Super Team or Super Dream?

Conor Klaers and Jade Gosar

Introduction

For the first time in 3 years, the NBA is looking forward to a full 82 game season with a standard post-season tournament featuring eight teams from both the Eastern and Western Conference. To further heighten the anticipation, there was a massive wave of free-agent signings and trades this past offseason that have seen the odds of teams like the Los Angeles Lakers skyrocket. While some teams looked to fill their roster through trades and draft picks, others have turned to players returning from injury, such as Klay Thompson for the Golden State Warriors, with high hopes that they will return as great as they left. Regardless of where team's focused their efforts in the off-season, the movement of players between teams in the past few months has set the stage for an NBA season unlike any other. It will feature some of the most talented teams to ever compete in the league as well as only three teams who are seen as actively trying to position themselves to be as high in the lottery as possible next season which is in contrast to the many of the previous years.[1] All in all, the motivation for our project stems from our interest in the dramatically different landscape of the NBA this year and the amount of impact that an off-season filled with free agency drama will have on a team's projected success.

Since player movement is such an integral contributor to success in the NBA, we want to explore the impact of free agency on team performance. With the landscape of the league drastically altered since the Milwaukee Bucks were crowned champions last season, we want to predict team wins using player data from updated rosters. By assessing the impact individual players have on a game's outcome, we will be able to predict next season's results beyond mere speculation. Additionally, with the start of the season less than a month away, our models will rely purely on data from the past two seasons. A two year window will enable us to add weight to recent performance while still accounting for players who missed last season with injuries. This will ultimately allow us to pull out individual-level statistics on players that were traded during free-agency to predict how much each will contribute to their new team.

To accomplish this we used data gathered from the nbaStatR package which contains a substantial amount of information on metrics such as blocks, turnovers, field goals attempted, field goals made, steals, and free throws attempted to name a few. We further manipulated the data to add metrics such as successful field goal percentage, successful three-point percentage, and free throw percentage. Ultimately, the outputs of our project will be a bagging model, a tuned xgboost model, as well as an individual player analysis that will allow us to draw conclusions on whether a traded player will have a positive or negative impact on their new team.

Related Work

Predicting how successful any given team will be in the upcoming season is not a new concept. As it currently stands, many analysts are predicting that the Los Angeles Lakers and Brooklyn Nets will be at the top of their respective conferences with the Milwaukee Bucks, Phoenix Suns, and

[1] Bontemps, Tim. "NBA Offseason Survey: Execs, Scouts on the Biggest Deals, Best Players and 2022 Title Favorites." *ESPN*, ESPN Internet Ventures, 19 Aug. 2021, https://www.espn.com/nba/story/_/id/32046072/nba-offseason-survey-execs-scouts-biggest-deals-best-players-2022-title-favorites.

Utah Jazz following closely behind.[2] These predictions are heavily based on how successful each of these teams were this previous season, as well as the talent that they acquired in the offseason. For example, the Lakers added MVP level talent in Russell Westbrook, along with the likes of Carmelo Anthony, Dwight Howard, and Rajon Rondo which, on paper, might be "among the best in modern Association basketball".[3] Even after an early first round exit in the Western Conference Finals, the Lakers are a heavy favorite to win the conference due to the players they added this offseason. Many have deemed these additions a more important factor than their performance last season. On the other hand, the prediction of the Bucks having success is based on the fact that they are the reigning champions rather than significant personnel additions. All in all, while the idea of predicting the likelihood of a team experiencing success is not a new problem, we aim to dive deeper into the analysis on an individual level in order to better understand the impact free agency will have on each team.

Other areas of the NBA that have been explored through machine learning are classifying players' importance into high and low in order to distinguish the players who are or should be marketed the most and using models to predict individual game outcomes which is highly prevalent when it comes to sports betting. In the first case, the models performed well when classifying the most important players in the high importance class at an accuracy level of around 68% for the logistic regression analysis. The next area that has been explored is using models to predict individual game outcomes based on team statistics. In this field, "the best NBA game prediction models only accurately predict the winner about 70% of the time"[4] which shows that the variable nature of basketball games makes the machine learning process slightly less accurate than it could be when applied to other sports. Overall, there has already been some work done when it comes to predicting what teams will succeed the most in the upcoming NBA season as well as recent attempts to classify players by their importance and predict individual game outcomes. However, we plan to distinguish our problem from these by assessing the impact individual players will have on a team's overall success this upcoming season.

