NBA 2023 PREDICTIONS

Spread, Total, OREB

Team 5 Contributors: Cesar Toro and Austin Cicale

I. Data Information

a) Collecting and Cleaning Data

The goal of this paper is to articulate how we used previous NBA data to create models predicting three variables - Spread, Total, and Offensive Rebound Total (OREB). Using our finalized models, we predicted these three outcome variables for all NBA games between April 4 and April 9 of 2023. We started our search for data with Nathan Lauga's *games* data set, which includes observations for each NBA game from the 2003 season to Dec 22 of the 2022-2023 season. We kept all variables except GAME_STATUS_TEXT and HOME_TEAM_WINS because they didn't have potential for importance in data collection or predicting Total, Spread, or OREB. Then we created variables for Total and Spread, adding PTS_away to PTS_home for Total and subtracting PTS_away from PTS_home for Spread. It was apparent that the OREB variable wasn't included in this data set, which was troubling because observations for OREB is necessary if we are going to create models predicting the variable. However, the *games_details* data set, also created by Nathan Lauga, included values for OREB and other potentially useful variables.

The first problem with simply merging the *games* and *games_details* data sets by common variables such as TEAM_ID and GAME_ID is that the *games* data set has observations for each game on the team level, but the *games_details* data set has observations for each game on the player level. To convert the *games_details* data set into team level observations per game, the sum of player level stats were taken for all players with the same team and game ID. This player level summation process was performed for eight variables in the *games_details* data set that we thought could be potential predictors for Spread, Total, or OREB: FGA, FG3A, FTA, OREB, STL, BLK, TO, and PF. A data set was created to include each TEAM_ID and

GAME_ID combination present in the *games_details* data set, but each of these observation combinations were now assigned one value for the eight selected variables rather than multiple for each player on the roster. This data set represents observations for each game on the team level.

After creating a set of data consisting of GAME_ID, TEAM_ID, and each of the eight variables selected from *games_details*, the next step to obtain one data set with all desired variables was to merge the eight predictor variables into the *games* data set by matching GAME_ID and TEAM_ID. However, the *games_details* data did not have the GAME_ID variable split into home and away teams like the *games* data set, so we had to temporarily rename the HOME_TEAM_ID variable in *games* to TEAM_ID to match the game identification variable in the *games_details* data set. After renaming the variable, we were able to successfully merge the data by TEAM_ID and GAME_ID. The TEAM_ID variable in the *games* dataset was then renamed back to HOME_TEAM_ID. This same process was repeated to gather away team data. Additionally, a similar renaming process was used to add the W_PCT variable from Nathan Lauga's *ranking* dataset to the *games* dataset. We were able to add W_PCT by merging *games* and *ranking* by team id and date of each observation.

The style and strategy of basketball has changed over time. Throughout the twenty-first century, there has been a noticeable shift in style of play that is less dependent on size. Importance of the traditional "Big Man" has diminished and it has been accompanied by the transition to point guards as more prominent scoring threats. One of the most influential players in the adaptation of this style and strategy of play is unquestionably Stephen Curry. His rookie year was the 2009-2010 season, so we decided to account for style and strategy differences by removing all data prior to the 2010 season. Additionally, data was removed for the 2020 and

2021 NBA seasons. The reasoning for removing data from 2020 is because the league went into a “Bubble” due to the pandemic and games were played at a neutral site without fans in attendance. For 2021, some arenas capped fan attendance or only allowed certain individuals in. We didn’t think either of these seasons were truly representative of the factors that play into home team advantage, so these observations were removed. Playoff game data was also removed as those games are played at a much higher intensity where teams go into the series with greater preparation for their specific opponent. Lastly, all observations before February of each season were ignored. Game plans, player minutes, and team composition can all change throughout the year so we focused specifically on data after the trade deadline which typically occurs in early February. This period is also when teams typically focus on making the playoffs if they’re potential contenders. After removing all these games, we were left with 5215 observations. This observation reduction accounted for any missing values if they were to exist.

