Dispute Outcome Expectations, v2.0

Robert J. Carroll Brenton Kenkel

July 22, 2019

Contents

1 Introduction **1**

2 Codebook **2**

3 How the Scores Are Calculated **3**
3.1 Data 3
3.2 Model Training 4
3.3 DOE Score Calculation 6

4 Replication **6**
4.1 Instructions 6
4.2 List of Files 6

5 Change Log **7**

1 Introduction

This document is the codebook and reference guide for the Dispute Outcome Expectations, or DOE, dataset. If you use the DOE data, please cite the following article:[1]

[1] Currently published online. Volume, issue, and pages not available as of 2019-07-22.

Robert J. Carroll and Brenton Kenkel. 2019. "Prediction, Proxies, and Power." *American Journal of Political Science*. https://doi.org/10.1111/ajps.12442.

You can download the current version of the DOE data from the DOE project Dataverse at https://dataverse.harvard.edu/dataverse/doe-scores. The replication code for the current version is available on GitHub at https://github.com/brentonk/doe-h2o.

The replication code and data used in the original *AJPS* article are available on Dataverse at https://doi.org/10.7910/DVN/FPYKTPP, though we recommend using the most recent version of the DOE data for any new empirical analysis.

Please contact Brenton Kenkel (brenton.kenkel@gmail.com) with any questions, corrections, or feature suggestions. Alternatively, you may post them to the project's "Issues" page on GitHub, https://github.com/brentonk/doe-h2o/issues.

2 Codebook

There are two data files: `doe-dir-dyad-2.0.csv` and `doe-dyad-2.0.csv`. These contain the DOE scores for all dyad-years (directed and undirected, respectively) from 1816 to 2012, where the universe of states is those tracked in the Correlates of War Project's National Material Capabilities data, v5.0 (Singer, Bremer and Stuckey, 1972).

Both datasets contain the same variables:

- `ccode_a`: Correlates of War country code for Country A.
- `ccode_b`: Correlates of War country code for Country B.
- `year`: Year of observation.
- `pr_win_a`: Estimated probability of victory by Country A.
- `pr_stalemate`: Estimated probability of a stalemate.
- `pr_win_b`: Estimated probability of victory by Country B.

In the directed data, Country A is assumed to be the initiator and Country B is assumed to be the target of the hypothetical dispute for which probabilities

are estimated. In the undirected data, values represent the average of the estimates for when Country A is the initiator and Country A is the target.

For convenience in data merging, the undirected data has two observations per dyad (one where the state with the lower country code is Country A, one where it is Country B). These contain the same observations, just with the labels flipped. For example, the entries for the USA (COW code 2) and the Soviet Union (365) in 1980 in the undirected data look like this:

```
   year ccode_a ccode_b pr_win_a pr_stalemate pr_win_b
1  1980       2     365   0.0553        0.872   0.0725
2  1980     365       2   0.0725        0.872   0.0553
```

Compare to the corresponding entries in the directed data, in which the Soviet Union has a higher predicted probability of success as the initiator than as the target:

```
   year ccode_a ccode_b pr_win_a pr_stalemate pr_win_b
1  1980       2     365   0.0573        0.899   0.0434
2  1980     365       2   0.102         0.845   0.0534
```

3 How the Scores Are Calculated

Broadly speaking, the process is the same as described in our *AJPS* article and its online appendix: we train an ensemble model on bilateral Militarized Interstate Disputes to find the function of material capabilities that best predicts the outcomes of these disputes. See the *AJPS* article for a detailed explanation and justification of this basic procedure.

DOE v2.0 (and onward) takes advantage of new machine learning technology that was not available when the original analysis for the *AJPS* article was run, as well as updated versions of the underlying datasets used for model training.

3.1 Data

We construct the training data for DOE 2.0 from the following sources:

- National Material Capabilities v5.0 from the Correlates of War project (Singer, Bremer and Stuckey, 1972)
- Gibler–Miller–Little Militarized Interstate Disputes[2] v2.1 (Gibler, Miller and Little, 2016)

Each observation in the training data is a dispute with exactly one state on each side. The response variable is the outcome of the dispute: victory by Country A (initiator), victory by Country B (target), or stalemate. Cases in which one side "yields" are coded as victory by the other side. Disputes that do not end in a victory, yield, or stalemate are excluded, leaving us with $N = 1{,}482$ training cases.

The training set contains 28 explanatory variables about each dispute:

- Year of observation
- Each country's CINC score (2 variables)
- Each country's raw value of the six CINC components: iron and steel production, military expenditures, military personnel, primary energy consumption, total population, and urban population (12 variables)
- Each country's share of the global total of each of the six CINC components (12 variables)
- Country A's share of dyadic CINC scores: $\text{CINC}_A/(\text{CINC}_A + \text{CINC}_B)$

Four of the CINC components (military expenditures, military personnel, primary energy consumption, and urban population) have some missing values in the training data. In all, just under 17% of the cases in our training data have a missing value for at least one explanatory variable.

