BBL (Basketball LePredictions)

Abrah Furbee, Adam Dameron, Danny Zepeda-Martinez, Henoc Codije, Madison Duffy, Vivian Moore

1. Data Information

In order to have an initial dataset to add our outside variables to, we had to first clean our given data. After downloading Nathan Lauga’s data from Dr. Mario's website, we decided to not use the *Teams* dataset. All of the variables in *Teams* could be found in the other four datasets (*Games, Game\_details, players, ranking*). We decided to drop the variables LEAGUE\_ID, SEASON\_ID, RETURNTOPLAY and CONFERENCE from the *Rankings* data, as well as renaming STANDINGSDATE to DATE. We deleted these variables because we know we will not need them for further analyses. In the *Games* dataset, we removed the variables GAME\_STATUS\_TEXT, TEAM\_ID\_away and TEAM\_ID\_home because they were not relevant to the dataset. We also renamed HOME\_TEAM\_ID to TEAM\_ID and GAME\_DATE\_EST to DATE in order to more easily work with the variables when we built our models. In the *Game\_details* set, we removed the variables COMMENT and TEAM\_ABBREVIATION, as well as removing any rows that had missing data. The TEAM\_ABBREVIATION was redundant with our other team-level identifying variables, and the COMMENT variable contained notes regarding the game itself which we felt was not relevant to the metrics we aimed to predict. We then narrowed our data selection down to the current 2022-2023 season to ensure we were using the most relevant player and game performance and statistics to best inform our predictions.

When joining together the four datasets, we found it best to create two datasets; one based on the player-level and one based on the game-level. To do this, we inner-merged the *Details* and *Players* sets by PLAYER\_ID and PLAYER\_NAME, removed the TEAM\_ID.y, and SEASON variables, and set the rows to all be distinct. We also inner-joined the *Details* and *Games* datasets by DATE and TEAM\_ID. We then added all our Home and Away TPG, TSG, TOG, PPG, SPG, and OPG variables to the data set, and created the Matchup\_ID variable.

Lastly, we added Left and Right versions of most of the variables in our dataset. Based on the Matchup\_ID, which will be further described in the next paragraph, the “Left” team variables in each game represent the team with the lower valued team ID while the “Right” team variables represent the team with the higher valued team ID. These variables were created and added to the final dataset in order to help eliminate potential variation due to home field advantage. The variables that relied on Home and Away team distinction, such as Spread, Predicted Spread, etc. were not converted to the Left/Right representation. This addition finalized the dataset we went on to use in building our models.

*Adam mentions the game where both teams have over 100 points and how we left it in bc it made the model better*

In order to increase the accuracy of our predictions we created multiple new variables that we added to our final dataset: Matchup\_ID, and all TPG, TSG, TOG, PPG, SPG, and OPG variables. Matchup\_ID is a four digit value made up of the last two values of the team identification numbers, with the last two digits of the team with the lower value team ID first (the team that becomes the “Left” team in the Left/Right variables) and the last two digits of the team with the higher value team ID after (the team that becomes the “Right” team in the Left/Right variables). This variable was created to aid in the joining of our data sets to ensure that the final data set could be created accurately and efficiently.

The TPG variables (Left\_TPG, Right\_TPG, Home\_TPG, Away\_TPG) are a sum of the total points scored by the team during past games, up to but not including the current game being described. The TSG variables (Left\_TSG, Right\_TSG, Home\_TSG, Away\_TSG) and the TOG variables (Left\_TOG, Right\_TOG, Home\_TOG, Away\_TOG) similarly describe the sum of the total steals and total offensive rebounds that the team has accomplished in all past games up until the game being described, respectively. These variables allow us to use the performance of the team up to the point of the game for which we are predicting point spread, total points, and offensive rebounds. The past statistics for the season are important in determining future performance, as the teams will likely perform similarly, on average, in a specific game as they have over past games.

The PPG variables (Left\_PPG, Right\_PPG, Home\_PPG, Away\_PPG) describe the averages of the total points scored by the team in the past up until the game being described (see Figure 1). The SPG variables (Left\_SPG, Right\_SPG, Home\_SPG, Away\_SPG) and OPG variables (Left\_OPG, Right\_OPG, Home\_OPG, Away\_OPG) describe the average of the total steals and average of the total offensive rebounds accomplished by the team in past games up to the game being described, respectively (see Figure 1). These variables were made to better measure average performance of the teams over time. By averaging the TPG, TSG, and TOG variables any outliers in team performance over the past games played up to the current point are smoothed over, which provides us with a metric that better reflects typical team performance and will increase the accuracy of our predictions.



