How Team Statistics Affect Wins Through the History of the NBA

Carter Huskinson
NC State
Raleigh, United States
wchuskis@ncsu.edu

Carson Orejola
NC State
Raleigh, United States
clorejol@ncsu.edu

Cameron Seitz
NC State
Raleigh, United States
cdseitz@ncsu.edu

ACM Reference Format:
Carter Huskinson, Carson Orejola, and Cameron Seitz. 2023. How Team Statistics Affect Wins Through the History of the NBA. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 8 pages. https://doi.org/Team3.CSC422

1 INTRODUCTION AND BACKGROUND

1.1 Problem Statement

There is disagreement over which basketball statistic best captures a team's performance and contribution to the season's success, despite the vast amount of data available. Furthermore, many agree that the game of basketball is constantly changing as different priorities are placed on different things. Finding a concrete answer backed by an experiment will allow insight into how basketball is changing through different eras and what statistics are most vital for basketball. Finding the most pertinent and useful metric for assessing team performance is essential as basketball becomes more data-driven. This study aims to determine which statistic is the most important for predicting success in basketball, and to provide a comprehensive analysis of the statistical metrics currently used in basketball. This is done using automated learning models which will predict and determine which team statistic is the most important for predicting the success of a season across different eras.

1.2 Related Work

Several works have been important to helping us understand this topic. In Cabarkapa and colleagues' study [1], they identified the statistics that lead to wins from a four year period to determine how style of play changes between the regular season and postseason. This study will have a similar exploration, but will use a wider gap of time to compare how style of play changed over several 5 year splits.

Another study vital to determining the most important basketball statistic spanning eras is Giarta's and Asavareongchai's study [2] which uses three different linear regression models to predict the win percentage of a team, given their season statistics. We will also use linear regression in our study to create an accurate model to see how close our predictions for wins will be.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, June 03–05, 2018, Woodstock, NY
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/Team3.CSC422

One of the more important studies we looked over was Horvat's study published in 2023 [3] which used feature extraction to compile their data and results. They also had several unique methods that were beyond the scope of our study which could possible help us understand our results better. They note that their model uses an optimal time window in tow with the rest of their study to formulate their results. This is something that we considered while analyzing our results for this study.

Another interesting study to look at is Lorenzo's study from 2019 [4] which analyzes the game statistics of elite basketball players. For the study, they used players from the ACB, a first division Spanish basketball league, and analyzed their statistical changes across their careers. This is different from our study which looks at NBA players. To visualize and interpret their results, they used a customized excel spreadsheet developed to monitor statistical changes. Having a good understanding of visualizing data and interpreting these figures was something which aided us greatly in this study.

Finally, we also looked over the study by Mikołajec and colleagues [5] to see how we could possible structure and model our own study. In this study, they use a regression model to determine which variables had the highest influence on NBA game performance. Like the present study, they aim to find the most essential indicators to success in the NBA. However, they only explored an eight year span while this present study will also explore different eras of the NBA. Additionally, a more statistical analysis focused approach was taken with their study, for this study, an approach using automated learning will be used.

Each of these works aided in the process of designing all of the processes to be used in this study. How each model was composed and how each data set was analyzed and used in machine learning was influenced from the works mentioned previously. Additionally, the works mentioned allowed us to get an idea of what results we might expect and how we may interpret them. Overall, these works were essential in forming this study.

2 METHOD

2.1 Novelty

For our project, we will be looking at the features in the team stats for the NBA that affect season win rates the most. We will also compare the results of the experiment from different years including past and present season to analyze what has changed in terms of significant statistics. We obtained these results by looking at teams' statistics over a period of twenty seasons. We will analyze the pattern of the weights in the regression to find the most significant statistic.

2.2 Approach

We used several lasso regression models to test our hypothesis. In order to find the best model, we tested our data sets on several regression models including a lasso regression model, and a second order polynomial regression model.

For each model, we used Lasso regression to do feature selection to weed out useless features and better identify the most important features. In this phase of the selection process, the tol value set to default.

For the 1st order lasso regression model, we used a basic linear combination of weights w_i to each feature value x_i plus a bias term w_0 which would result in the equation $y = w_1 * x_1 + ... + w_n * x_n + w + 0$. The weights are adjusted by the model for each iteration to reduce the functions mean squared error based on the test set.

For the second order polynomial regression model, we used a setup very similar to the 1st order by using a linear combination of weights to features for the first part of the equation, but then adding a linear combination of unique weights to the square of the input value. This resulted in the equation $y = w_1 * x_1 + ... + w_n * x_n + ... + w_1 * (x_1)^2 + ... + w_n * (x_n)^2 + w + 0$. In order to achieve this, modified data sets need to be created with each feature of each object to have an additional feature column with the square of the value to fit the equation of the second lasso regression. This model attempts to adjust the weights based on the same optimization function.