Data Description

For our dataset, we used the nbastatR package to pull NBA game logs from the 2019-2020 and 2020-2021 regular seasons. In order to avoid bias, specifically those introduced by the 2020 playoff bubble, we did not use playoff statistics in our models. The only clear weakness of our data structure was that it may undervalue teams with high draft picks. Since our predictions are based on data from previous seasons, teams with high impact rookies are likely to be more successful than our model suggests.

Broken down by player statistics for each game, these two seasons provided 45,447 observations measured over 48 variables, such as outcome of the game, whether it was the second game of back-to-back, and offensive and defensive statistics for each player. We then combined this

---

[2] Fitzgerald, Matt. "NBA Playoff Predictions: Bracket Picks & 2022 Finals Champion." *Sportsnaut*, 8 Sept. 2021, sportsnaut.com/nba-playoff-predictions-east-west/.
[3] Vincent.j.frank. "NBA Power Rankings: Nets, Lakers Top Dogs with Training Camp Opening." *Sportsnaut*, 14 Sept. 2021, sportsnaut.com/nba-power-rankings/.
[4] Weiner, Josh. "Predicting the Outcome of NBA Games with Machine Learning." *Medium*, Towards Data Science, 7 Jan. 2021, https://towardsdatascience.com/predicting-the-outcome-of-nba-games-with-machine-learning-a810bb768f20.

data into team statistics for each game, using a for loop to compile past averages. As a result, we were able to look at average offensive and defensive statistics over a specified period of time for each team prior to any game during the season (See Figure 3 for Summary Statistics). The new data frame consisted of 86 predictors, 43 for each team in the matchup. Each team's first game in the specified range had to be omitted, however, as there was no previous data from which to calculate averages. At this point we created some exploratory visualizations to plot turnovers and assists based on game outcome and points scored by outcome shown in the appendix. Obviously, winning teams will display higher average point totals, Figure 1 shows us that winning teams are concentrated around 120 points while losing teams usually score around 100 points. Figure 2 showed some outliers in the data in which teams with high turnover rates almost always lost the game. After gaining these initial insights, we moved on to creating our training and test data.

Since nbastatR is linked directly to the NBA Stats API, our dataset can be easily updated in real time by adjusting the dates from which data game logs are pulled. When this is done, the most recent game data will automatically be incorporated into the statistics used in our models. In our modeling process, we used 80% of our observations for training and 20% as a test set.

Methods

The first model we ran was a random forest predicting game outcome by all of the quantitative variables in the dataset. We used 200 trees and an mtry of 84 for each of the variables in the game stat data frame. Using this model, we predicted outcomes, in terms of win or loss, for our test data set. Considering the frequency of uneven matchups in the NBA, in which game odds are not 50/50, the level of accuracy we obtained, at only .588, was not very impressive; however, we wanted to have results to compare our XGBoost model against when it comes to what variables have the most predictive power. To accomplish this, we extracted the importance of each variable (Figure 3). While it is possible that we could have tuned this model for a better result, we decided it was not powerful enough to confine our predictions to wins and losses.

To increase the power and applicability of our predictions, we ran an XGBoost model to predict game point differential. With the large dataset we used, employing XGBoost could give misclassified samples more weight, and in turn avoid errors on the test dataset. Since there are so many factors determining the outcome of an NBA game, we thought this would be advantageous. XGBoost is also flexible in terms of how it classifies and evaluates models. Since we switched from binary to continuous response, using XGBoost would allow us to switch back with relative ease.

Results

The Random Forest we ran to make binary game outcome predictions displayed a 0.5907 accuracy when applied to our test data. Wins were set as the positive class, and the confusion matrix showed a sensitivity of 0.5872 and a specificity of .5940, making it fairly balanced but more likely to correctly predict a loss than a win (See Figure 4). We ran an importance matrix to determine the most impactful variables in the Random Forest model, and we found that blocks were actually the most prominent predictor of wins, followed by average points scored of both teams. This was unexpected, but it suggests that the winningest teams over the past two seasons likely had strong interior defenses. It would be interesting to see if blocks remained a key predictor when more seasons are incorporated into the model. Overall, however, considering the uneven nature of many NBA matchups, predicting games at an accuracy of almost 59% was not especially impressive.