We analyzed each of our current variables for outliers, and there did not appear to be any obvious, extraneous values. Reassuringly, final models for Spread, Total, and OREB were created using aggregate variables representing team data averages for the season in addition to the last 3 games. The creation of these aggregate variables, which will be further explained in the *Engineered Variables* section, should account for any of the individual outlier observations that were missed.

b) Engineered Variables

Once we compiled all of the data from Nathan Lauga, we calculated two additional variables we felt were important. For each variable we added, it is important to note that we created one for the home team and one for the away team. We created a possessions (POSS) variable, calculated as .96*(FGA + TO + .44*FTA - OREB). The .96 is used to account for the

fact that some possessions end in offensive rebounds, and not in turnovers or missed field goals. This formula is widely used to calculate possessions if the exact number is not known. The portion in parentheses should be familiar as part of Dean Oliver's calculation of turnover percentage. We used the possessions variable in that calculation. Our second variable is win percentage (W_PCT) which is calculated using the HOME_RECORD and ROAD_RECORD variables in the *ranking* data set.

We also transformed some variables to be more useful. Using Dean Oliver's Four Factors of Basketball, we calculated eFG% as (FGM + .5*FG3M) / FGA. FGM and FG3M were not directly available, so we calculated them using the FGA/FG3A and multiplied them by FG_PCT/FG3_PCT respectively. We also calculated Turnover Percentage (TO_PCT) as TO / POSS. We calculated Defensive Rebound Percentage (DREB_PCT) as DREB / (DREB + Opp OREB).

Once we had the variables that we felt could predict the spread, total, and offensive rebounds, we needed to create predictors of those variables for the next game. To do this, we made 8 predictor variables for each of the following variables:

Pre-Aggregated Potential Predictor Variables (each includes _H or _A to indicate home vs away)

PTS = points scored
FT_PCT = free throw %
FTA = free throw attempts
AST = assists
STL = steals
BLK = blocks
POSS = possessions
W_PCT = win %
DREB_PCT = defensive rebound %
PF = personal fouls
TO_PCT = turnover %
eFG_PCT = effective field goal %

The 8 predictor variables were defined as follows:

Aggregated Variables for Each Potential Predictor (VAR is replaced with specified variable)

AVG_VAR_TY_H = home team's average at home this year
AVG_O_VAR_TY_H = average let up by home team at home this year
AVG_VAR_L3_H = home team's average at home in last 3 games

AVG_O_VAR_L3_H = average let up by home team at home in last 3 games
AVG_VAR_TY_A = away team's average away this year
AVG_O_VAR_TY_A = average let up by away team while away this year
AVG_VAR_L3_A = away team's average away in last 3 games
AVG_O_VAR_L3_A = average let up by away team while away in last 3 games

c) Outside Data

For our outside data, we are using teamrankings.com which has all of the current NBA data for each date this season. More specifically, for each stat provided in the original dataset and our engineered variables, it gives a table of teams, their yearly average, last 3 average, last game, home average, away average, and previous year average. After the Sunday, April 2 games, we will use this data to calculate our predictions.