We evaluated each model themselves using k-fold validation where k = 10. Each fold is 0.2 of the total data set so for each fold of about 150 test objects the training data is 120 objects with the test fold being 30 objects. For the data points in each fold, we randomly sampled from each season so that we have a good average of how the stats affected regression for that entire time period. We used RMSE as the error function to evaluate each fold.

After we have found which model works best we adjusted the alpha value of the chosen model to find the best hyper-parameter. We used Cross Validation to evaluate each hyper-parameter to select the best parameter.

After the model and hyper-parameters had been chosen, we ran each of the 4 data sets from each time period and retrieve the resulting model. For each of these models, we compared the values of the weights for each feature. The features with the highest weights would be deemed as the most significant.

2.3 Rationale

We decided to do Lasso Regression because our data set has a lot of features for each team and season, and it is unlikely that all these features equally contribute to win rate. Lasso regression lets us reduce these features out of the regression function and compare only the features that are the most significant for our hypothesis.

We decided to use k-fold cross validation on our models because this would be the simplest to implement while also giving us a good idea of the generalization ability of each model. This utilizes the data the best without being too computationally expensive to calculate unlike the LOOCV method.

We decided to use Root Mean Squared error to evaluate as this is the value reduced by Lasso Regression and would therefore give us the best idea of how well each model performs.

We also needed to run models testing different hyper-parameters to make sure every part of the model is as accurate as possible. This ensures that the weights and the significance we are looking for will be the most accurate.

3 EXPERIMENT

3.1 Dataset(s)

The data set used for our experiment comes from the regular season data set at https://www.kaggle.com/datasets/mharvnek/nba-team-stats-00-to-18. It was collected by user Michael H from https://www.nba.com/stats. It has 626 entries, giving information about team statistics for every NBA team's season from the 2000-01 season to 2020-21. It has columns for the season, team name, number of wins, losses, win rate, and the season's game average minutes, points, field goals made, field goals attempted, field goal make rate, three-point shots made, three-point shots attempted, three-point shot make rate, free throws made, free throws attempted, free throw make rate, offensive rebounds, defensive rebounds, total rebounds, assists, turnovers, steals, blocks, blocks against, personal fouls drawn, +/- (score differential).

3.2 Hypotheses

One hypothesis that we will be testing is that the number of three-point shots made have become more positively correlated with winning percentages toward the latter seasons than the first ones. We expect this because there has been a recent trend in the NBA to shoot more three-point shots than teams have in the past.

3.3 Experimental Design

To test the hypotheses, the csv-formatted data was first turned into a data frame. Then, the features were selected to be all the columns except the team name, games played, wins, losses, minutes per game, winning rate, the arbitrary team-season id, rebounds, and the season. This is because the team name would take away from our goal of isolating each team's statistics, the games played only changes depending on the NBA organizational plans, wins and losses depend on the number of games played, the winning rate is our target, the id is just an auto incremented id that shouldn't be taken into account, and the season date will be taken into account later when we break the set into eras. The rebounds will not be considered since this statistic is already represented by offensive rebounds and defensive ones, so the inclusion of this variable would take away from the distinction between the other two. The target, therefore, is selected as the win rate column. This data frame represents the overall data set from 2000-21.

After the initial tests, a pattern emerged with +/- (average score differential) having a larger relative correlation in every experiment with all other features being much lower. Seeing as this feature was not related to our initial hypothesis and the chance that +/- (score differential) contained too much information we ran our experiments again excluding this feature. We also found that when removing this feature, the optimal alpha was 0.0001, so this value was used for all post removal experiments.

As part of the prepossessing, normalization of the features occurs. This uses scikit's MinMaxScalar class to help the effectiveness of the model.

Then, this data set is split into four, each having 5 seasons of data. The first has all the seasons from 2000-01 to 2004-05. The second 2005-06 to 2009-2010, and so on until the last one ends with the 2019-20 season (the 2020-21 season was not used). This was done to create four different models, representing 5 year "eras," and have enough data points per model to create meaningful conclusions while also having small enough "eras" to see change in how the games are played.

Next, the data is split into the training and test sets with a ratio of .2 in order to predict the winning rate without over-fitting. Once the sets are made, we created multiple lasso regression models form sklearn to find the ideal alpha hyper-parameter for the model. To test, a range of alpha values from .0001 to .0007 is used with the model. These variations are tested using 10-fold validation, and the alpha value with the approximate lowest average RSME is selected for our actual model.

With the ideal alpha value found, the best model is used to fit the training data and predict the test data. After our model is trained, metrics like the coefficient of determination and the mean squared error can be evaluated for model accuracy. This process can be done with the whole data set at large and the smaller eras.

To see which team statistics are most important for predicting win rates, the coefficients produced by the lasso regression model are extracted and listed in descending order. The larger the absolute value, the larger impact they had on the lasso regression model and their sign tells if they are positively or negatively affecting the win rates.