After creating the training and test matrices required by the XGBoost package, we then ran an initial model with nrounds set at 100. Since we were no longer running a binary model, root mean squared error was our evaluation metric, and this model returned an RMSE of 14.55. It appears, however, that this model overfit the data, as the training RMSE on round 100 was just 2.05. This disparity between training and test accuracy went away as we tuned our XGBoost hyperparameters.

We tuned the model to find the optimal max depth, minimum child weight, gamma, subsample, colsample_bytree, and eta parameters. Our final model was run with an eta of 0.1, a max depth of 3, a minimum child weight of 10, a gamma of 0.05, a subsample of 1, and a colsample_bytree of 0.6 (See Figures 5-8). With a new RMSE of 14.46, the accuracy of the new model was slightly improved. Although this did not explicitly predict win or loss, the outcome of any game can be determined by whether the point differential was positive or negative. In terms of practical application to fields such as sports betting, determining the margin of victory is more insightful than binary response. The tuned XGBoost model found points scored, as well as offensive and defensive field goal percentages to be especially predictive of point differential (See Figure 11).

We used our XGBoost model to test the impact of player upgrades on the likelihood of winning. As an initial exploration, we substituted Russell Westbrook, one of the biggest acquisitions of the NBA offseason into the roster of his new team, the Los Angeles Lakers. Based on expected performance above replacement, we predict Westbrook to increase the winning percentage by 11-12%. Effectively, this is an improvement from 7th seed to 2nd seed in the Western Conference. Following a first round playoff exit, this is significant.

Discussion

The main insights gained from our Random Forest and XGBoost model came from our feature importance analysis, in which we were able to pull out the variables in our dataset that impact the outcome of NBA games the most. In both models the most important predictors were blocks, average points scored for both teams, and a range of team defensive stats. Specifically, the Random Forest model showed defensive rebounds, points allowed, and field goal percentage against as powerful predictors for the binary outcome of win or loss. The fact that both of our models have similar variables with strong predicting power was promising in light of our accuracy and RMSE scores. Additionally, there were no variables in our XGBoost model that stood out from the rest in regards to the level of their importance. As shown in Figure 11, the feature with the most importance was points scored with other various point scoring measures following closely behind. Team rebounds and blocks also are important predictors with a value close to .2. All in all, the main insights we gained from our Random Forest and XGBoost model were what metrics in our dataset have the most predictive power when it comes to both game outcome and point difference. Although our models were not particularly accurate, the insights gained give us a better understanding of which aspects of NBA games are the most important for predicting a team's success.

After creating our models to determine team success, we turned our attention to evaluating individual players in order to predict the impact that they will have on their new team. With this being the ultimate goal of our project, we decided to focus on one player in particular: Russell Westbrook. Russell Westbrook is possibly the biggest name traded to the Lakers this off-season and has a statistics line that led us to believe he will be a major asset to the Lakers this upcoming season. After replacing an average player, such as Kentavious Caldwell-Pope, with Westbrook and comparing how the Lakers would perform in each situation we came to the conclusion that Russell Westbrook will likely have a

very large positive impact on the team's success this upcoming season. To put it in perspective, we predicted that if the Lakers had Westbrook last year they would have won 57 games as opposed to 41 and would have a win percentage of 81%. Although this is a very high win percentage for any team to achieve, it is not far off to assume that the Super Team that is the Lakers this year could get close. Further action to be taken off this insight would be to do a similar analysis on other players and compare to the predictions we have made for Westbrook's impact to gather a more comprehensive picture of how powerful the Lakers could be.

Conclusion and Future Work

In conclusion, our report focuses on predicting team wins and point differential of games with the ultimate goal of being able to assess the likely impact a given traded player will have on their respective team's success. In our analysis, we found that the most important predictors for our random forest model were team blocks, average points scored for both teams, and team defensive statistics such as field goal percentage against and points allowed. Our random forest model was focused on predicting team wins and losses and had an accuracy of .5872 which we decided was not powerful enough considering the difficulty that predicting wins in the NBA entails. When it comes to our XGBoost model, we focused on predicting point differential instead of a binary variable such as wins and losses due to the fact that it has more practical applications. In this case, our model had an RMSE of 14.46 after tuning the parameters and showed that average points scored, opponent team blocks, and team defensive statistics are the most important predictors when looking at point differential. Although both our models did not perform as well as we originally thought they would, we further researched what the typical accuracy score is of other models that have been created to predict NBA team wins and the upper bound ranges between 66-72% accuracy. We feel that with more time and resources we would be able to tune our model further to fully incorporate updated player rosters which would raise the accuracy of our model significantly going forward.

