**Predicting Player Impact on NBA Games Final Report**

Maria Elosua, Erin Jackson, Jade Gosar, Jose Ernesto Salas, Katerina Rodriguez, Sarah McCabe

**Motivation**

We want to understand what factors differentiate between winning and losing in the NBA. Since player movement is such an important part of the NBA, we want to explore the real impact that a given player has on the game's ultimate outcome. To do this, we will investigate how minutes played, days of rest, and whether they are on a back-to-back schedule affects specific teams and players. By doing so, we will assess individual players’ contributions to NBA teams’ wins and losses.

We are motivated to research this topic because we want to understand the motivations behind the many free-agent signings in past seasons; we also hope to use our insights to evaluate individual player’s contributions to a win or lose to capitalize on the sports betting industry’s recent, exponential growth. The NBA league has grown its revenue more than $585 million, thanks to collective and individual partnerships with sports betting companies. Moreover, the increase in sports betting is projected to increase the consumption of NBA’s offerings, thus increasing teams’ profits.

**Review of Previous Work**

In 2021, UPenn students Josh Weiner, Jackson Joffe, and Jack Rosener used machine learning to predict the outcome of NBA games. The group’s random forest model had a testing accuracy of 67.15%; the model predicted game outcome, based on team average stats over the ten most recent games and Elo ratings. Similarly, Bryant University’s Matthew Houde’s Gaussian Naïve Bayes model achieved 65.1% accuracy with thirteen numeric NBA stats as predictors. Among these stats are team win percentage, field goal percentage, rebounds, assists, and turnovers. As Houde notes, “a model that can perform and successfully predict the outcome of a game greater than 50% of the time is an achievement.” Finally, in 2016, researchers at Xiangtan University predicted the outcome of NBA matches with 74.4% accuracy, using the Maximum Entropy principle and *k*-means clustering to create a model. The model is built upon 29 features of each team, including field goals made, offensive rebounds, and personal fouls.

Each of the previous works acknowledges that accurate prediction is difficult, due to unforeseeable factors like upsets and injury. We assume that we will encounter similar limitations in our own predictions.

**Problem Framing**

Fans love predicting winners—this is the most basic prediction in watching sports. Win prediction is usually done based on the number and strength of past wins. However, we intend to dive deeper into these metrics to really understand which human performance variables are most central to winning. We use metrics from NBA player performance in the 2018-2019 season, the last fully completed season before Covid-19. After we select the important variables, we will more closely examine three players who are central to their team’s success to see their real impact over the point differential of a game.

We intend to analyze how scheduling impacts performance. The NBA season consists of about 82 games that are played in 5 and a half months. Coaches and players have pointed out that playing too many consecutive games negatively affects players’ physical and mental health and scoring. Additionally, a rested and well-recovered team has a strong performance advantage over an over-scheduled, tired team. Because of this, we are interested in analyzing how rest days and back-to-back days affect the player’s performance. Play time also impacts player performance. Even the best players suffer from fatigue and cannot constantly be at peak performance so we aim to consider these factors in addition to traditional metrics, like win percentage and average fouls.

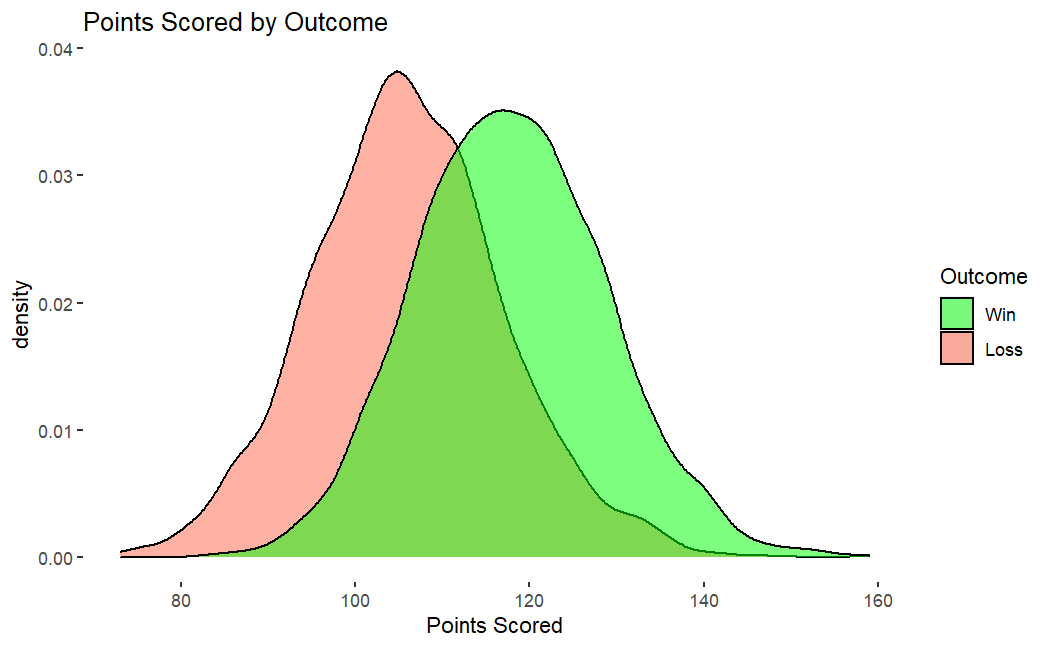
Our project’s outputs are a bagging model to predict wins and losses, a tuned xgboost model to predict point differential, and an individual player analysis that allows us to attribute win probability proportions to individual players.

