**Pick the best model**

At the very beginning, we tried various models and picked the model with the highest prediction accuracy, which is XGBoost. The steps to train and test each model are as following:

We picked the 2015 Regular Season data and 2016 Regular Season data from www.basketball-reference.com/ as training and testing data respectively.

We trained the model based on the 2015 Regular Season data, with parameters selected from 10-fold cross-validation. For training the model, we treated each game as one observation. And for each game, we used the performance of both teams, say Team A and Team B, and other game related data (such as the players’ information and the weather the game is played in Team A’s home field) as features.

Based on our model, we predicted the outcome of each game on 2016 Regular Season. For each team, we found all the historical games it played and used weighted average of the history data as the measure of performance of one team. For predicting the results for each game, knowing which two teams will play, we combined the historical performance of the two teams and added other information as test features. Once a game finished, we updated the historical performance of both participated teams, and continued the process until the last game.

Following the above procedure, we finally found the XGBoost had the best prediction accuracy, which was 70.2%.

**XGboost**

For predicting the champion for the 2017 playoff, we retrained the XGBoost Model, on the 2015 Regular Season, 2016 Playoff, and 2016 Regular Season data, with the best parameters (depth, eta) selected from cross-validation.

After training the model, we could predict the wining probability for Team A in each game as long as we have the information for that game.

**Simulation Description:**

In our project, instead of only giving the wining probability for one team in each game, we intended to utilize the probabilities to run the simulation and predicted which team would be the winner.

According the NBA rules, each pair of teams will play 7 games, with the 1st,2nd,5thand 7th games played in Team A’s home field, and others in Team B’s home field. One round will end until one of the team has won 4 games and become the winner.

Based on our model, given two teams Team A and Team B, we will have two probabilities based on whether Team A plays in the home filed. According to the schedule and playoff rules, we simulated the result for each game and predicted the winners for each layer, from the first round: 8 teams in both Western and Eastern Conference, to the final layer: only 1 team left in both Conference and fight for the championship.

Importantly, in each layer, we can update our features and wining probabilities by incorporating the most up-to-date data. And as the real games go on, we can correct the winner of each round and continue the simulation process.