The key finding of the model that we created to predict an individual player's impact on their team was that it seemed to be more accurate than the XGBoost model, although it is important to note that we were focused on point differential in our XGBoost model and win percentage in our player evaluation. We discovered this by running a player analysis on Russell Westbrook and predicting Laker wins if he had replaced KCP, an average player, last season. This model performed very well, predicting that the Lakers would win 41 games with KCP when they actually won 42. Furthermore, our model predicted win percentage in 2021 to be .5575 when it was actually .5915 which gave us the confidence to assume that running the model with Westbrook replacing KCP would show accurate results. Overall, the key takeaway from this analysis is that Russell Westbrook likely will have the largest positive impact of players traded to the Lakers and will significantly raise their chances of making a deep run this upcoming season. Given more time, we would be interested in applying the player model to all traded players in order to predict the ones that will have the most impact on their teams and attempt to factor in that particular ones may need their stats lowered to match new playing time expectations.

Contributions

Conor focused heavily on working with the data to get it into a usable form and building the models needed for our team-level analysis while Jade concentrated on making the model to determine how much impact an individual player will have on their respective team. We worked together to finish

both the short report and final report and Jade worked on creating the design for the presentation while Conor imputed all necessary information.

Bibliography

Bontemps, Tim. “NBA Offseason Survey: Execs, Scouts on the Biggest Deals, Best Players and 2022 Title Favorites.” *ESPN*, ESPN Internet Ventures, 19 Aug. 2021, https://www.espn.com/nba/story/_/id/32046072/nba-offseason-survey-execs-scouts-biggest-deals-best-players-2022-title-favorites.

Fitzgerald, Matt. “NBA Playoff Predictions: Bracket Picks & 2022 Finals Champion.” *Sportsnaut*, 8 Sept. 2021, sportsnaut.com/nba-playoff-predictions-east-west/.

Vincent.j.frank. “NBA Power Rankings: Nets, Lakers Top Dogs with Training Camp Opening.” *Sportsnaut*, 14 Sept. 2021, sportsnaut.com/nba-power-rankings/.

Weiner, Josh. “Predicting the Outcome of NBA Games with Machine Learning.” *Medium*, Towards Data Science, 7 Jan. 2021, https://towardsdatascience.com/predicting-the-outcome-of-nba-games-with-machine-learning-a810bb768f20.

https://www.cbssports.com/nba/news/nba-win-totals-for-2021-22-season-nets-bucks-projected-for-most-victories-in-league-lakers-jazz-lead-west/

https://www.nba.com/news/nba-player-movement-2021

https://fadeawayworld.net/nba-trade-rumors/nba-rumors-30-players-that-could-be-traded-until-the-2021-deadline