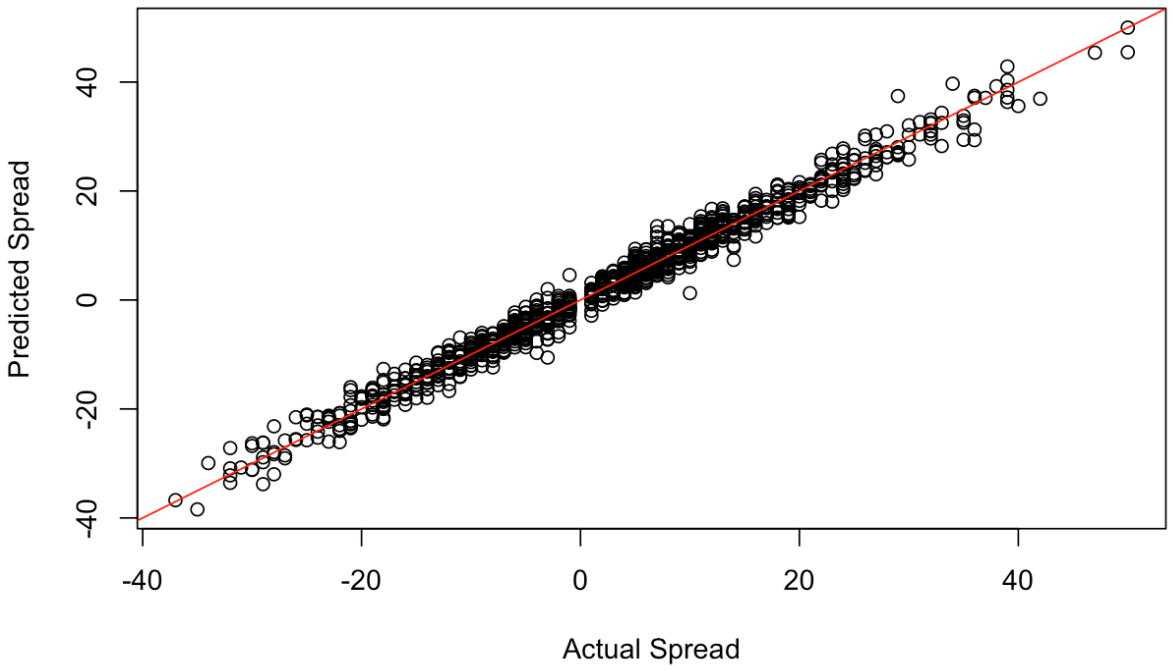
II. Methodology

We began the process of creating models for Spread, Total, and OREB by splitting the data into training and testing sets. The training data includes all games from our cleaned data set that occurred in February or March. The testing data includes all games in April. This split left us with 4067 observations in the training data and 1148 observations in the testing data. We split the data up this way to ensure that our models could properly predict April games after being trained to predict games in February and March, since the games we will be predicting for the project will be those occurring in April. The in-depth methodology steps for predicting each of the three variables will be discussed in their respective section below.

a) Spread

Before creating our model for Spread, we used a bi-directional stepwise progression to identify potential independent variables that would be good predictors of spread using the training data set. This was performed with all of the pre-aggregate variables as potential predictors. This model was cross validated on the testing data and accurately predicted the

testing values. The zero mean, variability, and normality conditions were met for the training models ability to predict the testing data. Additionally, the shrinkage value was extremely close to zero. This training models' ability to predict the testing data is seen in the graph below. We will proceed with using the aggregate variables that represent each of the variables that this model identified as good predictors of Spread.



Bi-directional stepwise progression was now used again, only taking into consideration aggregate values for the variables we isolated as good predictors of Spread in the previous step. After creating this model, we checked for outliers by analyzing studentized and standardized residuals, leverage values, and cook's distance. There did not appear to be any significant outliers influencing the model. This model was then cross validated to see how accurately it could predict the testing data. The model appeared to predict fairly well, with an rmse of 12.91263.

Our predictions will be graded based on MAE, so we want our finalized models for each of the three variables to minimize rmse. In an attempt to create a model with a lower rmse, we

recreated a model using the aggregate values for the variables we isolated as good predictors with LASSO cross validation. This is a better model selection process because it standardizes all variables and oftentimes results in models with smaller rmses. This model resulted in an rmse of 12.8276, which is slightly smaller than our previous model. Both models included similar variables but resulted in different coefficients. All variables included were significant at the 95% confidence level. This model was used to make our final predictions of Spread.