Figure 1: *Formula used for all PPG, SPG, and OPG variables.*

To supplement the data we used to create our models, we added a predicted spread variable gathered from a [Project FiveThirtyEight dataset](https://projects.fivethirtyeight.com/2023-nba-predictions/games/?ex_cid=rrpromo). This variable was joined into our final data set by matching the team name to the respective TEAM\_ID given in our original datasets. This variable describes the predicted point spread by a model created by professionals at FiveThirtyEight. We added this variable in order to better inform our own predictions and compare the values formulated by our models to their predicted spread for the same team match-ups. Additionally, we used the R package [nbastatR](https://www.rdocumentation.org/packages/nbastatR/versions/0.1.10131) to gather the Date, Home\_ID, Away\_ID, Home\_pts, Away\_pts, Home\_oreb, Away\_oreb, Home\_FGPct, Away\_FGPct, Home\_3Pct, Away\_3Pct, Home\_Steals, and Away\_Steals variables we used in our final data set. These variables feature information regarding performance statistics on a team level, and we included it because it is updated more frequently and will therefore contain values more consistent with current team performances. As a result, our predictions will likely be more accurate as the models were constructed based on this more recent collection of statistics.

Part 1:

* Steps taken to get data cleaned, joined, sourced
* How did we handle missing data?
* How did we handle outliers/identify them?
* Did we remove any games? Why?
* How did we get team names and OREB into the dataset?
* **Chronological Order!**
* Rubric: Clearly outlined and described
* Why we chose not to use player data
* Why we

Part 2:

* Discuss engineered variables
* Why we engineered those variables - how they helped with prediction of 3 outcome variables
  + averages
* Mathematically represent metrics (formulas); Written descriptions
* Rubric: Innovative, well defended

Part 3:

* Discuss all outside data joined into set
  + Project 538 data (predicted matchup spread), NBA stat R
* Where we got it, why we included it
* Rubric: Innovative, well defended

1. Methodology for ***Spread***

Going into this we had a couple separate model plans for spread, doing two linear regressions, a K-Nearest Neighbor (Or KNN) Regression, and polynomial regression. These would all be tested using either MAE (Mean Absolute Error) and RMSE (Root mean squared error) and with the best one being chosen as the final model to run our future predictions on. For the first one, we ran all our initial variables through to see if a linear regression was even worth doing, as if it came up with a high MAE or RMSE that means that even with additional variables that we wouldn’t have it would not be a good measure of spread. To do this we performed a forward selection. It ended up giving an overall MAE of 1.612544. This meant that it was a good indicator of spread and we moved on to recreating it without the use of additional variables.

The models after the initial one used the following variables: Home\_ID, Away\_ID, Spread\_Pred, Home\_TPG, Away\_TPG, Home\_TSG, Away\_TSG, Home\_TOG, Away\_TOG, Home\_PPG, Away\_PPG, Home\_SPG, Away\_SPG, Home\_OPG, Away\_OPG, Matchup\_ID, Left\_ID, Right\_ID, Left\_TPG, Right\_TPG, Left\_TSG, Right\_TSG, Left\_TOG, Right\_TOG, Left\_PPG, Right\_PPG, Left\_SPG, Right\_SPG, Left\_OPG, Right\_TSG, Left\_TOG, Right\_TOG, Left\_PPG, Right\_PPG, Left\_SPG, Right\_SPG, Left\_OPG, and Right\_OPG. Along with this, we split the data into two groups: one that used the Right and Left and all their interactions and the Home and Away and all their interactions. The interactions were done between each of the variables in groups of two, with one for the Home and Away pairs and one for the Left and Right pairs. Ultimately, we found that the interactions had no real impact on the final model chosen.

The variables were chosen because they were either ones that we could have on hand or simply calculate for the teams that were doing to play. These were placed into a table with the spread variable and then ran through a stepwise selection linear model. This caused the best output to have an MAE of 5.489101. However, instead of predicting using the variables that we expected, it focused on using the betting data and prediction data from sports blogging sites. To test if this was the best way to go, we reran it with two more model types.

After the linear model, we followed it up with a KNN model to see if we would get good results from it. It worked better than expected given that spread had a much larger range that is normally used for KNN models, giving us a spread MAE of 10.24442. Being lower than the RMSE of the secondary linear model, it was waived in favor of the second linear model.

The final model that we checked was a polynomial regression using a degree of 3. When we used a stepwise model selection to check this one, we ended up getting a model that also used the betting and prediction data. Because of this we decided to go with the model that used the predictions from the betting and sports blogging data. This had around the same MAE of 5.62863.