**Data Overview**

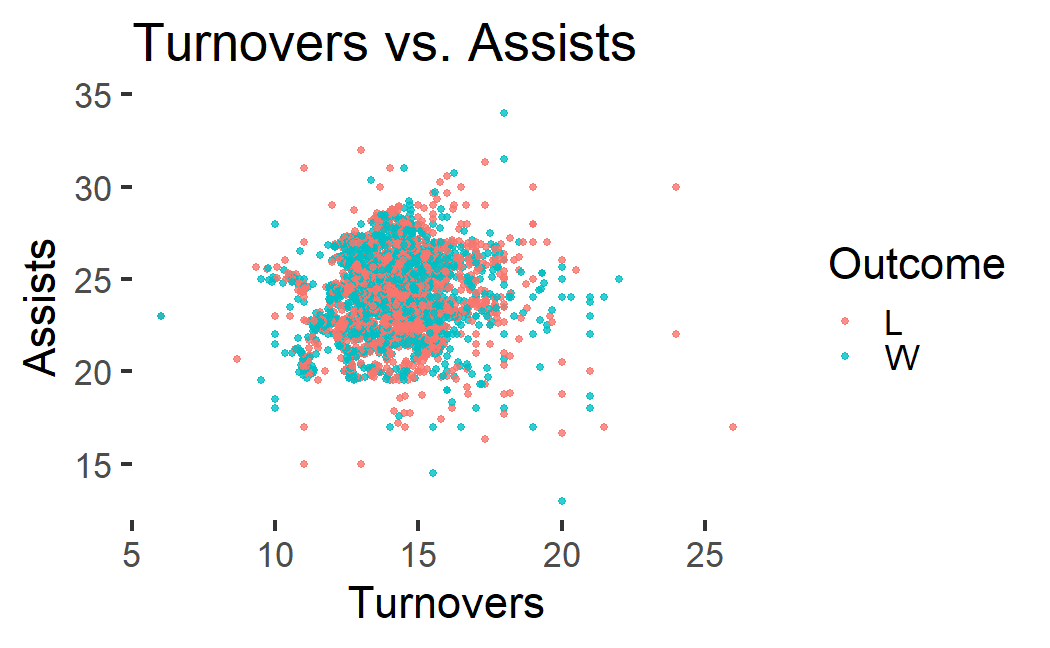
**Dataset Description & Pre-Processing**

The dataset we are using in order to solve this problem is based on the nbastatR package which is linked directly to the NBA Stats API. This means the dataset can be updated easily depending on our needs and focus. Through this package, we pull game logs from the NBA regular season/s of our choice both on a team level as well as a player level. The data extracted from the nbastatR package is organized by player statistics per game. This player data is combined into team statistics per game, using a for loop to compile past averages. With this, we can now look at statistics at the team level such as average offensive and defensive stats. We created two main “rough version” datasets: game\_data and player\_df. Game\_data has 93 columns (variables) and over 4,000 entries (samples). Player\_df has 45 columns (variables) and over 45,000 entries (samples).

