Technical Methodology

Jiaxi Wu

1. Data management and overall workflow

For the flight delay, we estimate the performance of Regression Tree and Random Forest, Classification Tree and Random Forest. We follow the following workflow:

(1). Data preprocessing: remove the records of cancelled flights. Delete carriers with less than 100000 records.Then only retain columns about time of day, carrier and airport ID information. We also treat negative values on departure time as 0 and add a categorical variable named "delay" to denote whether or not the flight was delayed. Finally restore the whole data set for usage of models.

(2). Divide the data into a training sample and a hold-out sample.

(3). Run each prediction algorithm on training sample. For classification algorithms, we treat "delay" as dependent variable, while for regression, we treat "DepDelay" as dependent variable. We also tune empirically on the training sample by cross-validation.

(4). After storing all the prediction functions, we run each algorithm on the hold-out sample and produce the statistics we are interested in such as precision, recall and R^2.

2. Training & Out-of-Sample Datasets

It's necessary to separate a hold-out(testing) sample to assess the model performance since good in-sample performance may overstate the overall performance and overfit the model. In this assignment, we separate 20% of the original data as a hold-out sample to test the model performance.

3. Model Parameters

For classification algorithms, we treat categorical variable delay as dependent variable, while for regression algorithms, we treat numerical delay time as dependent variable. Then we run the algorithm on the full training sample and store the result.

We choose complexity parameters like tree depth and minimum units in each non-terminal node by cross-validation within the training sample. We randomly divide the training data into 5 folds. For each fold, we fit the model for every tuning parameter value on all other folds and predict on the given fold. Then we obtain one prediction for every tuning parameter on each fold and average the prediction loss over the full training sample for each tuning parameter. Based on the loss, we choose the tuning parameter and then fit the model with the chosen tuning parameter on the full training sample. Finally evaluate the fitted function on the hold-out sample to test the performance of model.

When considering multiple tuning parameters, such as the depth of the tree and the minimal size of a non-terminal node in this assignment, we choose these parameters jointly; that is, we run the above procedure over a grid of multi-dimensional parameter. As a result, we relax the parameters due to the over-simplication of this problem.

4. Fit & Performance

As a result, classification tree has precision of 0.61 and recall of 0.51 on hold-out sample. Regression tree has hold-out R^2 around 0.025. Classification random forest has precision of 0.61 and recall of 0.55 on hold-out sample. And regression random forest has hold-out R^2 around 0.0158. So we may conclude classification random forest has the best performance on this data set.