Coefficients for Final Model Predicting Spread

variable	coefficient	variable	coefficient
(Intercept)	-13.30814526	AVG_ODREB_PCT_PG_L3_A	4.366478856
AVG_OAPG_TY_H	-0.072106521	AVG_FTA_PG_TY_H	0.122642287
AVG_OAPG_L3_H	-0.103666887	AVG_FTA_PG_L3_H	0.077608842
AVG_OAPG_TY_A	0.042637047	AVG_OFTA_PG_L3_H	-0.036737117
AVG_OAPG_L3_A	0.049176553	AVG_FTA_PG_TY_A	-0.001378882
AVG_OBPG_L3_H	-0.13986924	AVG_OFTA_PG_TY_A	0.085735411
AVG_BPG_TY_A	-0.088320676	AVG_FTA_PG_L3_A	-0.025032657
AVG_OBPG_TY_A	0.139479482	AVG_OFTA_PG_L3_A	0.020809502
AVG_PF_PG_TY_A	0.011651005	AVG_TO_PCT_PG_TY_H	-1.72319762
AVG_W_PCT_L3_H	20.60682519	AVG_TO_PCT_PG_L3_H	-5.001424119
AVG_OW_PCT_L3_H	1.057883976	AVG_OTO_PCT_PG_L3_H	9.606789165
AVG_W_PCT_L3_A	-18.45639539	AVG_eFG_PCT_PG_TY_H	11.13426769
AVG_OW_PCT_L3_A	0.134629753	AVG_OeFG_PCT_PG_TY_H	-15.93220816
AVG_FT_PCT_PG_L3_H	1.170784287	AVG_eFG_PCT_PG_L3_H	13.32261216
AVG_FT_PCT_PG_TY_A	-4.242103743	AVG_eFG_PCT_PG_TY_A	-21.23159103
AVG_DREB_PCT_PG_TY_H	15.9503027	AVG_OeFG_PCT_PG_TY_A	6.121580963
AVG_DREB_PCT_PG_L3_H	6.443208308	AVG_eFG_PCT_PG_L3_A	-7.189350363
AVG_ODREB_PCT_PG_L3_H	-5.804245875	AVG_OeFG_PCT_PG_L3_A	7.6837874
AVG_DREB_PCT_PG_TY_A	-1.755818397		

Calculation Method

To calculate our predictions for total, spread, and offensive rebounds, we used Google Sheets. For each of our variables, we imported the teamrankings.com data into separate sheets using the IMPORTHTML function, refreshing this function each hour to keep the information up

to date. The first six rows of the first sheet look like this:

	A	B	C	D	E
1	NBA Total Projections				
2	**Stats**	**Home**	**Away**	Weight Total H	Weight Total A
3	Team	Charlotte	Chicago		
4	Intercept	1	1	-81.88133815	0
5	Avg Possessions Per Game L3	103.5	100.8	0.181999261	0.450822104
6	Avg Possessions Per Game H/A	104	102.7	1.427794267	1.232838613

Each variable is listed in column A and the coefficients for the total model are listed in column D and E, corresponding to Home and Away respectively. For spread and offensive rebounds, the coefficients are in similar columns with their respective coefficients but due to space are not included in the image above. If a coefficient is not used in that model, as the models are different, the "Weight" is 0. Columns B and C use the VLOOKUP function to search for the corresponding value given the team name and search in the corresponding variable's sheet. For example, B5 looks for Charlotte in the 'Possessions Per Game' sheet and returns the value in the 4th column as that is the column with the Last 3 data as provided by teamrankings.com.

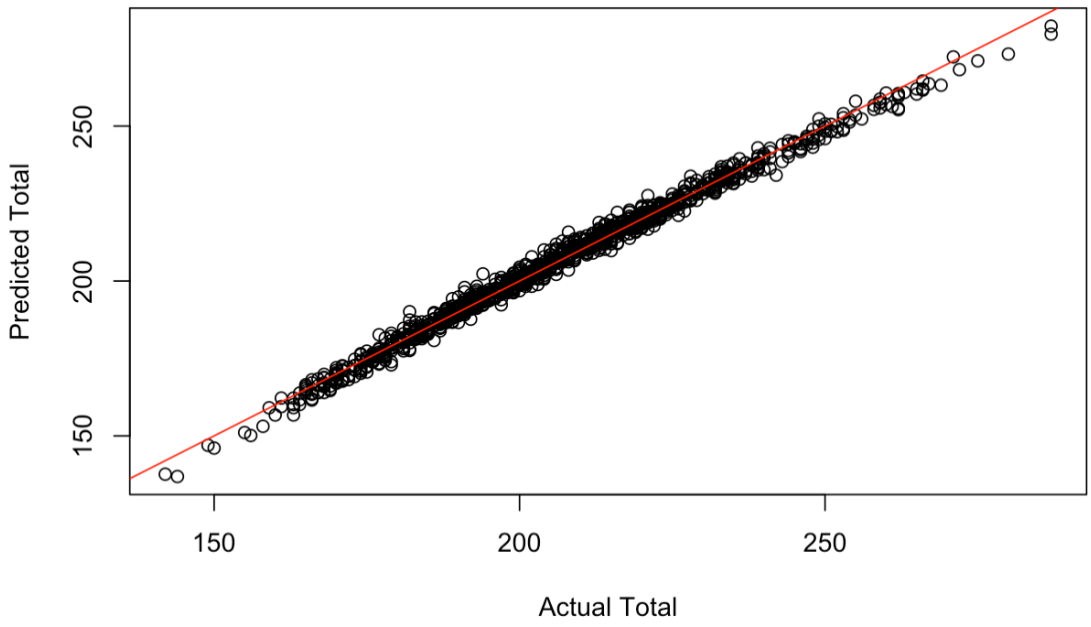
Finally, to calculate the total predictions, we multiplied the cells in column B by column D and column C by column E and added each of the cells together. A similar process was done using the weights for spread and offensive rebounds.