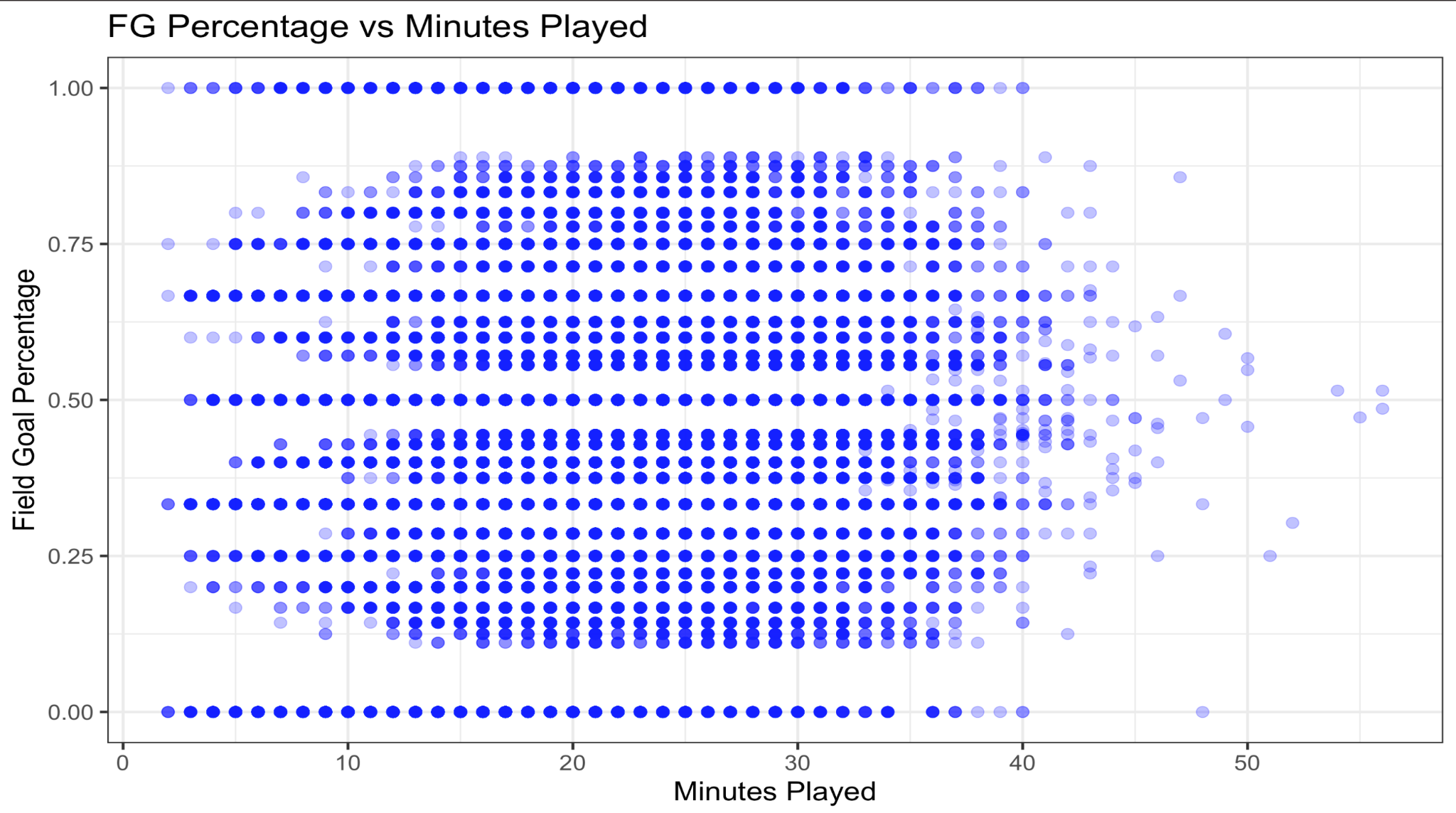
**Exploratory Visualizations**

****

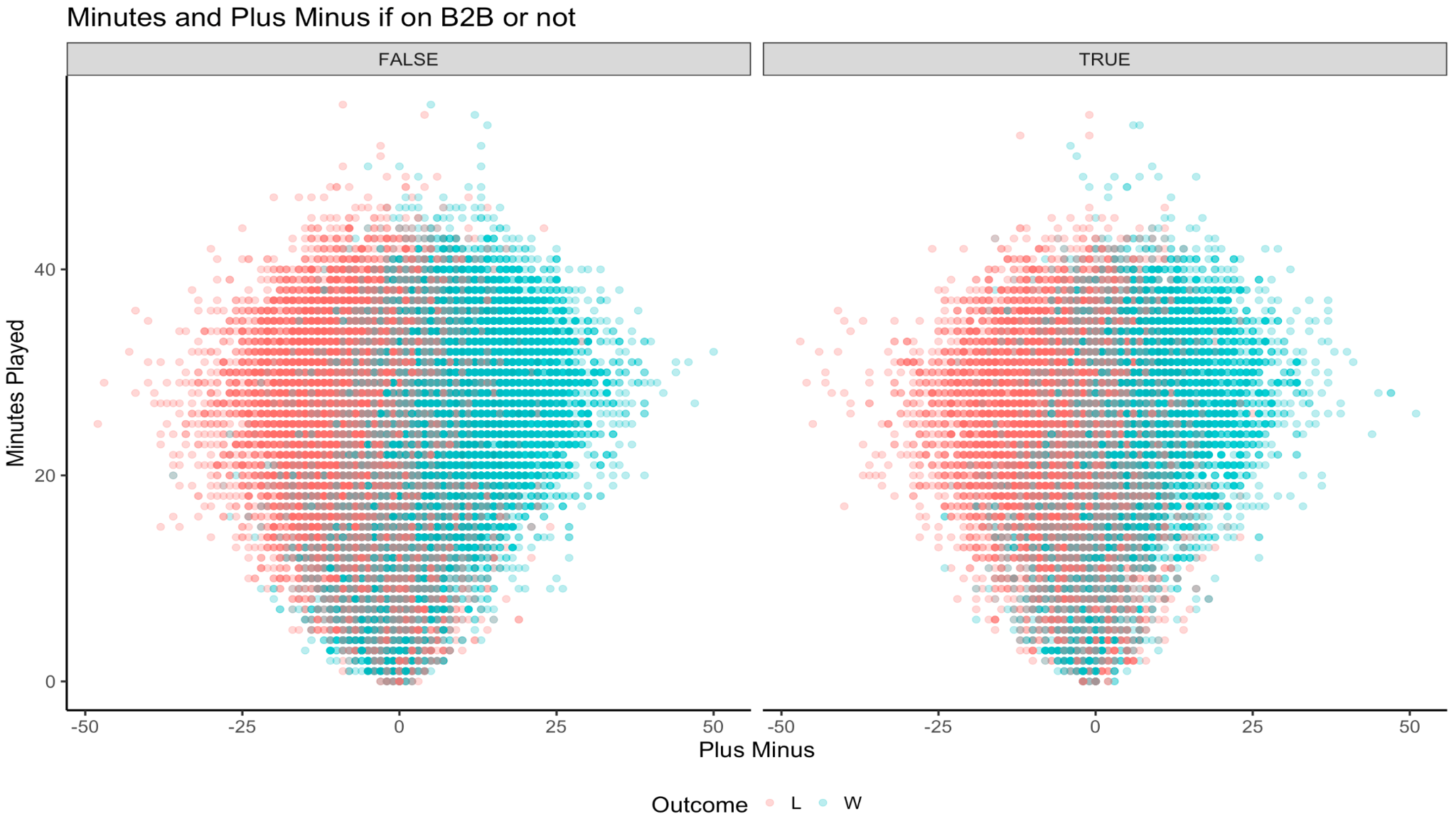
Points Scored (as is intuitively expected) is higher for teams that won the game. Teams that win the game are usually around 120 points. Those who lose around 105 points.