This process is then duplicated, this time adding squared columns for each of the features. This second order polynomial regression model lets us see if any of the features, once squared, make a significant impact on the winning rate. Like before, the new data frame is scaled and fitted to the lasso regression. With this completed, the most important characteristics are extracted from the regression model.

These coefficients are then be analyzed and conclusions drawn as to how they change throughout the different eras. One way this is done is through making a line chart, graphing their values across the four eras. This visualizes how they change in both magnitude and sign. Additionally, for the most important statistic for our hypothesis, another line chart is made to graph the league's average value for that statistic across the seasons to see how that feature has changed through time.

In addition to charting the changing team statistics, the data points are plotted on graphs with colors for each era, with one axis being a selected important statistic and the other the win rate. This visualizes the correlation of these statistics and reveals any significant clustering and changes over time if they exist.

4 RESULTS

4.1 Results and Discussion

Before needing to fit a linear regression model on our data set, we looked at the relationship between the win rate and different statistics we are looking at. Figure 1 shows this relationship with the average number of points per game for each team's season. We see what we expect: a positive correlation between the number of

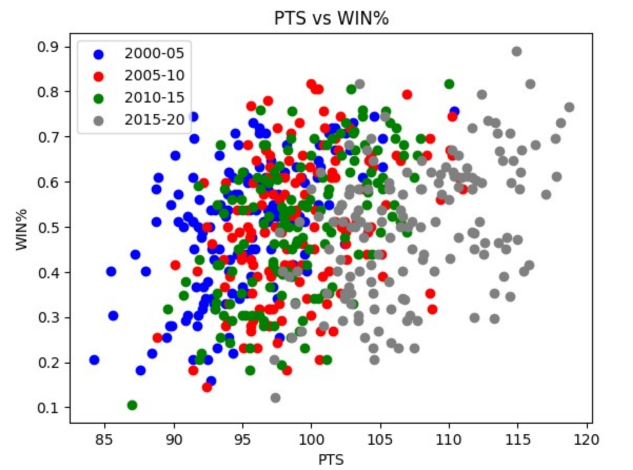

Figure 1: Correlation Between Average Points per Game and the Win Rate

points and the number of wins. The interesting part is the color-coded aspect. By giving each of the four eras its own color, we can see how the average number of points per game has changed from the early 2000's to now. Taking the first and last eras as an example, we can see the majority of the first era's points lower than 100, whereas the most recent era has most of its points higher than 100. This shift to the right going from the first to the last era shows how NBA teams have become more efficient offensively as time passed.

Similarly, in Figure 2, we see a positive correlation between a team's average three-point shot make rate and their winnings. However, we do not see a large shift to the left or right like we saw in Figure 1. Instead, there seems to be one homogeneous group in the middle where all the eras have a similar correlation. This suggests not much change in the accuracy of three point shots during the entire data set date range. There is more variance for the first era and more data points on the left side of the group, however.

Lastly, compare these two to Figure 3, where we see the correlated between the win rate and the average number of three point shots attempted. There is less of a positive correlate between these two variables compared to the first two figures, but we still see the change in number of attempts as time went on as in Figure 1. Later eras have a higher average number of three point shots attempted. Knowing that the accuracy of three point shots have mostly stayed the same from Figure 2 and that teams have been shooting more three point shots recently, it is reasonable to assume that some of the increase in points from Figure 1 can be attributed to this relationship shown between Figures 2 and 3.

Figure 4 shows how we landed upon using .0001 as our alpha hyper-parameter for lasso regression. This was the resulting graph from the experimentation described earlier.

Figure 5 shows the resulting accuracy metrics for our lasso regression models across all 20 seasons. For the first order model (model a), it gets around .0022 MSE, .94 for R^2, and the RSME was around .047. At this point in experimentation, model a suggests that the average game's score differential has the highest impact

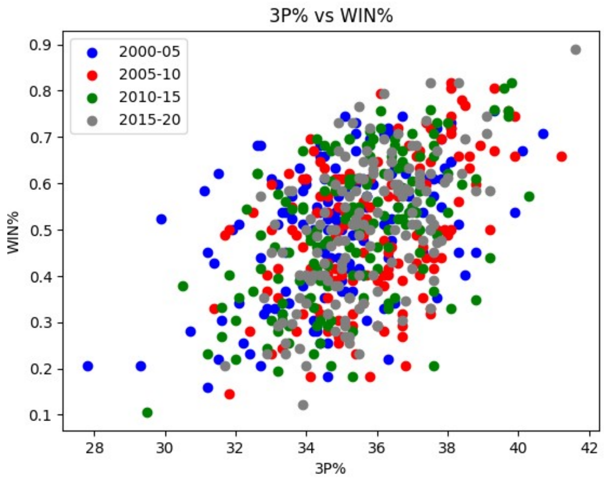

Figure 2: Correlation Between Average Three-Point Shot Percentage per Game and the Win Rate