3.2 Model Training

Model training was performed in a Linux environment (Ubuntu 18.04.2 LTS) on a machine with an Intel i7-7820X processor (8 cores, 16 threads). The analysis requires at least 16GB of RAM.

We use the following software:

- H2O v3.24.0.5 (run via OpenJDK v11.0.3)
- R v3.6.1 with packages:

[2]This differs from the original analysis, which used the Correlates of War MIDs.

- assertr v2.6
- caret v6.0-84
- foreach v1.4.4
- h2o v3.24.0.5
- tidyverse v1.2.1

All machine learning is performed in H2O.[3] We first train 409 component models, each predicting the dispute outcome as a function of the 28 aforementioned explanatory variables:

- 200 random forests (Breiman, 2001), randomly sampling across three tuning parameters (maximum depth, number of columns randomly selected at each split, sample rate).
- 108 gradient boosting machines (Friedman, 2001), exhaustively searching across three tuning parameters (number of trees, learning rate, sample rate).
- 101 elastic net–regularized multinomial logistic regressions (Zou and Hastie, 2005), exhaustively searching across alpha, with lambda optimally chosen by grid search for each component model.

Missing data is handled according to the H2O default for each model.[4] For the tree-based models (random forests and GBMs) this entails treating missingness as a categorical variable. For the logistic regression this entails mean imputation.

After training the component models, we use H2O's "stacked ensemble" function to create a super learner ensemble of these component models (van der Laan, Polley and Hubbard, 2007). We employ 10-fold cross-validation to identify the optimal weights, and we restrict each model to have non-negative weight. The final model's proportional reduction in loss is 0.195, roughly the same as in the original analysis.

[3]This differs from the original analysis, which used the caret package in R and a slightly different set of component models.

[4]This differs from the original analysis, in which we used multiple imputation to train and average ten separate ensembles.

3.3 DOE Score Calculation

We construct a dataset of all 1,761,031 directed dyad–years in the state system between 1816 and 2012, including the aforementioned material capability variables for each side of each observation. We then use the ensemble trained in the last step to predict the probability of each outcome (A wins, B wins, stalemate) in case the dyad were to have a dispute that year.

As noted above, the undirected DOE scores are calculated by averaging the directed scores, assuming each side has the same chance of being the initiator.

4 Replication

4.1 Instructions

Make sure you have installed the software listed in the previous section, as well as Git and GNU Make. Then, assuming you don't mind having your computer tied up for a few days, run the following series of commands at the terminal:

```
git clone https://github.com/brentonk/doe-h2o.git
cd doe-h2o
make
```

The DOE scores, along with various intermediate files, will be stored in the `results` subdirectory.

Our code sets seeds wherever possible, but we cannot guarantee that training results will be exactly identical across computing environments.

If you have access to a computing cluster, you can speed things up by running the DOE score calculation in parallel. See `predict.slurm` for an example submission script for a cluster that uses the SLURM scheduler.

4.2 List of Files

- `.gitignore`: List of files to keep out of Git version control tracking.

- `0-download.r`: R script to download required data.
- `1-clean-data.r`: R script to prepare data for training and prediction.
- `2-train.r`: R script to train the super learner ensemble.
- `3-assemble.r`: R script to assemble the year-by-year directed-dyad DOE predictions into a single file and calculate the undirected-dyad predictions.
- `4-compare-to-v1.r`: R script to calculate the canonical correlation between the original and current DOE scores (0.942 as of DOE 2.0).
- `5-prl.r`: R script to calculate the proportional reduction in loss of the super learner (0.195 as of DOE 2.0) and an ordered logistic regression solely on the capability ratio (0.018).
- `doe.bib`: Bibliography for references in this file.
- `LICENSE`: MIT license for the code included in the replication repository.
- `Makefile`: Instructions to run the complete analysis via GNU Make.
- `predict.slurm`: Sample submission script for running the post-training predictions in parallel on a computing cluster.
- `README.md`: Markdown source code for this file.

5 Change Log

- **2019-07-22.** DOE 2.0 released.

References

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29:1189–1232.

Gibler, Douglas M., Steven V. Miller and Erin K. Little. 2016. "An Analysis of the Militarized Interstate Dispute (MID) Dataset, 1816–2001." *International Studies Quarterly* 60:719–730.

Singer, J. David, Stuart Bremer and John Stuckey. 1972. Capability Distribution, Uncertainty, and Major Power War, 1820–1965. In *Peace, War, and Numbers*, ed. Bruce Russett. Beverley Hills, CA: Sage.

van der Laan, Mark J., Eric C. Polley and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6(1).

Zou, Hui and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.