****

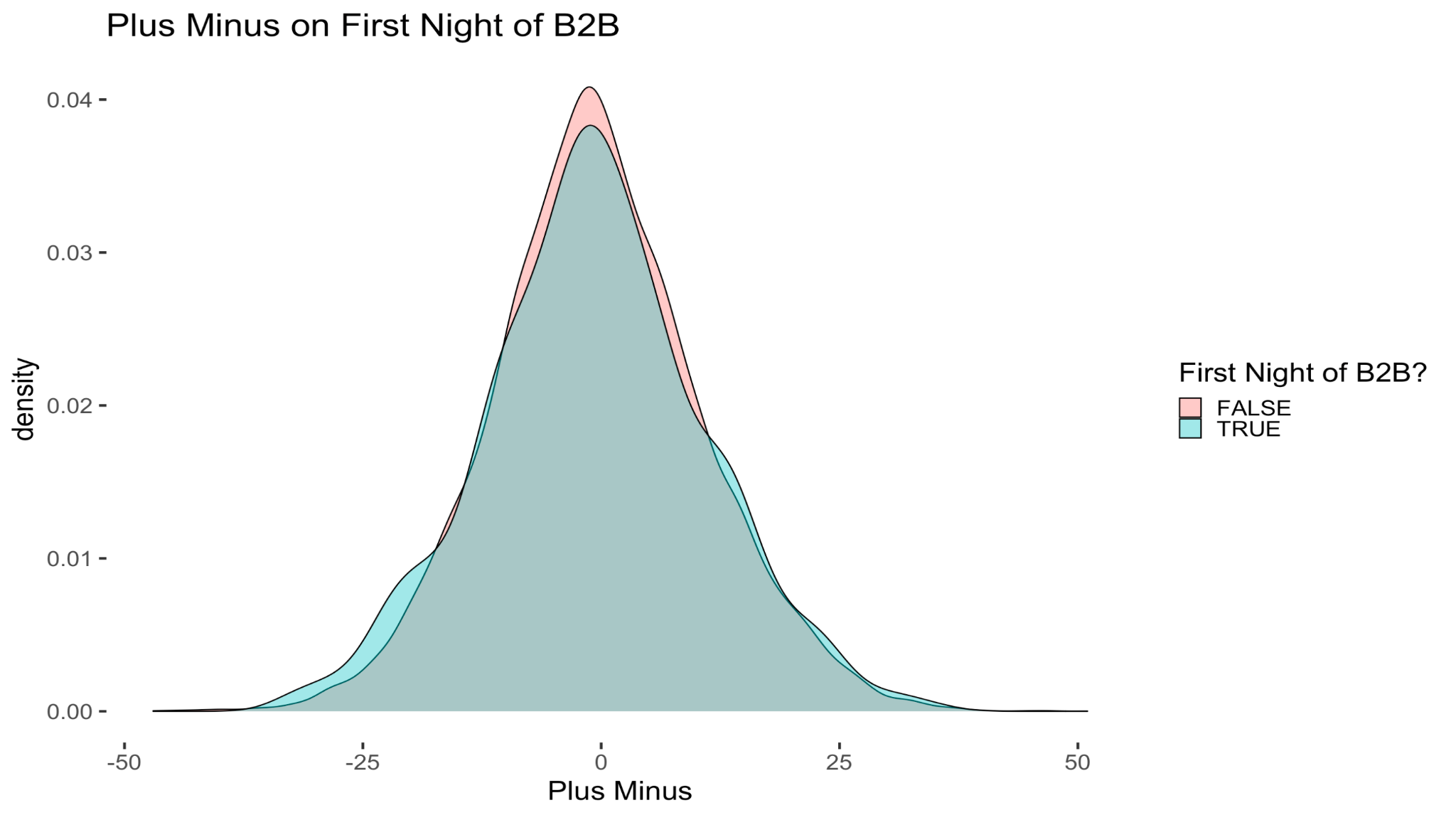
We cannot really derive any significant information from this visualization. Yet, we can see a couple outliers with regards to turnovers in particular, where high turnovers translate to losing the game or better said is correlated with losing the game.