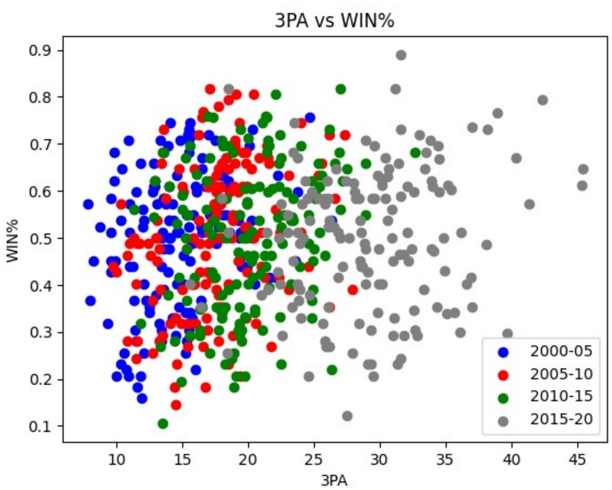

Figure 3: Correlation Between Average Three-Point Shot Attempts per Game and the Win Rate

on the prediction by far. The second highest impact is the team's accuracy with field goals, which makes sense knowing that offensive efficiency is key to winning games. The third highest is the team's turnovers, though negatively correlated. For the second order model (model b), it gets around .0028 MSE, .91 for R^2, and the RSME was around .047. At this point in experimentation, model b again suggests that the average game's score differential has the highest impact on the prediction by far. The second highest impact is the square term for the team's field goals accuracy, which agrees with model a. The third highest is the team's field goal attempts, though negatively correlated. This is interesting knowing field goals are essential to converting to field goal makes. The reasoning for this is that information for actually making the points is contained

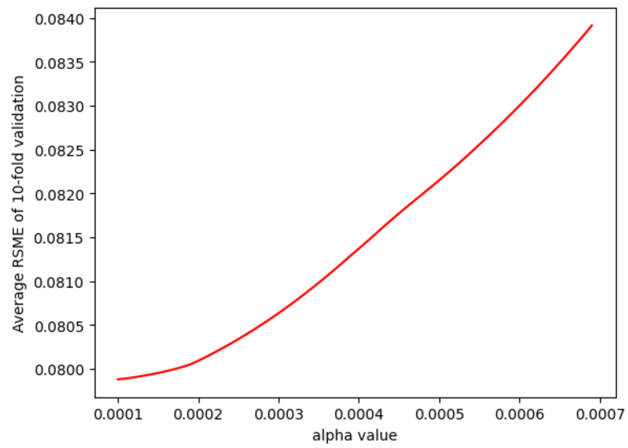

Figure 4: Hyper-parameter Tuning for Lasso Regression's Alpha post removal of the +/- (score differential)

in other variables, while this one also carries information about misses.

Figure 6 shows the results of our accuracy metrics for our lasso regression models across all 20 seasons, except the score differential (+/-) feature has been removed from the data set to observe any new patterns. The first order model (model a) had an MSE of 0.0058, an R^2 of 0.83, and an RSME of 0.08. The increase in the MSE and RSME values are likely due to the score differential feature artificially inflating the initial metrics with its unusually high correlation. In this model, a new pattern emerges with offensive and defensive rebounds having the highest positive impact on predictions. This suggests a teams ability to keep and gain possession is the most important factor in a game. Field goals attempted is shown to be highly negatively correlated to a teams success. The second order lasso regression (model b) shows a very similar pattern. MSE and RSME both increased than the initial model with rebounds once again showing a high positive correlation. Steals also have a high positive correlation to wins in the second order model once again suggesting possession is most important.

In Figure 7, the accuracy metrics and coefficients for each of the 5-year spanning lasso regression models can be found. All have similar R^2 values, around .90, similar MSE values around .003, and similar RSME values around .05. For all of them, the score differential is again the most important metric for determining how many wins a team will have that season. This makes sense knowing how often you outscore your opponent defines wins. Most surprising at this point in experimentation is the change in influence each of the other features have throughout the years. For example, three point shot attempts goes from the second most influential positive factor in the first era to becoming the most negative one in the most recent one. Before conclusions on our hypothesis about the change in three point efficiency can be confirmed or denied, the second approach to disregard the score differential must be examined.

