National Child Development Study

What It Is

The National Child Development Study is an influential dataset that has been used in over 900 publications in health and social science journals. The datasets associated with the study track a group of 17,000 people form birth through the present day.

Our Objective

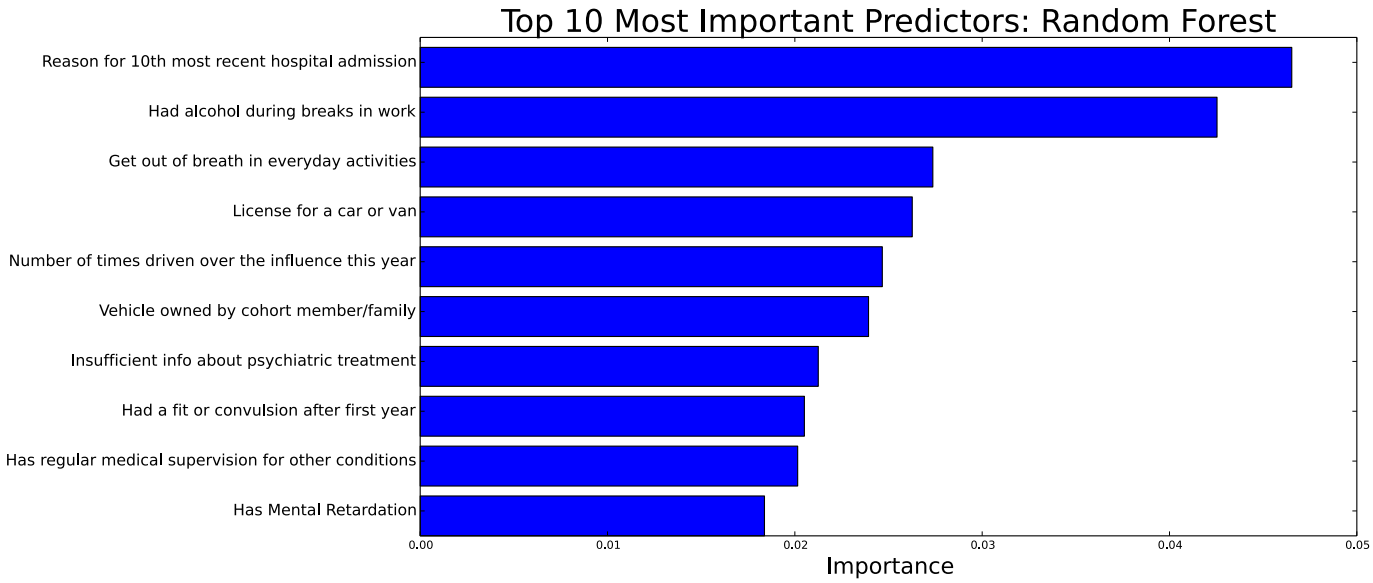
We hoped to use these datasets to learn about the most important features in predicting a patient has having or not having epilepsy.

Data Cleaning

The datasets are messy and contain thousands of predictors, so we did some initial filtering by choosing predictors from categories of particular interest. We also needed to remove predictors that directly related to the patient's epilepsy condition, such as whether or not the patient had visited a physician for epilepsy. We concatenated the datasets - captured at various points over the past fifty years - according to each patient's unique ID and set to work on fitting classification models to the data.

Modeling and Results

We focused on fitting logistic regression and random forest classification models to the NCDS data. After tuning the hyperparameters, we managed to achieve a cross-validated classification accuracy of 73% on both patients with and without accuracy. Given that some of the predictors for each patient included whether or not he or she had experienced fit or convulsions, we had hoped to achieve greater classification accuracy. Regardless, we were able to identify some important factors, displayed here.



Predicting Epilepsy from other Diseases

Question
If epilepsy is correlated with other diseases, can we use this correlation to give a better-than-chance diagnosis of epilepsy in NCDS patients?

Method
We proceeded by constructing a dataset with fifteen binary disease predictors. In the adjacent figure, we show the correlation matrix corresponding to the sixteen disease variables.

We fit and tuned a Random Forest classification model to the data. The average ROC AUC score was 0.526, corresponding to a classifier that is only marginally better than a coin flip. We would not recommend using this classifier as the sole determinant of a patient's diagnosis.

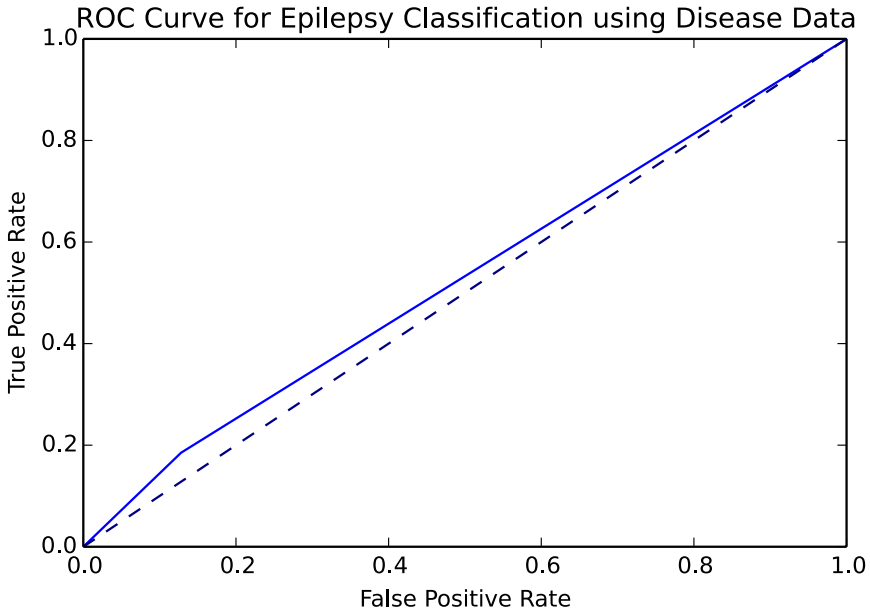


Figure: On average, a Random Forest classification model fit to the disease predictors barely outperforms a random coin flip in diagnosing epilepsy.