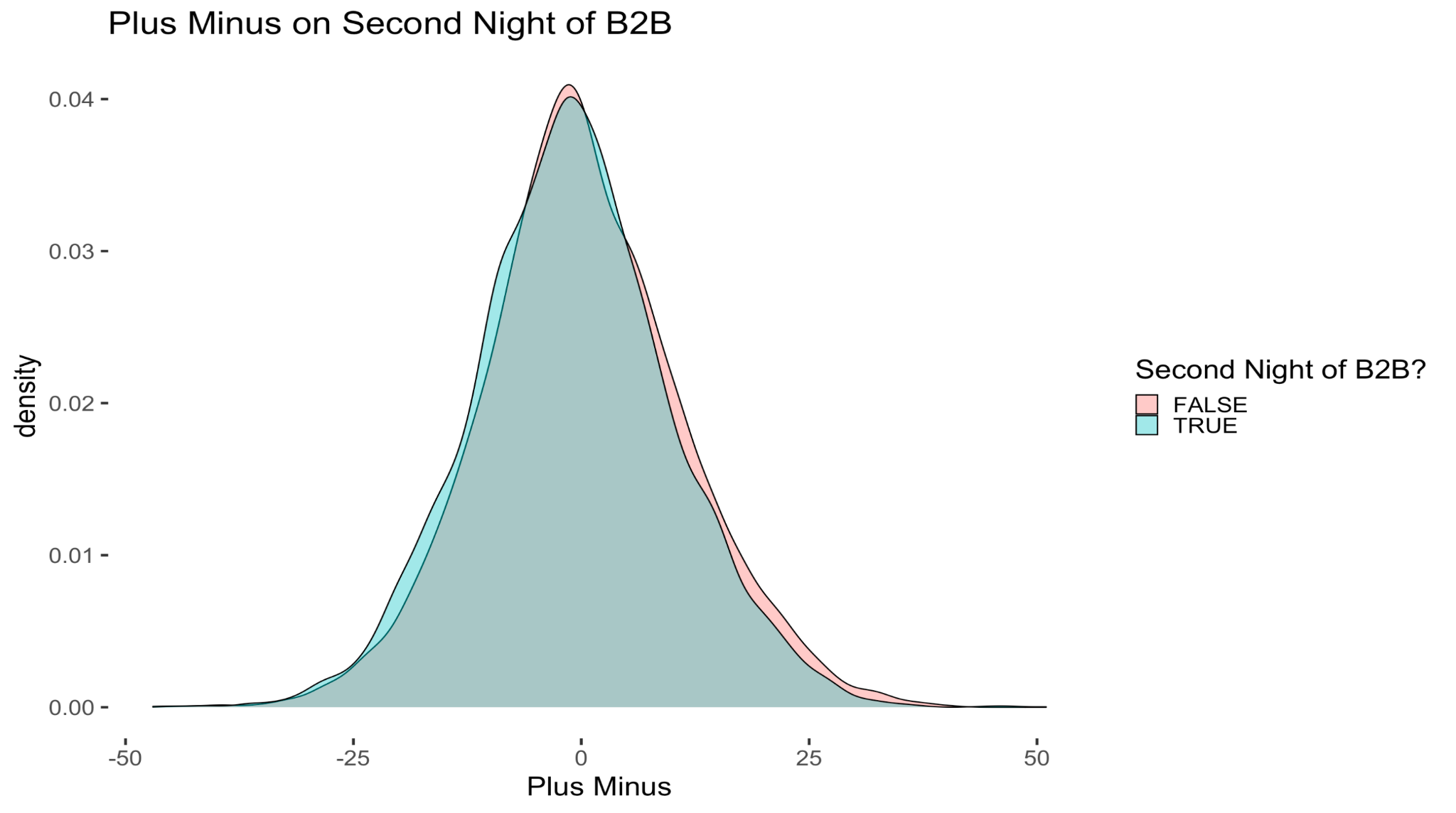
The data appears to be consistent. Arguably, FG attempts are concentrated within 10 to 35 minutes played as this is the “prime” time of the players. Percentage is all over the place. Some outliers are visible.



A plus/minus difference compares how many points a team scored, compared to how many points they allowed, while a player is on the floor. Unsurprisingly, winning teams typically have more players with a positive plus/minus difference than those who lose, particularly as minutes pass. These players contribute to outscoring the rival. We see no discernible difference between well-rested players and players in a back-to-back game.



A player’s involvement in a B2B game does not affect their performance, as measured by the plus/minus differential.



Although you would expect the number of B2B games played to impact performance, there seems to be little, if any, difference in performance. As the number of B2B games increases, performance is slightly negatively impacted.

**Methods**

We initially ran a random forest model to predict game outcome, using all of the quantitative variables in the dataset. For this random forest model, we used 200 trees and an mtry of 84 for each variable in the game stat dataframe. This model looked at previous wins and losses to predict outcomes, but its accuracy level was only 0.575. We compared our random forest results to an XGBBoost model to determine which variables have the most predictive power. To achieve this, we extracted each variable’s importance, as seen in Appendix Figure 1. Ultimately, we decided that the random forest model was not powerful enough to accurately predict wins and losses.

We ran an XGBoost model to predict game point differential and to increase the power and applicability of our predictions. XGBoost gives misclassified samples more weight and avoids error on the test dataset, while maintaining flexibility in classification and evaluation. Thus, we find XGBoost ideal for a dataset as large as ours.

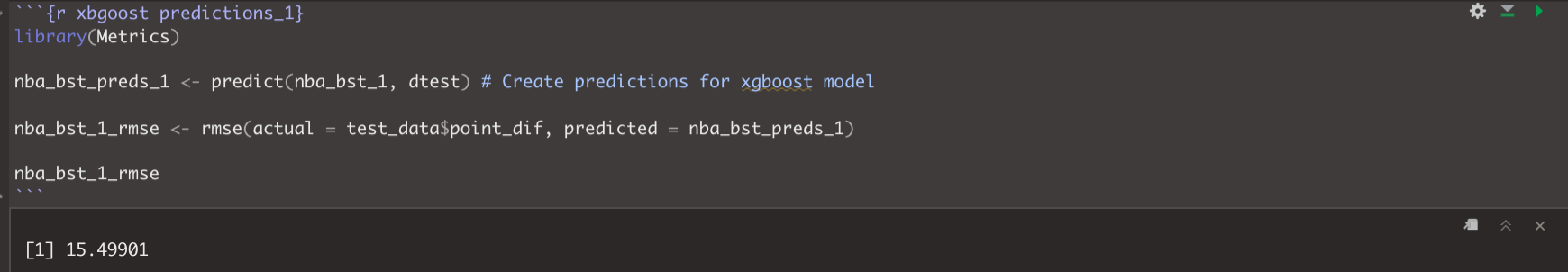
**Results**

The Random Forest and XGBoost models revealed several different important features within the dataset. Running our model on the test data yielded an unimpressive 0.575 accuracy.