In Figure 8, the coefficients for all 5 year spans for this model are shown, excluding score differential (+/-) to find the regression. All have similar R^2 values, around .80 and similar RSME values around

mse: 0.0022257993975254766
r2: 0.938340067489696

	feature	coefficient
19	+/-	1.020122
3	FG%	0.028965
15	BLK	0.000021
9	FT%	-0.000000
18	PFD	-0.000000
14	STL	0.000000
12	AST	0.000000
11	DREB	0.000000
1	FGM	-0.000000
0	PTS	-0.000000
8	FTA	-0.000000
7	FTM	-0.000000
6	3P%	0.000000
5	3PA	0.000000
4	3PM	0.000000
10	OREB	-0.000000
16	BLKA	-0.000592
17	PF	-0.003874
2	FGA	-0.007246
13	TOV	-0.014044

rsme: 0.04685453021198095

(a) Coefficients for the first order lasso regression

mse: 0.0028091557824866
r2: 0.9149539913110656
rsme: 0.04683943658061661

	feature	coefficient
19	+/-	1.023106
23	FG%^2	0.015146
28	FT%^2	0.000000
21	FGM^2	-0.000000
22	FGA^2	-0.000000
24	3PA^2	-0.000000
25	3P%^2	0.000000
26	FTM^2	-0.000000
27	FTA^2	-0.000000
29	OREB^2	-0.000000
1	FGM	-0.000000
30	DREB^2	0.000000
31	AST^2	0.000000
33	STL^2	0.000000
34	BLK^2	0.000000
35	BLKA^2	-0.000000
36	PF^2	-0.000000
37	PFD^2	-0.000000
20	PTS^2	-0.000000
0	PTS	-0.000000
18	PFD	-0.000000
10	OREB	-0.000000
3	FG%	0.000000
4	3PM	-0.000000
5	3PA	-0.000000
6	3P%	0.000000
7	FTM	0.000000
8	FTA	0.000000
9	FT%	0.000000
11	DREB	0.000000
12	AST	0.000000
13	TOV	-0.000000
14	STL	0.000000
15	BLK	0.000000
16	BLKA	-0.000000
38	+/-^2	-0.000000
17	PF	-0.009997
32	TOV^2	-0.010647
2	FGA	-0.012744

(b) Coefficients for the second order lasso regression

Figure 5: Coefficients extracted from the twenty season - long lasso models

mse: 0.0058173602021842135
r2: 0.8391812620872532

	feature	coefficient
11	DREB	0.881689
10	OREB	0.735532
3	FG%	0.448153
14	STL	0.375631
4	3PM	0.331717
6	3P%	0.167059
12	AST	0.134052
9	FT%	0.129309
15	BLK	0.050187
17	PF	0.007343
0	PTS	0.000000
1	FGM	0.000000
8	FTA	-0.000000
7	FTM	-0.000000
5	3PA	0.000000
18	PFD	-0.020218
16	BLKA	-0.055295
13	TOV	-0.542663
2	FGA	-0.926311

rsme: 0.08039076863124824

(a) Coefficients for the first order lasso regression post removal of the +/- (score differential)

mse: 0.006600662053254992
r2: 0.813252199550592
rsme: 0.07909314801902712

	feature	coefficient
11	DREB	0.923718
10	OREB	0.726406
32	STL^2	0.399127
4	3PM	0.304996
36	PFD^2	0.277724
3	FG%	0.242682
22	FG%^2	0.186333
24	3P%^2	0.183448
12	AST	0.124454
9	FT%	0.122128
15	BLK	0.056657
35	PF^2	0.004080
34	BLKA^2	-0.000000
33	BLK^2	0.000000
30	AST^2	0.000000
29	DREB^2	0.000000
28	OREB^2	0.000000
27	FT%^2	0.000000
25	FTM^2	-0.000000
23	3PA^2	0.000000
0	PTS	0.000000
20	FGM^2	0.000000
19	PTS^2	0.000000
1	FGM	0.000000
17	PF	0.000000
14	STL	0.000000
13	TOV	-0.000000
8	FTA	-0.000000
7	FTM	-0.000000
6	3P%	0.000000
5	3PA	0.000000
16	BLKA	-0.043361
26	FTA^2	-0.053398
18	PFD	-0.249388
2	FGA	-0.376620
31	TOV^2	-0.542310
21	FGA^2	-0.558399

(b) Coefficients for the second order lasso regression post removal of the +/- (score differential)

Figure 6: Coefficients extracted from the twenty season - long lasso models post removal of the +/- (score differential)

0.8 - 0.1. The MSE values are for more varied ranging from 0.004 - 0.01. Though the accuracy measures are worse when excluding the +/- feature due to the high amount of information contained within that statistic, this model is more important because of the obviousness contained within that particular feature. It minimized the importance of the other statistics which now come to light. Following a similar pattern to the total set of data, offensive rebounds and defensive rebounds remain the highest for most eras. This suggests that possession is and continues to be the most important factor when determining wins. In terms of interesting evolution's through the different eras, FG% steadily moved higher in terms of relative value from fourth to second for the last era. Steals starts out at third and moves down to fourth most positively correlated. In terms of our hypothesis, three-point shots attempted and three-point shot percentage remain fairly close to zero relative to the rest. However, the importance of the number of three points made increased consistently from the second to the third and fourth era. This consistent increase recently confirms our hypothesis that making three point shots has become more and more important in the most recent seasons, compared the ones at the beginning of the 2000s.