1. Methodology for ***Total***

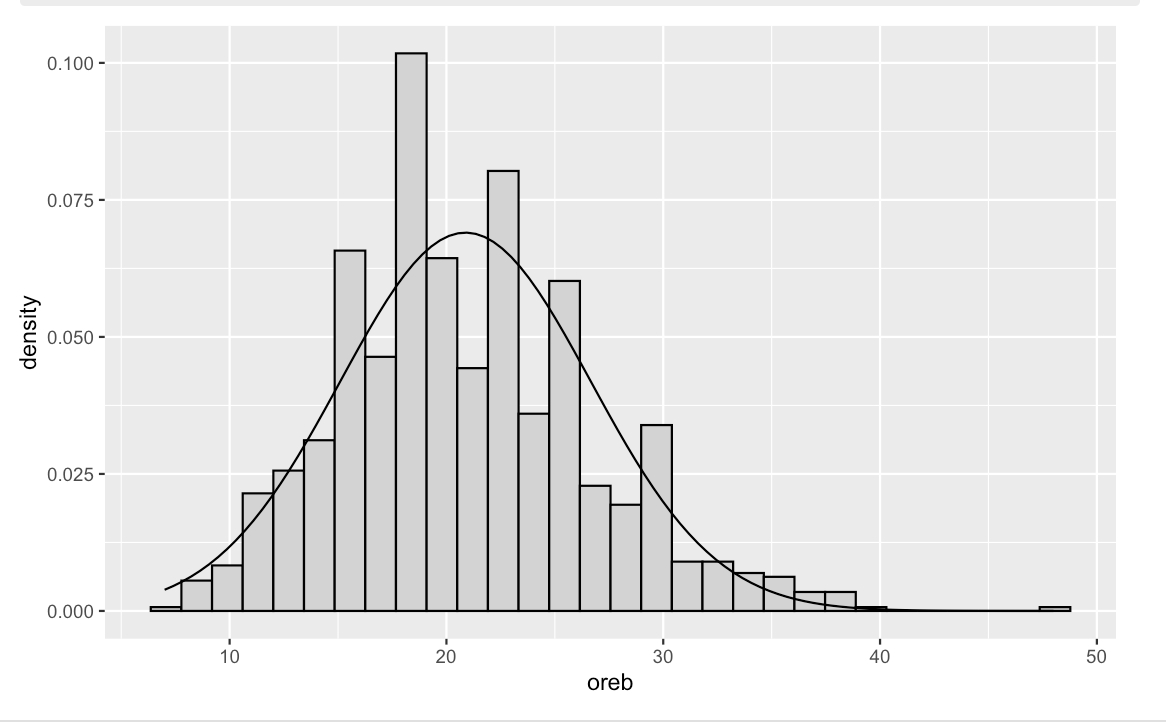
* Describe best model for predicting *Total,* and why we picked it
* Describe steps taken to create the model
* Need out-of-sample testing for consideration in model selection and potential regularization (discuss the results of this)
* **Chronological Order!** (Describing steps taken for picking best model including out-of-sample testing)
* Which variables were useful or useless for predicting *Total* (how used dataset with the model)
* **Optional:** add tables/figures to defend choices
* Rubric: Excellent description of best model, Methodology innovative, clearly described, thorough

Anova showed that IDs are significant in predicting (p = 0.01091)

To start the model creation process we first considered two saturated linear models that had Total Points as a response. One model had home and away IDs, predicted spread, and season statistics as predictors, and the second model had Matchup ID, and Left and Right team season statistics. Both models also had the day in the season as a predictor as well, to account for team improvement over the course of the season.

The first model, with home and away statistics

1. Methodology for ***OREB***



To create an optimal model, the first step that was taken was to examine the distribution of the dependent variable (OREB) to see what kind of models (normal distribution, poisson, etc.) would work best, and if necessary, apply transformations to improve the distribution of the dependent variable. Looking at Figure 1.1, the distribution of the dependent variable seems to more closely resemble a normal distribution than a Poisson distribution (even though the variable is a count variable). Therefore, we will experiment with models that use the Gaussian distribution as their base. Next, a series of transformations were applied to this variable to see if the OREB variable could more closely approximate a normal distribution. These transformations were taking the log of the OREB variable, taking the square root of the OREB variable, and reciprocating the OREB variable. The log transformation didn’t change the distribution much (if at all) and the reciprocal made the distribution more skewed than it initially was. The only transformation that improved the distribution was the square root transformation, which makes sense as a square root transformation corrects for right skewness. By performing a shapiro test, which tests for normality, we found that for a square root transformation, the value was highest at .005, and therefore closer to normal than without a transformation, which was valued at . Based on the results of this test, the dependent variable for all of the models will be transformed into the square root of OREB (and for MAE, the fitted values will be squared for the calculation).