With “wins” as the positive class, a confusion matrix showed a sensitivity of 0.5832 and a specificity of 0.5667. These numbers mean that though our model is fairly balanced, the model is more likely to predict a loss over a win. In addition, we ran an importance matrix to determine the most influential variables in the Random Forest model. The importance matrix revealed that the most prominent variables were rebounds (both offensive and defensive), field goal (average of two and three pointer) percentage, and steals as is shown in Figure 1.

Subsequently, we created an XGBoost model, using root mean squared error (RMSE) to evaluate its success. We set nrounds to 100 and the model yielded a 15.499 RMSE, which we further tuned in order to increase accuracy.

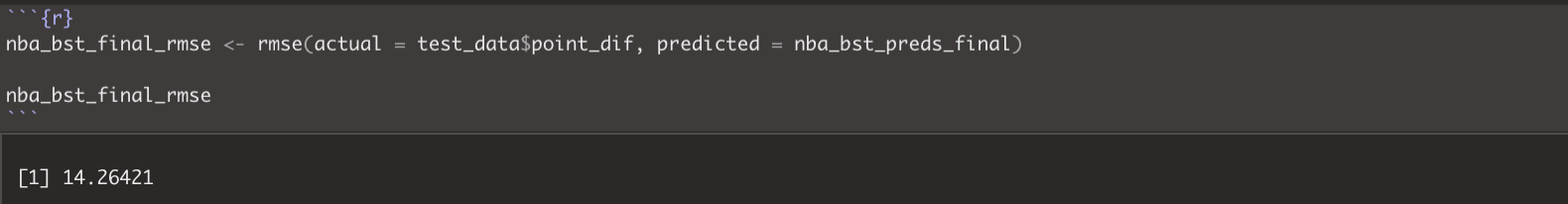
**Original XGBoost model RMSE without tuning**



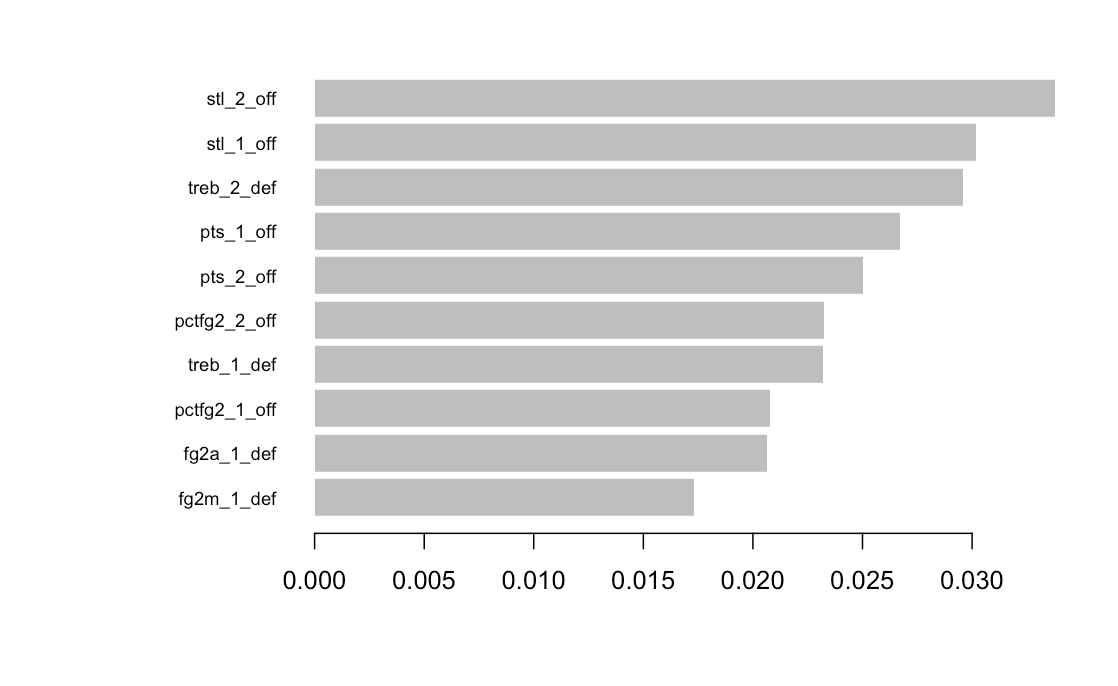
In order to calibrate the model’s parameters, we tuned the model and set the maximum depth to 5, minimum child weight to 3, gamma to 0.1, subsample 0.7, and colsample\_bytree to 1. With these parameters, the new RMSE was 14.26421, so tuning improved the model’s accuracy. Though these results did not explicitly predict win or loss, the results can determine a game’s outcome by determining if the point differential between home and away teams is positive or negative.