These trends can be visually seen in Figures 9 and 10, which track the coefficient of the features across the four eras (except +/-). Figure 9 has all of them, while Figure 10 takes the 10 which have the highest mean absolute value. This eliminates the ones that have little effect on the outcome of the games. Like before, we see what we expect. Turnovers are consistently negatively correlated

Era: 1
mse: 0.0050373230463494146
r2: 0.9092786969773268
rsme: 0.053814966879066485

	feature	coefficient
19	+/-	0.973795
5	3PA	0.029285
17	PF	0.023250
14	STL	0.016642
7	FTM	0.011068
18	PFD	0.008100
11	DREB	0.007401
13	TOV	0.001866
15	BLK	-0.000000
12	AST	0.000000
0	PTS	0.000000
1	FGM	0.000000
9	FT%	0.000000
8	FTA	0.000000
4	3PM	0.000000
3	FG%	0.000000
2	FGA	-0.000000
10	OREB	-0.000000
6	3P%	-0.011659
16	BLKA	-0.026096

(a) Coefficients for the first order lasso regression for era 1

Era: 2
mse: 0.0022787229907827843
r2: 0.9550365471630211
rsme: 0.053866651291143944

	feature	coefficient
19	+/-	0.898448
11	DREB	0.047690
10	OREB	0.038497
15	BLK	0.034635
18	PFD	0.025513
12	AST	0.025243
3	FG%	0.000000
1	FGM	-0.000000
14	STL	-0.000000
0	PTS	-0.000000
9	FT%	-0.000000
8	FTA	-0.000000
7	FTM	-0.000000
5	3PA	-0.000000
4	3PM	-0.002598
6	3P%	-0.011859
17	PF	-0.020536
2	FGA	-0.036196
16	BLKA	-0.042417
13	TOV	-0.061281

(b) Coefficients for the second order lasso regression for era 2

Era: 3
mse: 0.004582355284906789
r2: 0.8901157445690999
rsme: 0.05473854802124606

	feature	coefficient
19	+/-	1.045963
3	FG%	0.027019
14	STL	0.007690
6	3P%	0.005333
15	BLK	0.002464
11	DREB	-0.000000
18	PFD	0.000000
16	BLKA	-0.000000
13	TOV	-0.000000
12	AST	0.000000
0	PTS	0.000000
1	FGM	0.000000
9	FT%	0.000000
8	FTA	-0.000000
7	FTM	-0.000000
5	3PA	-0.000000
4	3PM	0.000000
10	OREB	-0.000000
17	PF	-0.032501
2	FGA	-0.035034

(c) Coefficients for the first order lasso regression for era 3

Era: 4
mse: 0.002553587178755485
r2: 0.9429395163761489
rsme: 0.050075472272238894

	feature	coefficient
19	+/-	0.840903
6	3P%	0.043218
3	FG%	0.027646
17	PF	0.003361
13	TOV	0.000134
1	FGM	0.000000
18	PFD	-0.000000
16	BLKA	-0.000000
14	STL	-0.000000
11	DREB	0.000000
0	PTS	-0.000000
9	FT%	-0.000000
8	FTA	0.000000
7	FTM	-0.000000
4	3PM	-0.000000
10	OREB	0.000000
15	BLK	-0.005320
12	AST	-0.008553
2	FGA	-0.008775
5	3PA	-0.016137

(d) Coefficients for the second order lasso regression for era 4

Figure 7: Coefficients extracted from each five-year-long era lasso models

Era: 1
mse: 0.01061999098984859
r2: 0.8177413106509631
rsme: 0.10204817795758367

	feature	coefficient
10	OREB	0.595873
11	DREB	0.586957
14	STL	0.560552
3	FG%	0.462065
5	3PA	0.196641
9	FT%	0.171883
4	3PM	0.156349
17	PF	0.111911
15	BLK	0.103808
12	AST	0.096904
18	PFD	0.027817
1	FGM	0.000000
8	FTA	-0.000000
6	3P%	-0.000000
0	PTS	0.000000
16	BLKA	-0.033607
7	FTM	-0.041226
13	TOV	-0.750197
2	FGA	-0.769196

(a) Coefficients for the first order lasso regression for era 1 post removal of the +/- (score differential)

Era: 2
mse: 0.005920070228865872
r2: 0.8810398642603287
rsme: 0.0826144262895686

	feature	coefficient
11	DREB	0.530835
18	PFD	0.501456
10	OREB	0.458350
3	FG%	0.404236
14	STL	0.389496
5	3PA	0.175257
12	AST	0.131705
6	3P%	0.119554
9	FT%	0.080700
15	BLK	0.051757
4	3PM	0.049215
1	FGM	-0.000000
7	FTM	-0.000000
0	PTS	0.000000
17	PF	-0.002329
16	BLKA	-0.022754
8	FTA	-0.408890
13	TOV	-0.584365
2	FGA	-0.779457

(b) Coefficients for the second order lasso regression for era 2 post removal of the +/- (score differential)

