Introduction to Machine Learning and Text Mining for Trade Policy

Problem: Lecture 1.

Ben Shepherd, Principal.

1. The demonstration script uses WDI data to predict the LPI. Let's imagine instead that we want to use LPI data to classify countries that are likely to see future rapid trade growth, and those that are not. We define "rapid" trade growth as a percentage increase in total exports that is in the top quintile of observed rates over a five year period.
2. The file "WDI data for export booms.csv" is arranged in the same way as the demonstration data. The variable of interest is export_boom, which is one for countries in the top quintile, zero for others. All other variables have been lagged by one period.
3. Import the data, clean it, and use interpolation to fill in missing values. You can do all of this by adapting the demonstration code.
4. Separate the data into training and testing subsamples.
5. Use the Lasso, Ridge, and 50-50 Elastic Net to estimate a model using levels and interactions of the available WDI variables to predict export booms. Use the training data only.
6. For the three models, compute prediction accuracy for the testing sample. Compare results. Which model do you prefer, and why?
7. How could this model be used in policy settings?