After determining what kind of distribution the dependent variable was and figuring out what type of models may be best at estimating this variable, the next step was to select all possible independent variables that could be incorporated into the model. Given that we are predicting how many offensive rebounds both teams make in a game, it doesn’t make too much sense to incorporate variables that measure this value on a game-by-game basis. This is because we won’t know information like the team's free throw percentage for the game or how many points they score in advance. It would be better to use variables that incorporate this information over time, such as the average points that the team scores in a game or the total points they have scored so far in a season. Another important factor to take into consideration is the concept of home court advantage. Using the variables we have created, we will be able to test for this effect by testing two kinds of different models; one that uses home/away variables (where home court advantage is important) and another that uses left/right variables (assuming there is no difference where the teams play. The variables for the home/away models that will be used are as follows: Home.ID , Away.ID + Day + Home\_TPG + Away\_TPG + Home\_TSG + Away\_TSG + Home\_TOG + Away\_TOG + Home\_PPG + Away\_PPG + Home\_SPG + Away\_SPG + Home\_OPG + Away\_OPG. The variables for the left/right models that will be used are as follows: Matchup\_ID + Day + Left\_TPG + Right\_TPG + Left\_TSG + Right\_TSG + Left\_TOG + Right\_TOG + Left\_PPG + Right\_PPG + Left\_SPG + Right\_SPG + Left\_OPG + Right\_SPG

The methodology of this variable will closely mirror that of the total variable. Fully saturated models are a good starting point for any project dealing with linear regression. Based on this idea, we created 4 different types of fully saturated models: one predicting home OREB using the home/away set of variables, one predicting away OREB using the home/away set of variables, one predicting overall OREB using the home/away set of variables, and one predicting overall OREB using the left/right set of variables. The results of the regressions that predicted home and away OREB separately were added together to get a resulting home + away MAE. The first set of MAE values can be found here:

| **Model Type** | **MAE** |
| --- | --- |
| Home + Away | 3.945736 |
| Home/Away | 3.928675 |
| Left/Right | 3.267374 |

After doing this, we still felt that our group could do more to create better models (given that the fully saturated models had quite a few variables). A drop1 f-test was performed on each of the fully saturated models, and our suspicious proved to be correct. In all of the models, coefficients were found to be statistically insignificant under the drop1 f-test. Based on the results of these F-tests, it became clear that some kind of selection technique could prove beneficial to refine the models. Stepwise regression was then performed on all of the fully saturated models. The second set of MAE values can be found here:

| **Model Type** | **MAE** |
| --- | --- |
| Home + Away | 3.955161 |
| Home/Away | 3.935538 |
| Left/Right | 4.336732 |

The selection technique did not improve the MAE values much, and in some cases made them even worse (although they did improve the MAE values). There was one final modeling technique we used for the OREB variable, and that was making a mixed effects model (where the Day variable became a random effect). The justification for trying this technique is the same as the justification listed in the section for the total variable. (insert mini summary of justification here). This technique was performed on both the full set of independent variables, and a the reduced set selected by the stepwise technique. The final set of MAE values can be found here:

**home/away fully saturated with random effects**

| **Model Type** | **MAE** |
| --- | --- |
| Home + Away | 3.898627 |
| Home/Away | 3.931958 |
| Left/Right | 3.26291 |

**home/away fully “best choice variables” with random effects**

| **Model Type** | **MAE** |
| --- | --- |
| Home + Away | 3.883751 |
| Home/Away | 3.93028 |
| Left/Right | 4.321122 |

After looking at all of the MAE values, the model with the lowest value for OREB was the fully saturated left/right mixed effects model. This is a bit surprising, given that fully saturated models are usually not the optimal choice for a regression model. Additionally, when stepwise regression was conducted on this set of variables, the MAE increased. This is probably due to the fact that Matchup\_ID had many different value options, yet was removed from the “best model” by selection techniques. Nonetheless, these predictions are going to be evaluated by MAE values. Therefore, the best model by MAE results is in fact the fully saturated left/right mixed effects model. The formula for this model is: lmer\_lr1 = lmer(sqrt(oreb)~ Matchup\_ID + Left\_TPG + Right\_TPG + Left\_TSG + Right\_TSG + Left\_TOG + Right\_TOG + Left\_PPG + Right\_PPG + Left\_SPG + Right\_SPG + Left\_OPG + Right\_OPG + (1|(Day)), data = KnownLeftRightStats). The predictions yielded by this model for the upcoming game are:

General Details/Misc. Rubric Comments:

* Organized and discussed in a way an audience with only basic understandings of statistics and the sport will understand
* At least 5 pages long