Era: 3
mse: 0.006453629253506226
r2: 0.8504441239927291
rsme: 0.08803832148004694

	feature	coefficient
10	OREB	0.610442
11	DREB	0.538026
3	FG%	0.497689
14	STL	0.384369
4	3PM	0.234400
9	FT%	0.228796
18	PFD	0.160253
6	3P%	0.080658
12	AST	0.076042
15	BLK	0.052331
7	FTM	-0.000000
1	FGM	0.000000
5	3PA	0.000000
0	PTS	0.000000
16	BLKA	-0.027077
17	PF	-0.102842
8	FTA	-0.168577
13	TOV	-0.475019
2	FGA	-0.591819

(c) Coefficients for the first order lasso regression for era 3 post removal of the +/- (score differential)

Era: 4
mse: 0.008080890185734447
r2: 0.6886922454201195
rsme: 0.08124529399630201

	feature	coefficient
11	DREB	0.787409
3	FG%	0.530235
10	OREB	0.477857
14	STL	0.379530
4	3PM	0.330015
9	FT%	0.090852
5	3PA	0.089290
6	3P%	0.086168
15	BLK	0.052719
18	PFD	0.045116
16	BLKA	0.030894
17	PF	0.026414
1	FGM	-0.000000
7	FTM	-0.000000
0	PTS	0.000000
12	AST	-0.001580
8	FTA	-0.115907
13	TOV	-0.421517
2	FGA	-0.735616

(d) Coefficients for the second order lasso regression for era 4 post removal of the +/- (score differential)

Figure 8: Coefficients extracted from each five-year-long era lasso models post removal of the +/- (score differential)

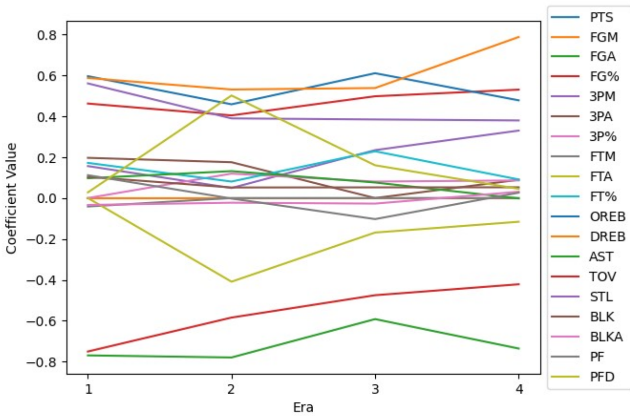

Figure 9: Change in Coefficients for each feature across each era post removal of the +/- (score differential)

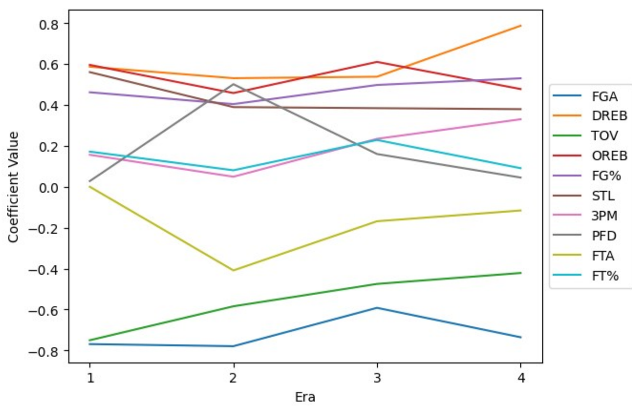

Figure 10: Change in Coefficients for the top 10 features across each era post removal of the +/- (score differential)

because it causes teams to lose possession of the ball and field goal attempts contains information for missing shots. The most consistent positively correlated stats are defensive and offensive rebounds, field goal accuracy, and three points made.

Now that we have evaluated our hypothesis and seen an increase in the importance of making three-point shots, one might ask: "Have NBA teams noticed this as well?" Looking at Figure 11, there has been a constant increase in the number of three point shots teams have been attempted. However, in the most recent era evaluated, around 2015-2020, there is a massive jump. This is exactly what we would expect following the increase in importance of the three-point shots made statistic going from the second to third era. It appears that teams across the league recognized the evolution of basketball as we have and increased focus on attempting to shoot more three point shots.