Appendix

Figure 1

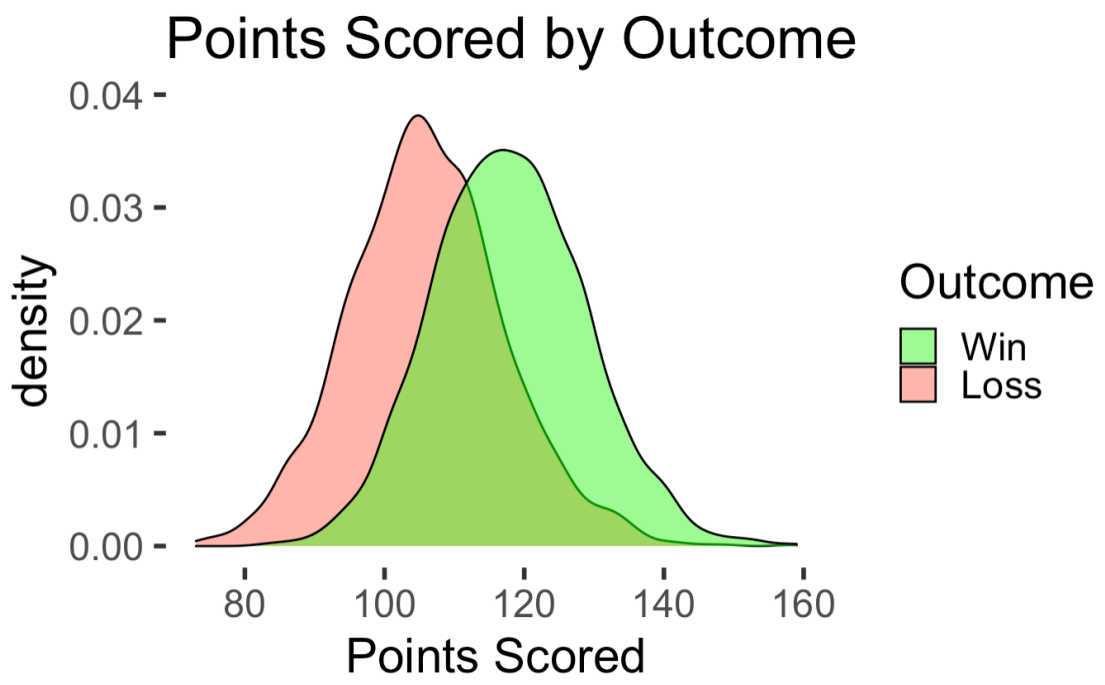

Figure 2

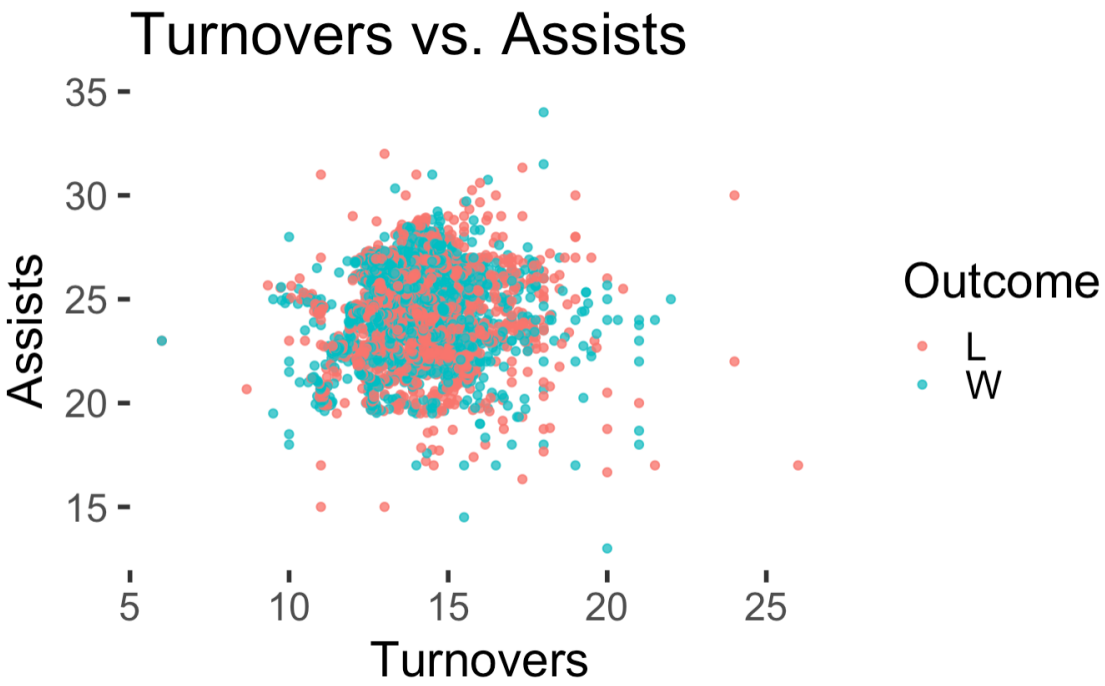

Figure 3

```
   blk_2_def    blk_1_def  pctfg_1_def pctfg2_2_def    pts_1_off    pts_1_def  pctfg_2_def    pts_2_off
    52.48568     37.64646     37.24836     35.32821     33.00245     32.45445     32.04612     29.96062
 pctfg_1_off   dreb_1_off    ast_1_off  pctft_1_off  pctft_1_def pctfg3_1_off  pctft_2_off   fg2m_1_def
    26.69240     24.74582     23.35104     22.78089     22.32310     22.11854     21.97120     21.89931
pctfg2_1_off   oreb_2_def   treb_1_off    ast_2_def    blk_1_off   oreb_1_off   dreb_2_off    tov_1_off
    21.21803     21.08573     20.88756     20.82953     20.81697     20.67583     20.56928     20.53162
pctfg2_2_off pctfg3_2_off pctfg2_1_def    pts_2_def  pctft_2_def     pf_1_def    blk_2_off    ast_2_off
    20.48073     20.29806     20.22104     20.05406     20.04627     19.92391     19.91394     19.89125
  fg2m_2_def    ast_1_def pctfg3_1_def    tov_2_off    stl_1_def  pctfg_2_off    fgm_1_off pctfg3_2_def
    19.67309     19.62830     19.56325     19.53791     19.49046     19.43920     19.31445     19.30529
   tov_2_def   fg3a_1_def   treb_2_off   oreb_2_off    stl_2_off   oreb_1_def    ftm_2_off     pf_2_def
    19.13511     19.11455     19.06422     18.92465     18.77266     18.69290     18.65553     18.44950
  dreb_1_def    stl_1_off    tov_1_def   dreb_2_def    fgm_2_off   fg3m_1_def    ftm_1_def    fgm_1_def
    18.39917     18.05233     17.76868     17.68891     17.57619     17.56836     17.35332     17.31219