Once we had our predictions for spread, we looked at the average spread across all of our predictions. Our model predicted that all home games would win by an average of 5.5056, which is an increase of 2.7923 from the average spread of 2.7133 across the rest of the 2022-23 season. Using our data dating back to 2010, the spread increases only .1066 in April compared to the other months. We concluded that our model overestimates home advantage by 2.7133-.1066 =

2.6857. Therefore, we subtracted this number from each of our spreads to come to our current spread predictions which now range from -7.45 to 12.83 with an average of 2.82.

b) Total

Before creating our model for Total, we used a bi-directional stepwise progression to identify potential independent variables that would be good predictors of Total using the training data set. This was performed with all of the pre-aggregate variables as potential predictors. This model was cross validated on the testing data and accurately predicted the testing values. The zero mean, variability, and normality conditions were met for the training models ability to predict the testing data. Additionally, the shrinkage value was extremely close to zero. This training models' ability to predict the testing data is seen in the graph below. We will proceed with using the aggregate variables that represent each of the variables that this model identified as good predictors of Total.



Bi-directional stepwise progression was now used again, only taking into consideration aggregate values for the variables we isolated as good predictors of Total in the previous step.

After creating this model, we checked for outliers by analyzing studentized and standardized residuals, leverage values, and cook's distance. There did not appear to be any significant outliers influencing the model. This model was then cross validated to see how accurately it could predict the testing data. The model appeared to predict fairly well, with an rmse of 19.37847.

Creating our model for spread appeared to result in a lower rmse value when LASSO cross validation was used, so we will attempt to use this selection method again. This model selected all the variables that were selected in the previous model, but the coefficients appeared to be very different. However, all of the selected variables were still significant at the 95% confidence level for both selection methods. The models ability to predict the testing data resulted in an rmse of 18.90788, which is less than our previous model. This will be the final model we use to predict Total.

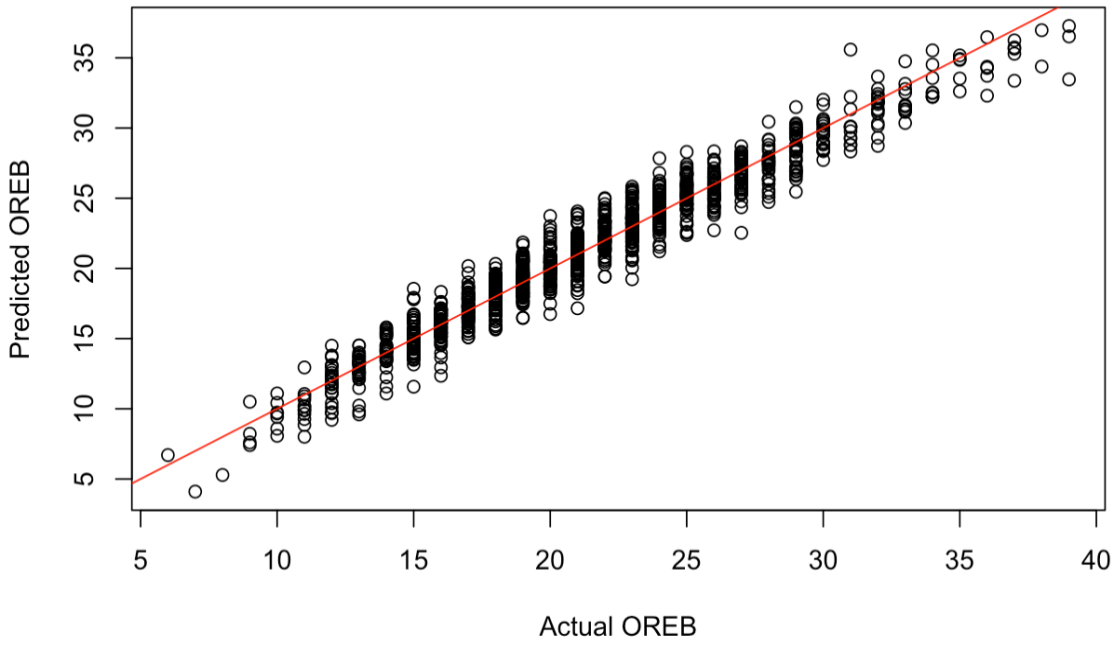
Coefficients for Final Model Predicting Total