To review the final results, we discovered that the data post the removal of the score differential is more relevant to our hypothesis and forms the basis of our model. When analyzing our hypothesis, we found that we were correct that three-point shots did increase

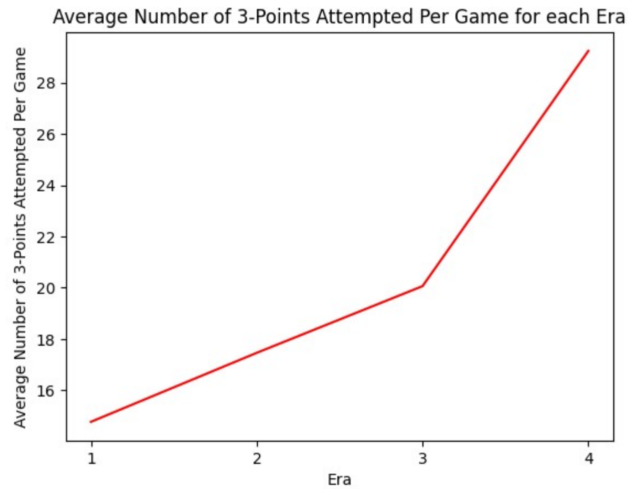

Figure 11: The Average Number of Three-Point Shots Attempted Per Game During Each Era

in significance. However, it was still overall less importance than field goals in general. 3P% starts out at zero and does gain a positive correlation but still remains relatively low. The average number of three-point shots made, however, ended up becoming the fifth most important positive feature. This would support the idea that three-point shots are becoming more important and winning teams are relying on them more. We saw this reliance in Figure 11 where we saw a constant increase in the number of three-point shots taken, especially in the most recent era. The increase in FG% throughout each time period also shows that shooting accuracy as a whole (including three-point shots) is increasing in significance for the NBA. It might be necessary to adjust are hypothesis to consider shooting as a whole becoming more important rather than just three-point shots.

Our results partially reflect previous works but differ in several ways as well. Much like Cabarkapa's results [1], we also see that field goal percentage and general shooting efficiency has a very high positive impact on winning NBA games. Additionally, studies like [2] also show that three points mad is also a very significant statistic. However, in their study they also analyzed things like the number of All-Star players (an award for exceptional players), which ended up being the most important statistic in their study. This trend of offensive efficiency being incredibly important is also seen in [5] which determined that offensive efficiency and win percentage were the most key statistics to winning NBA games. However, it should be noted that each of the studies that we have mentioned have taken different approach to ours. Our results offer another angle to either challenge or affirm what other studies might have already learned. With this, we can continue to unpack and determine how NBA basketball is changing and how the game can be optimized statistically with data science and automated learning.

5 CONCLUSION

Through this project, we learned the importance of pre-processing data and discovered through the inclusion and exclusion of the score differential how features which capture most of the target information can cloud the importance of other statistics from the data set.

We also learned practical applications of a lasso regression model, and how it differs from a traditional linear regression model. We were able to experiment using both models and see how Lasso regression narrows the results of our experiment and makes interpreting easier.

In terms of the problem itself it was interesting to see how significant patterns could be found by processing data over a period of time. We were able to explore a more complex problem and solution simply by adding another dimension to how we used data and machine learning approaches.

Our results can be refined and used by basketball teams around the world to focus on the team statistics which contribute the most to wins. We also got to see how teams in the NBA have changed playing styles, important for historical reference when comparing players across different eras. This research can also be generalized and adapted to other sports.

In terms of problems we would like to explore in the future with this data set, we would further condense redundant features such as 3PA, 3PM, 3P% and see if this improves results or creates even more interesting patterns in our regression models.

6 MEETING SCHEDULE

- April 17, 2023 at 4:15PM for 1 hour | Attended: Cameron, Carter, Carson
- April 19, 2023 at 4:15PM for 1 hour | Attended: Cameron, Carter, Carson
- April 21, 2023 at 4:15PM for 1 hour | Attended: Cameron, Carter, Carson
- April 24, 2023 at 4:15PM for 1 hour | Attended: Cameron, Carter, Carson

REFERENCES

[1] Dimitrije Cabarkapa, Michael A. Deane, Andrew C. Fry, Grant T. Jones, Damjana V. Cabarkapa, Nicolas M. Philipp, and Daniel Yu. 2022. Game statistics that discriminate winning and losing at the NBA level of basketball competition. *PLOS ONE* 17, 8 (August 2022). https://doi.org/10.1371/journal.pone.0273427

[2] Evan Giarta and Nattapoom Asavareongchai. 2015. Predicting Win Percentage and Winning Features of NBA Teams. http://cs229.stanford.edu/proj2015/117_report.pdf

[3] Tomislav Horvat, Josip Job, Robert Logozar, and Časlav Livada. 2023. A data-driven machine learning algorithm for predicting the outcomes of NBA games. *Symmetry* 15, 4 (2023), 798. https://doi.org/10.3390/sym15040798

[4] Jorge Lorenzo, Alberto Lorenzo, Daniel Conte, and Mario Giménez. 2019. Long-Term Analysis of Elite Basketball Players' Game-Related Statistics Throughout Their Careers. *Frontiers in Psychology* 10 (February 2019). https://doi.org/10.3389/fpsyg.2019.00421

[5] Kazimierz Mikołajec, Adam Maszczyk, and Tomasz Zając. 2013. Game indicators determining sports performance in the NBA. *Journal of Human Kinetics* 37, 1 (July 2013), 145–151. https://doi.org/10.2478/hukin-2013-0035