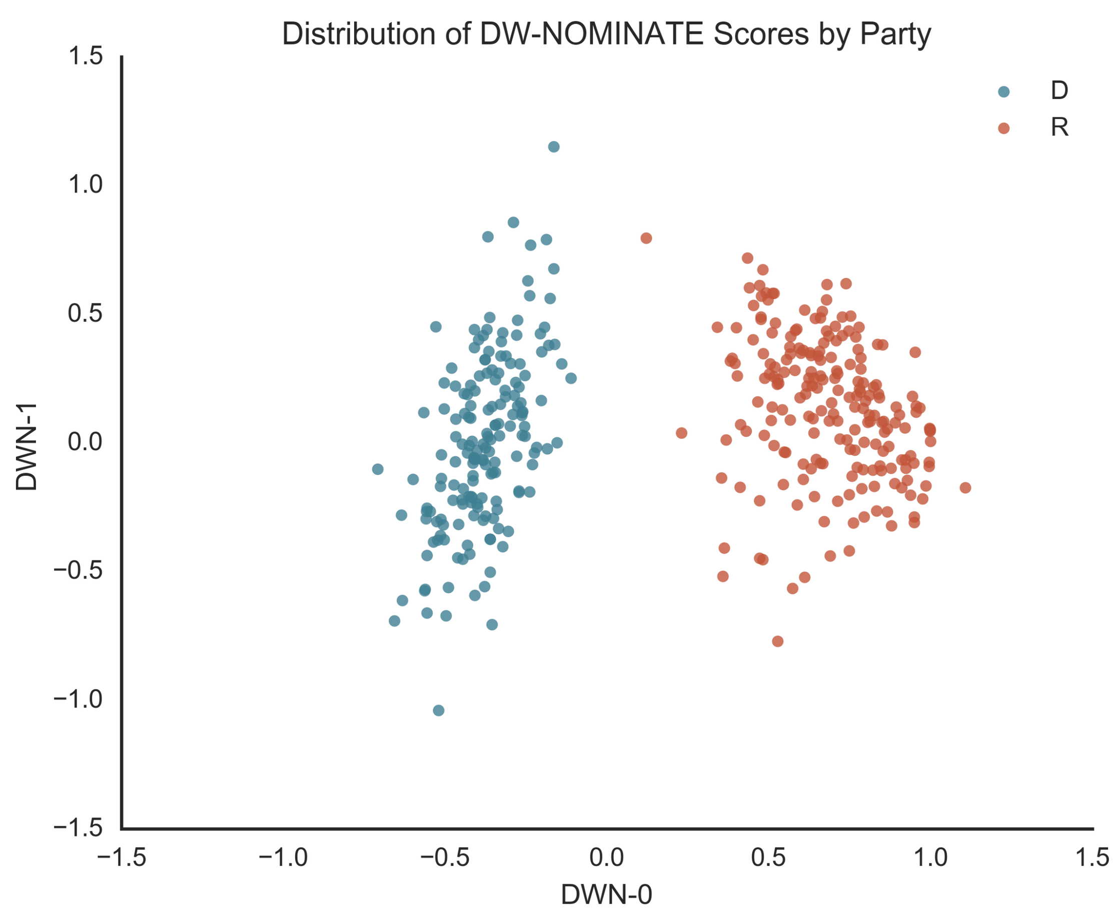
Department of Chemistry, Stanford University

Background

The research of Poole and Rosenthal has focused on quantifying political ideology [2] via the DW-NOMINATE method. For this method, ‘ideal point’ coordinates were obtained for legislators by maximizing the log likelihood function

$$\mathcal{L} = \sum_{t=1}^{T}\sum_{i=1}^{p_t}\sum_{j=1}^{q_t}\sum_{\tau=1}^{2} C_{ij\tau t}\ln P_{ij\tau t}$$

where $P_{ij\tau t}$ is the probability of voting for choice τ and $C_{ij\tau t} = 1$ if that probability accurately predicts the vote [1]. Indices j, i, and t sum over roll call votes, legislators, and legislative sessions, respectively. A common interpretation of the first coordinate is that it reflects the divide between the Republican and Democratic parties, whereas the second coordinate is more highly correlated with intra-party division. The distribution of these scores is illustrated below.