   fga_1_off     pf_2_off    ftm_1_off    stl_2_def   fg3a_2_def    fga_2_def     pf_1_off   treb_1_def
    16.89285     16.83390     16.80079     16.71615     16.59804     16.55676     16.24061     16.12322
  fg2a_2_def    ftm_2_def   fg2a_2_off    fga_2_off   treb_2_def   fg2m_2_off    fgm_2_def   fg3m_2_def
    16.07747     16.05886     15.86150     15.80884     15.69260     15.68540     15.62422     15.35833
  fg3a_1_off    fta_2_off   fg3m_2_off    fta_2_def    fta_1_off   fg2m_1_off   fg3a_2_off    fta_1_def
    15.17905     15.00646     14.93306     14.79916     14.76203     14.66470     14.55324     14.35412
  fg2a_1_def    fga_1_def   fg3m_1_off   fg2a_1_off
    14.29138     14.14218     14.04825     13.33379
```

Figure 4

```
bag_preds   L   W
        L 259 168
        W 177 239

               Accuracy : 0.5907
                 95% CI : (0.5567, 0.6242)
    No Information Rate : 0.5172
    P-Value [Acc > NIR] : 1.055e-05

                  Kappa : 0.1811

 Mcnemar's Test P-Value : 0.6667

            Sensitivity : 0.5872
            Specificity : 0.5940
         Pos Pred Value : 0.5745
         Neg Pred Value : 0.6066
             Prevalence : 0.4828
         Detection Rate : 0.2835
   Detection Prevalence : 0.4935
      Balanced Accuracy : 0.5906

       'Positive' Class : W
```

Figure 5

```
bst_preds_1 <- predict(bst_1, dtest) # Create predictions for xgboost model

bst_1_rmse <- rmse(actual = test_data$point_dif, predicted = bst_preds_1)

bst_1_rmse
```

```
[1] 15.09869
```