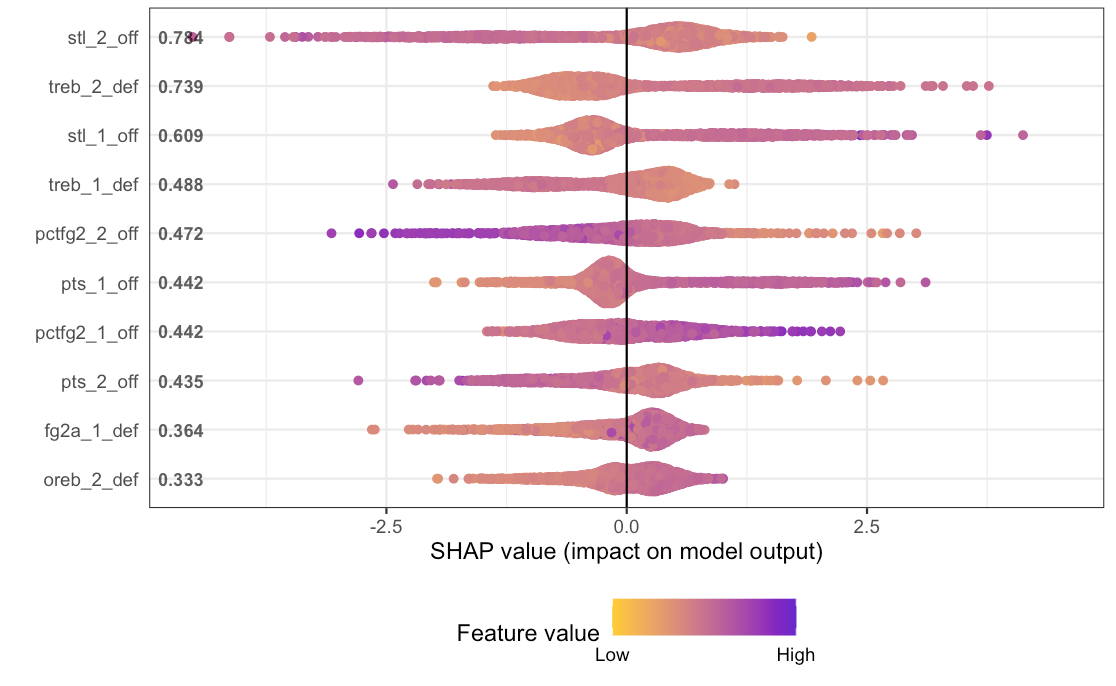
This model may be most applicable to point spread betting. The tuned XGBoost model displayed that offensive stats like field goal percentage, points scored, steals, and rebounds are the most useful to predict the point differential. Ultimately, both our models found similar results for variable importance which is a promising indicator of its applicability to real game data and the sports betting industry in general.

**Tuned XGBoost model RMSE**



**XGBoost importance matrix**

****

**XGBoost shap plot**

**Action**

We used the Random Forest and XGBoost models to collect the most valuable variables that impacted the overall game and then used that to evaluate individual performance. Both models prioritize steals, rebounds, and two point field goal percentage (though these values were emphasized in a different order). Though the feature importance graph reveals that steals are the most powerful metric, no variables in our XGBoost model were vastly more important than the others. Though both models have mediocre accuracy and high RMSE scores, they include the same variables, which implies that these metrics *are* important. Using these variables, we can now determine which aspects of NBA games predict a team’s success, and therefore see how individuals contribute to these metrics. We decided to focus on three players in particular: Jayson Tatum, Nikola Jokic, and Chris Paul.

Jayson Tatum is a three-time NBA all star and an Olympic gold medalist, as well as a forward for the Boston Celtics. As a 24 year old, he is a rising star and has the potential to be an important player for years to come. The data used in this project occurred during his sophomore and third seasons in the NBA. While he is a stronger player now, our model asserts that he had a large impact on his team during the 2018-2019 season. For example, on November 5, 2018, against the Denver Nuggets, the Celtics lost by 8 points, with a final score of 107-115. The model predicted that the final score would be 105-115; without Tatum’s contributions, the model predicted a 96-115 point differential. Though the Celtics still lost, Tatum largely impacted his team’s score. This metric could be useful for coaches to decide when to play Tatum to make their record appear stronger, or when to rest him when the stakes are less high.

Nikola Jokic, also known as the Joker, has played for the Denver Nuggets since his rookie season in 2015-2016. The 27 year old is already in the top 10 all-time players with the most triple doubles. He also is a four-time NBA all star and was voted the NBA’s most valuable player in 2021. However, his value did not show on February 2, 2019 against the Timberwolves, in a back to back game after playing the Rockets. The Nuggets won by 10 points; our model predicts a Nuggets win by 3, though that differential increases to 8 without the Joker playing. The talented Jokic struggled in this back to back game, and the model apparently detected some pattern to reach this same conclusion. Moving forward, we would investigate how long Jokic takes to recover.

Legendary player Chris is considered one of the greatest point guards of all time, with a number of awards and accolades too long to list. During the 2017-2018 and 2018-2019 seasons, Paul started for the Houston Rockets. For one game against the Utah Jazz, the model predicts a loss of 12 points without Paul, but a 3 point loss with him. The model was one basket away from the actual outcome, with a 5 point loss prediction. This pattern follows in other games: the Rockets will lose by less and win by more with Paul in the game.

However, since Paul is the oldest NBA player in the 2021 season, he needs more rest than younger players. Therefore, while he is a key player, coaches should use this model to prevent Paul’s fatigue and burnout.