variable	coefficient	variable	coefficient
(Intercept)	-81.88133815	AVG_FTA_PG_L3_H	0.135752684
AVG_OAPG_TY_H	0.077091625	AVG_FTA_PG_L3_A	0.066740442
AVG_OAPG_TY_A	0.362325068	AVG_OTO_PCT_PG_TY_H	-67.01561331
AVG_PF_PG_TY_H	-0.160709709	AVG_OTO_PCT_PG_L3_H	2.58354359
AVG_OPF_PG_L3_H	-0.202179587	AVG_TO_PCT_PG_TY_A	5.09389288
AVG_PF_PG_TY_A	-0.176600613	AVG_OTO_PCT_PG_TY_A	-63.65698621
AVG_OW_PCT_TY_H	-9.732582619	AVG_TO_PCT_PG_L3_A	-23.58324522
AVG_OW_PCT_L3_H	-2.163801101	AVG_OTO_PCT_PG_L3_A	-4.526913398
AVG_FT_PCT_PG_TY_H	13.55807381	AVG_POSS_PG_TY_H	1.166596211
AVG_FT_PCT_PG_L3_H	2.762910482	AVG_OPOSS_PG_TY_H	0.261198056
AVG_OFT_PCT_PG_L3_H	0.260591065	AVG_POSS_PG_L3_H	0.181999261
AVG_OFT_PCT_PG_TY_A	-13.77410523	AVG_POSS_PG_TY_A	1.232838613
AVG_OFT_PCT_PG_L3_A	2.307009591	AVG_POSS_PG_L3_A	0.450822104
AVG_DREB_PCT_PG_TY_H	-52.4460992	AVG_eFG_PCT_PG_TY_H	10.01726833
AVG_ODREB_PCT_PG_TY_H	-6.328858625	AVG_OeFG_PCT_PG_TY_H	75.88895758
AVG_DREB_PCT_PG_L3_H	-4.291892107	AVG_eFG_PCT_PG_L3_H	29.96772041
AVG_ODREB_PCT_PG_L3_H	-8.387312731	AVG_OeFG_PCT_PG_L3_H	3.354211852
AVG_DREB_PCT_PG_TY_A	-49.91711275	AVG_eFG_PCT_PG_TY_A	36.12803906
AVG_ODREB_PCT_PG_TY_A	-29.69011116	AVG_OeFG_PCT_PG_TY_A	56.05645074
AVG_DREB_PCT_PG_L3_A	-1.398356639	AVG_eFG_PCT_PG_L3_A	9.265267102
AVG_FTA_PG_TY_H	0.061383383	AVG_OeFG_PCT_PG_L3_A	7.748827133

The calculations for total were done using the *Calculation Method* described in the spread section. Similarly to spread, we looked at the average total across all of our predictions. Our model predicted an average total of 251.94, which is an increase of 22.67 from the average total of 229.27 across the rest of the current season. Using our data dating back to 2010, the total increases only 0.29 in April compared to the other months. We concluded that our model overestimates the total by 22.67-.0.29 = 22.38. Therefore, we subtracted this number from each of our totals to come to our current total predictions which now range from 218.03 to 240.96 with an average of 229.56.