Figure 6

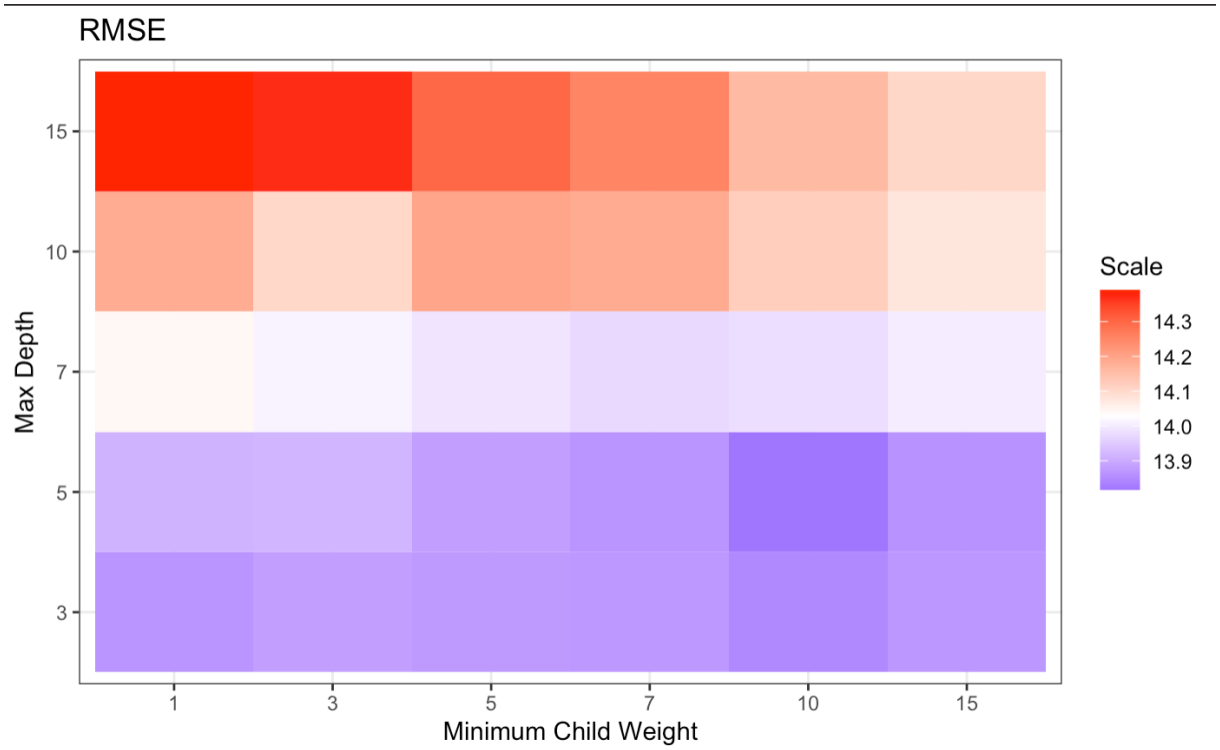

Figure 7

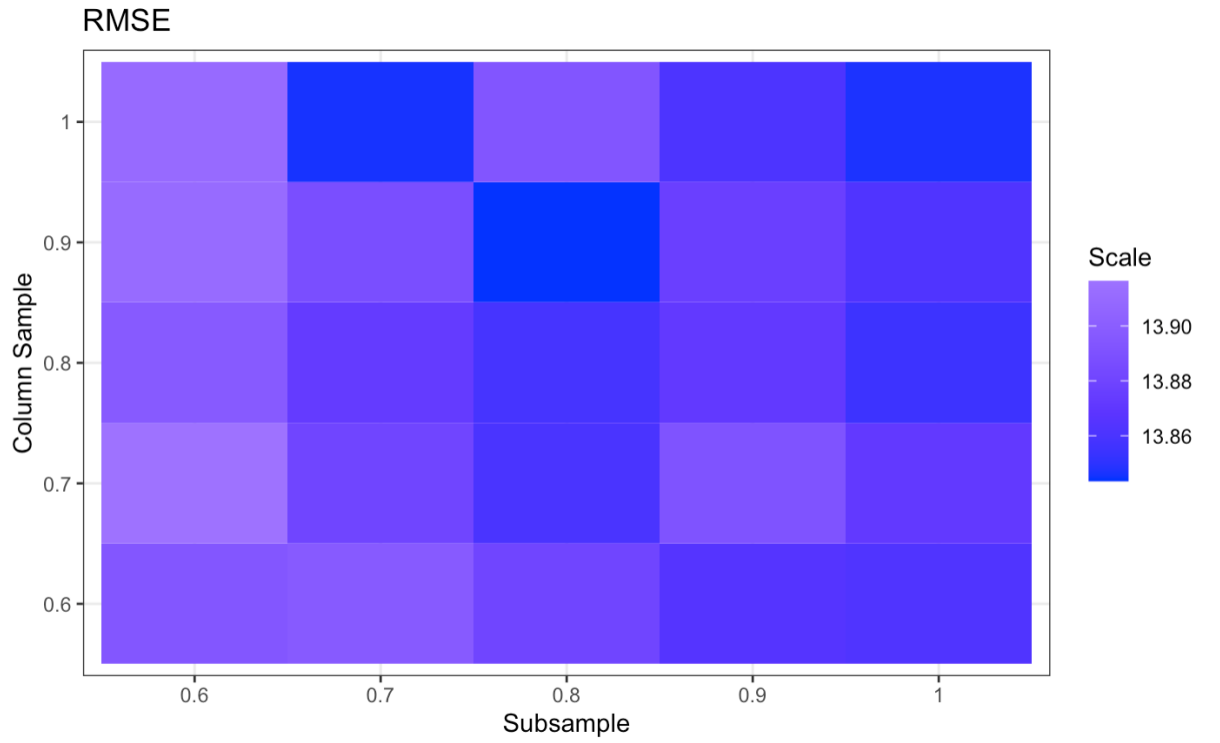

Figure 8

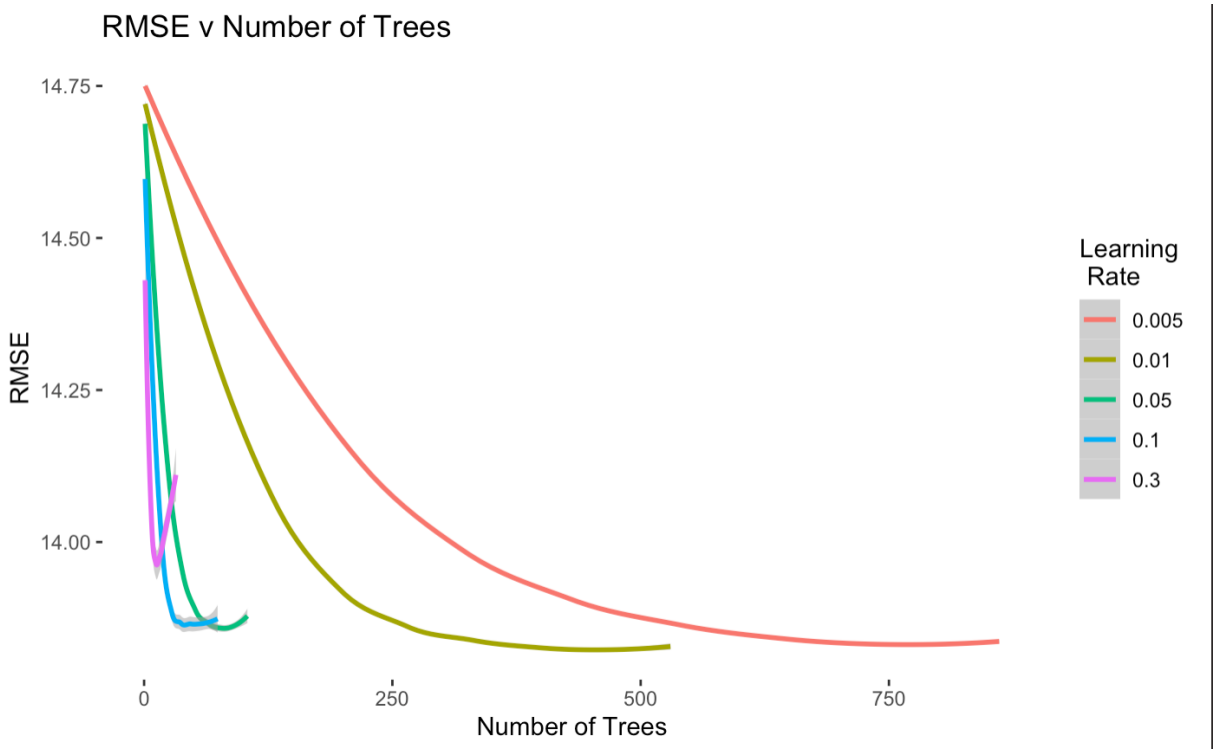

Figure 9

Figure 10

```
bst_final_rmse <- rmse(actual = test_data$point_dif, predicted = bst_preds_final)

bst_final_rmse
```

```
[1] 14.56953
```

Figure 11