**Conclusion and Future work**

Our team sought to create a model that could predict NBA teams’ wins and losses, using extensive game data. Offensive and defensive rebounds, two point field goal percentage, and steals are the most influential predictors in our Random Forest model. Players who consistently offensively rebound likely have high two point field goal percentages, so our model did detect existent patterns. However, overall accuracy is low at 0.575.

Rather than create a model to outright predict wins and losses, we created an XGBoost model to predict point differentials per game. The “stl\_2\_off” and “treb\_2\_def” variables are XGBoost’s most important predictors for predicting point differential. Stl\_2\_off records the visiting team’s average steals over the past 10 games; treb\_2\_def tracks the average rebounds that the visiting team allows over the past 10 games. Unsurprisingly, XGBoost prefers moving averages to individual game values, since aggregate metrics show a team’s consistency (or lack thereof).

We then used the XGBoost model to predict the outcome of individual games and attribute outcomes to individual player contributions. To start, we evaluated contributions from Jayson Tatum, Nikola Jokic, and Chris Paul. We accomplished this by running the model twice—once with the original player’s metrics, and once with average player statistics—and comparing the difference in point differential predictions to determine player contribution. The model, which consistently values individual players more in lost games, correctly predicted whether a player’s team would win or lose for every game we analyzed.

In the future, we believe that our model could be applied to individual players across the to analyze their actual impact on an individual game level. Additionally, we would want to explore if this model could be applied to evaluare if players are a necessity to keep on the roster or, due to their overall influence being lower than expected, could be traded for more impactful players. Lastly, given more time and resources, we would like to look deeper into metrics such as days rest, minutes played, or if on a back-to-back game to see if performance changes significantly for players if they do not have proper rest and recovery time between games.

**Contribution**

**JES -** Data Overview: I looked at the data composition as well as the sourcing of the data. Attempted to understand the data and be comfortable with how the datasets were created. A few visualizations were included with their respective interpretations and I created the slide deck for our final presentation.

**Maria Elosua**: I worked on the Motivation and Methods part of this project. I brainstormed our main motivation and researched sports betting and how that is affecting the NBA.

**Katerina Rodriguez:** I did some research on how the amount of rest days and back-to-back games can affect a player’s and team’s performance. I then worked on the results section.

**Jade Gosar:** I worked on creating the final r-notebook.

**Sarah McCabe:** I worked on the problem framing and action aspect of the project.

**Erin Jackson:** I looked into the previous work done on this topic, wrote the conclusion and edited the paper

**Bibliography**

<https://towardsdatascience.com/predicting-the-outcome-of-nba-games-with-machine-learning-a810bb768f20>

<https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1000&context=honors_data_science>

<https://www.researchgate.net/publication/312236952_Predicting_the_Outcome_of_NBA_Playoffs_Based_on_the_Maximum_Entropy_Principle>

<https://barcainnovationhub.com/what-are-the-effects-of-more-or-less-rest-days-between-two-games-in-elite-basketball/>

**Appendix**

Figure 1

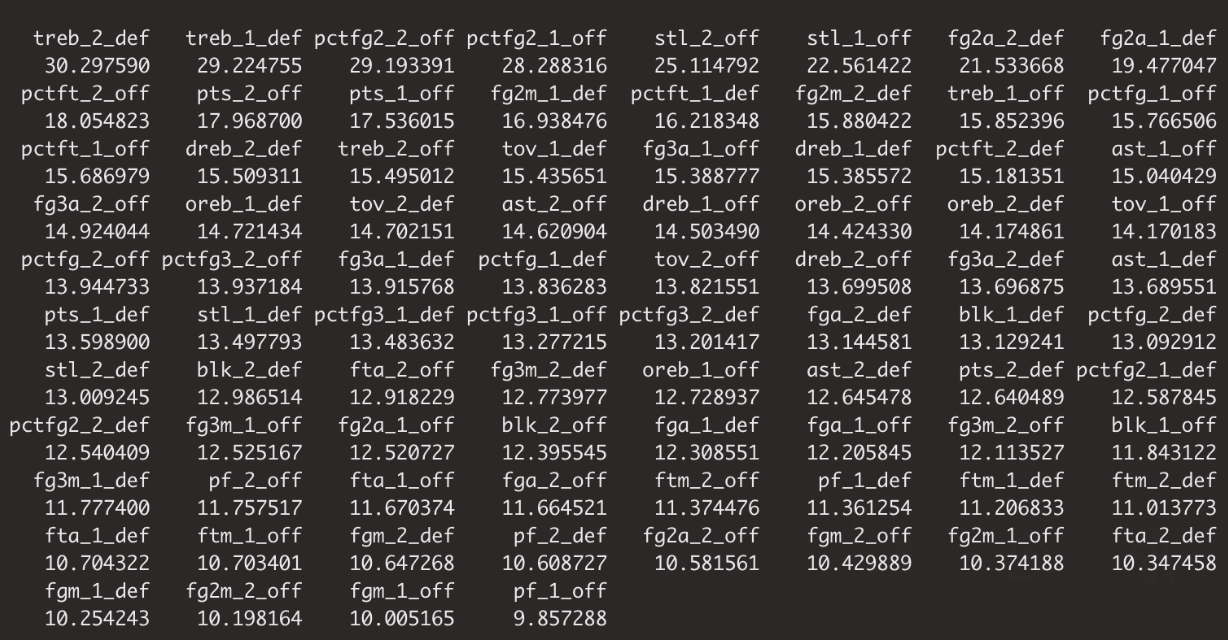


Figure 2