Here we aim to use machine learning methods to predict DW-NOMINATE scores using a candidate’s campaign finance information. It would be useful to be able to predict the ideal point of a candidate even before they have established a congressional voting record. Furthermore, the ability to do so would help elucidate a relationship between monetary contributions to candidates and the voting patterns those contributions may effect.

References

[1] Royce Carroll et al. “Measuring bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap”. In: Political Analysis 17.3 (2009), pp. 261–275.

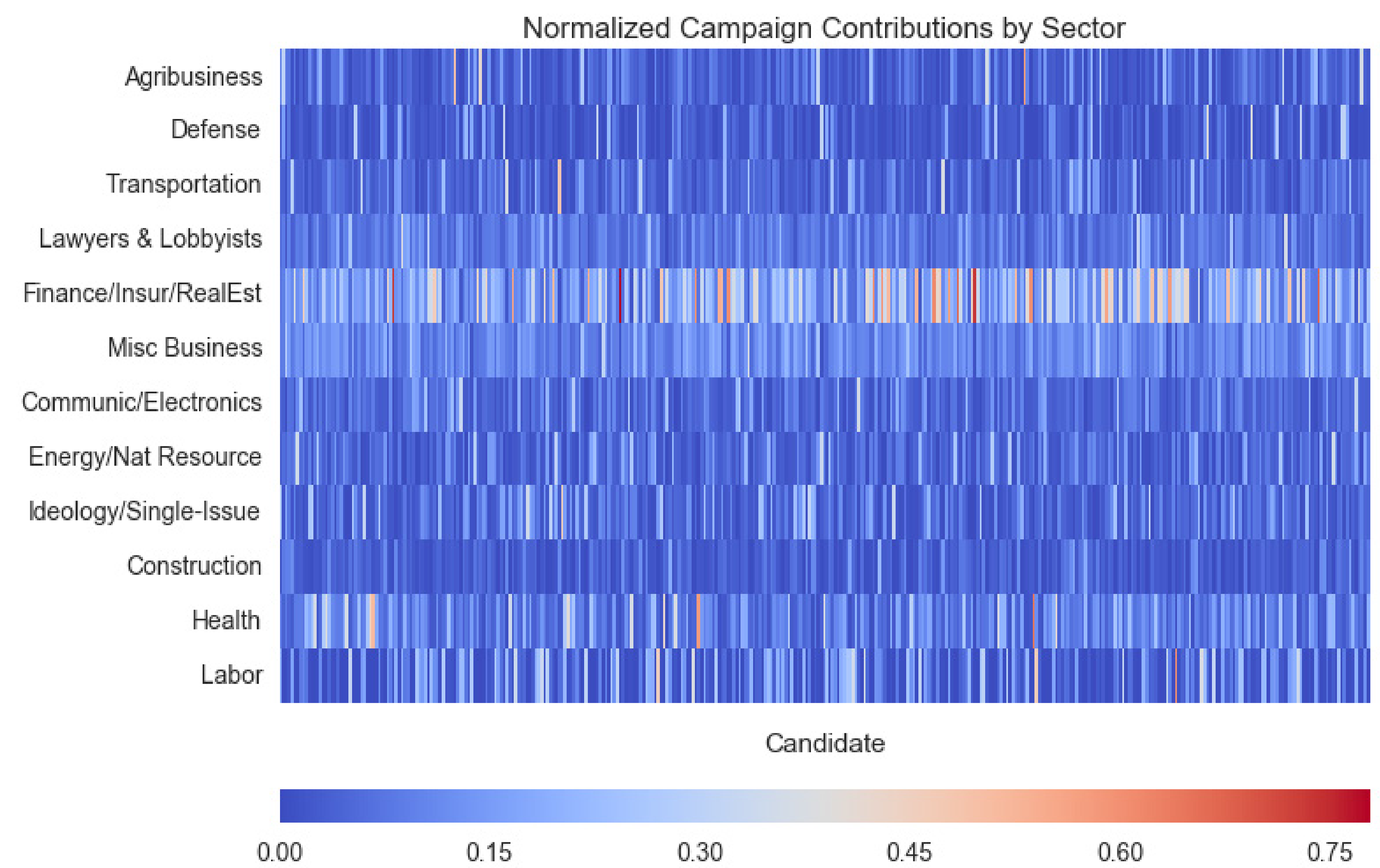
[2] K.T. Poole and H. Rosenthal. Congress: A Political-Economic History of Roll Call Voting. Oxford University Press, 2000.