c) Offensive Rebounds

We started the process of creating a model to predict OREB with the same steps as seen when creating the models for Spread and Total. A bi-directional stepwise progression was used to identify potential independent variables that would be good predictors of OREB using the training data set. This was performed with all of the pre-aggregate variables as potential predictors. This model was cross validated on the testing data and accurately predicted the testing values. The zero mean, variability, and normality conditions were met for the training models ability to predict the testing data. Additionally, the shrinkage value was extremely close to zero. This training models' ability to predict the testing data is seen in the graph below. We will proceed withing using the aggregate variables that represent each of the variables that this model identified as good predictors of OREB.



Once again, bi-directional stepwise progression was used to predict OREB with the aggregate values for the variables we isolated as good predictors. After creating this model, we checked for outliers by analyzing studentized and standardized residuals, leverage values, and cook's distance. There did not appear to be any significant outliers influencing the model. This model was then cross validated to see how accurately it could predict the testing data. The model appeared to predict well, with an rmse of 5.508327.

While this appears to be a decent rmse value, we are going to create another model to predict OREB using LASSO cross validation in hopes of potentially reducing the rmse. The LASSO model selected similar variables that were selected in the previous model, with slightly different coefficients compared to the other model. All variables included in this model were significant at the 95% confidence level. The models ability to predict the testing data resulted in an rmse of 5.41755, which is slightly different than our previous model. This will be the final model we use to predict OREB.

Coefficients for Final Model Predicting OREB

variable	coefficient	variable	coefficient
(Intercept)	68.39849456	AVG_ODREB_PCT	-23.32485999
AVG_OPPG_TY_A	0.01275291	AVG_ODREB_PCT	-1.191351138
AVG_OAPG_TY_A	0.034982158	AVG_DREB_PCT_	-6.464886614
AVG_APG_L3_A	0.017633546	AVG_ODREB_PCT	-32.81383277
AVG_BPG_TY_H	0.097986928	AVG_DREB_PCT_	-0.129343945
AVG_BPG_TY_A	0.22297751	AVG_ODREB_PCT	-2.143514255
AVG_BPG_L3_A	0.05464023	AVG_FTA_PG_TY	-0.062078741
AVG_OBPG_L3_A	0.006233945	AVG_OFTA_PG_T	0.0272745
AVG_OSPG_TY_H	0.20790378	AVG_FTA_PG_L3	0.007307369
AVG_SPG_TY_A	0.067253499	AVG_FTA_PG_TY	0.014623673
AVG_OPF_PG_TY_A	-0.003468967	AVG_FTA_PG_L3	0.000632629
AVG_W_PCT_TY_A	-0.770490689	AVG_TO_PCT_PG	-1.000523261
AVG_OW_PCT_TY_A	0.660452144	AVG_OTO_PCT_P	1.175939678
AVG_W_PCT_L3_A	-0.104465381	AVG_POSS_PG_T	0.048939976
AVG_OW_PCT_L3_A	0.94866249	AVG_POSS_PG_T	0.125686415
AVG_FT_PCT_PG_TY_H	-8.174492558	AVG_POSS_PG_L	0.017974979
AVG_OFT_PCT_PG_TY_A	-0.95516359	AVG_eFG_PCT_P	-5.754252368
AVG_FT_PCT_PG_L3_A	-0.896461835	AVG_eFG_PCT_P	-3.372861497
AVG_DREB_PCT_PG_TY_H	-12.42543167		

The calculations for offensive rebounds were done using the *Calculation Method* described in the spread section. Similarly to spread and total, we looked at the average offensive rebounds across all of our predictions. Our model predicted an average of 25.76, which is an increase of 4.91 from the 2022-23 season average of 20.85. Using our data dating back to 2010, offensive rebounds increase only 0.099 in April compared to the other months. We concluded that our model overestimates the number of offensive rebounds by 4.91-.0.099 = 4.811. Therefore, we subtracted this number from each of our model's predictions to come to our current offensive rebound predictions which now range from 16.90 to 25.10 with an average of 20.95.