1/23/2018

Just found out my Random Forest performed worse than Linear Regression on the four year MLB dataset. This could be for two reasons:

1. The features have a smooth, linear dependence on the covariates, leading the linear regression to model the dependence better than random forests, which will approximate a linear curve with an ugly irregular step function.
   1. Perhaps add more trees?
2. Or, it could be that my Random Forest algorithm is overfitting the training data, which should be systematically like the test data, and thus the test predictions won’t be as good.
   1. Well, we do know that there was Random Forest led to a significant overfit on the housing data example. So, we may want to figure out a way to tackle the overfit.