[3] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

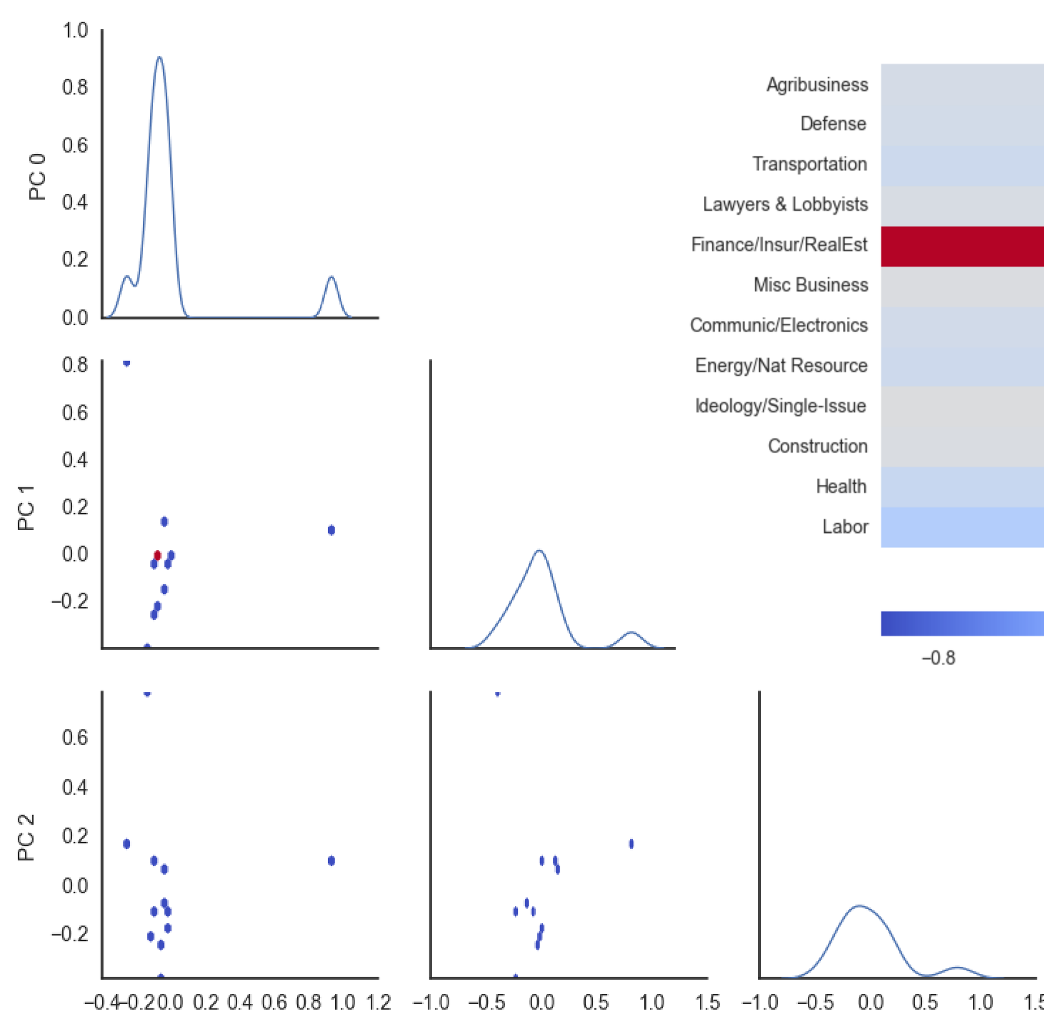
[4] Data provided by OpenSecrets.org Center for Responsive Politics

