Executive Summary

Jiaxi Wu

In this assignment, we build tree models to predict flight departure delays in minutes. To achieve this goal, we use data about 2m flights containing the carrier and airport using information. To eliminate features that can leak information about prediction targets or not relevant to the analysis, we only retain time of day, carrier and airport ID information. Treat negative values on departure time as 0. We also add a categorical variable named "delay" to denote whether or not the flight was delayed. When assessing the performance of prediction, it's reasonable to consider R^2 for regression and precision/recall for classification as measures. It's also necessary to separate a hold-out(testing) sample to assess the performance since good in-sample performance may overstate the overall performance and overfit the model. When training model, we use cross validation to choose the complexity parameters such as the depth of a tree.

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.