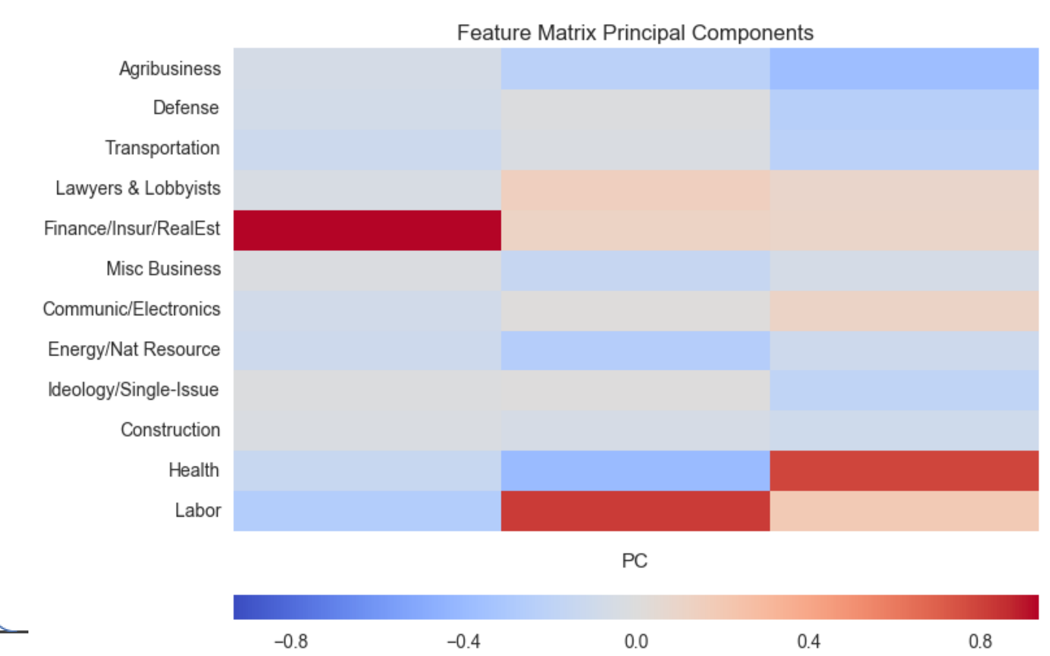
Training

Campaign finance data for a set of candidates (as determined by sampling from the DW-NOMINATE scores) was obtained using the Open Secrets python API [4]. Of the available data, we chose to partition finance streams by sector. For each sector, individual and PAC contributions were combined and normalized on a per candidate basis. A heatmap of the total feature set is illustrated below.



In order to evaluate the structure of the feature data set, PCA was performed on the above matrix. It appears the financial, labor, and health sectors specify the dimensions of largest variance.

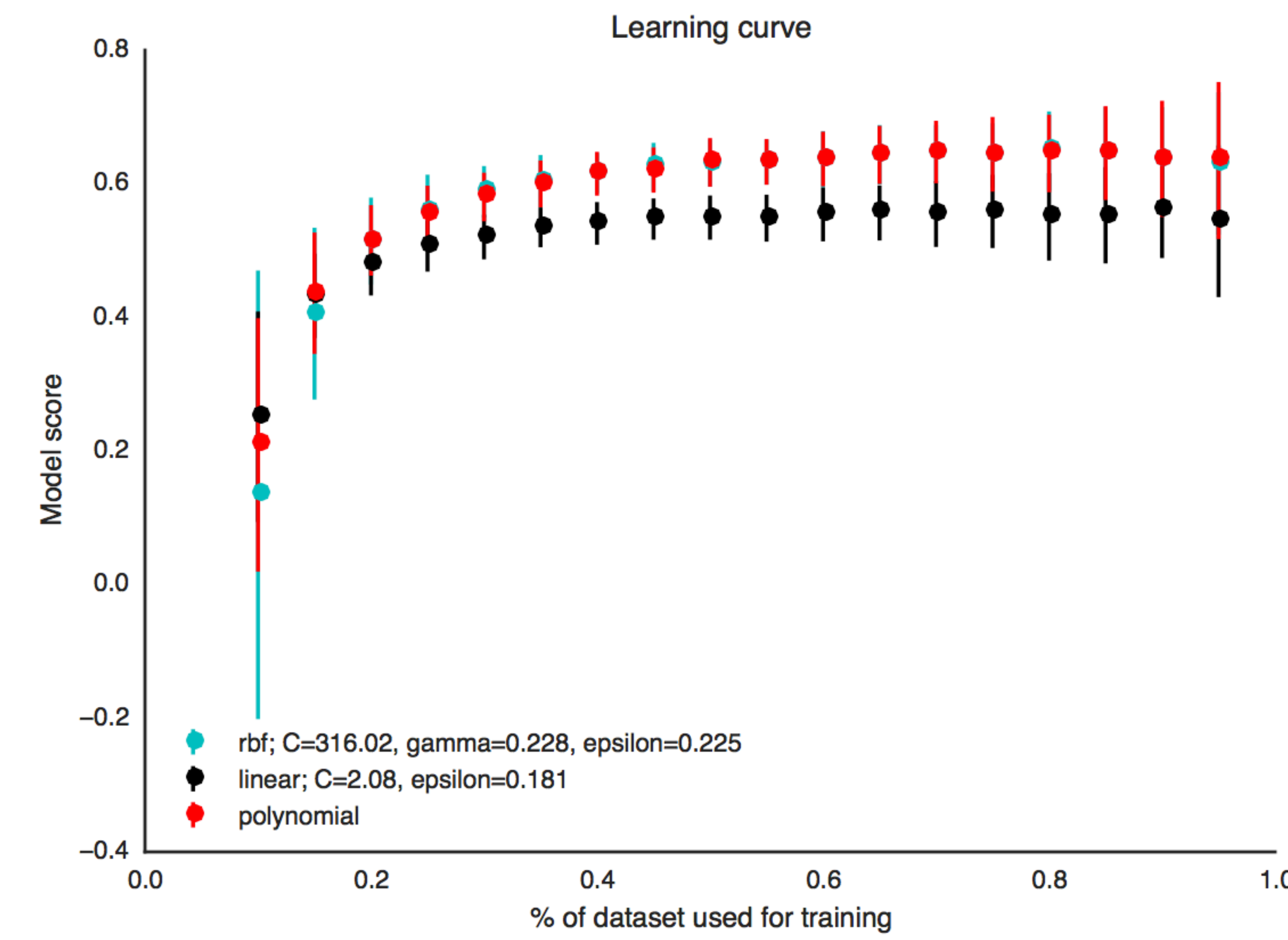




This data was used to train a range of support vector regression (SVR) models with variable kernel types, over a range of hyperparameters using the SciKitLearn cross_validation and GridSearch methods [3].

Validation

The optimized model was tested on a subset of data not used for training. In order to investigate the dependence of model score on training set size, learning curves, as a function of feature set size, were calculated for each model type. This curve is shown below. It was found that the gaussian kernel outperformed all other model types.



The support vectors for the RBF SVR model are depicted below.