Predicting NBA Wins Based on Performance in Previous 10 Games

CSE 158 Assignment 2

Christopher Liu
UC San Diego
cmliu@ucsd.edu

Derek Shibata
UC San Diego
dkshibat@ucsd.edu

Brandon Tsui
UC San Diego
bhtsui@ucsd.edu

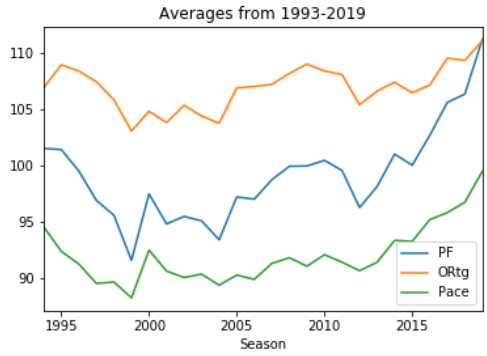
1 DATASET

The data we used were advanced game logs of every NBA regular season game from the 1993-1994 season to the 2018-2019 season. These advanced game logs, pulled from https://www.basketball-reference.com/ scraped using Beautiful Soup 4 include the fields in Tables 1 and 2.

Exploratory Analysis

Despite the league currently having 30 teams since 1993, there have been expansion teams added and many teams have moved locations or changed names, some on multiple occasions.

One of the most obvious trends we saw when looking at the initial data are the differences in play style and statistics between NBA seasons. The average of these statistics are not consistent between seasons, as we seen in the following graphs:



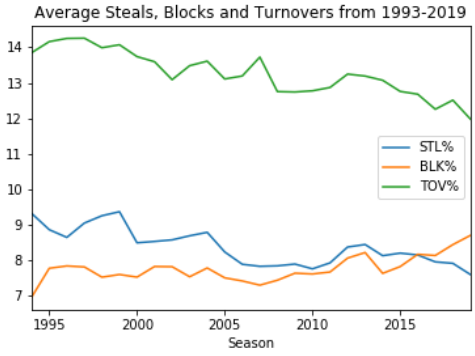
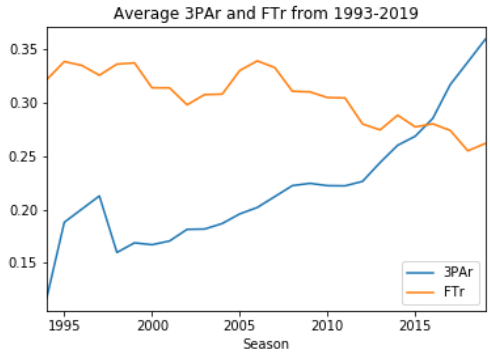
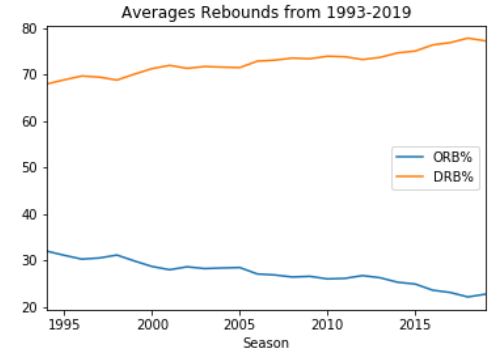


Table 1: NBA Game Raw Data Fields

Field Abrev	Field Name	Description
G	Game Number	Game in the season
Date	Date	Day of game
Away	Home or Away Game	1 if away game, else 0
Opp	Opponent	Opposing team
PF	Points For	Points scored
PA	Points Against	Points allowed
ORtg	Offensive Rating	An est. of pts scored per 100 possessions
DRtg	Defensive Rating	An est. of pts blocked per 100 possessions
Pace	Pace	An est. of possessions per 48 minutes
FTr	Free Throw Attempt Rate	Number of free throw attempts per field goal attempts
3PAr	3 Point Attempt Rate	Percentage of field goal attempts from 3 point range
TS%	True Shooting Percentage	Shooting efficiency accounting for type of shot
TR%	Total Rebound Percentage	Percentage of available rebounds grabbed by the whole team
AST%	Assist Percentage	Percentage of field goals that were assisted
STL%	Steal Percentage	Percentage of opponent possesions that ended in a steal
BLK%	Block Percentage	Percentage of 2 point field goals attempts that ending in a block
eFG%	Effective Field Goal Percentage	Field goal % adjusting for difference between 2 and 3 point field goals
TOV%	Turnover Percentage	An est. of turnovers per 100 possessions





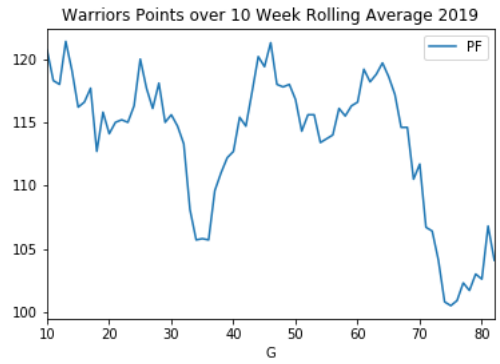
In these graphs it is shown how pace and scoring has drastically increased over the years partly due to the increase in three point attempts. Other stats such as steals and turnovers have decreased while blocks have increased.

Many of these trends were to be expected based on the highly documented rise in outside shooting. Three pointers and shots within 4 feet are the most efficient shots, and as data analytics rose in the NBA, teams started taking more of these shots, which resulted in a higher scoring and faster paced play style accompanied by the other changes noted earlier. These findings highly motivate how we created our features. However, we cannot compare game performance to total averages but need to adjust for the year in which the game is played. The simplest way to do this is by making features using the difference between the statistics of both teams.

Another observation/concern found while exploring the data are the trends within a given season for particular teams. One very important factor that this data does not capture are injuries, free agent signings and other changes to the composition of a team throughout a season. Typically, we see a drop off in performance when star players exit the lineup which of course would change the prediction we are trying to make.

Table 2: NBA Game Raw Data Fields (Continued)

Field Abrev	Field Name	Description
ORB%	Offensive Rebound Percentage	% of available offensive rebounds grabbed by the team
DRB%	Defensive Rebound Percentage	% of available defensive rebounds grabbed by the team
FT/FGA	Free Throws per Field Goal Attempt	Free throws per field goal attempt
Opp_eFG%	Opponent's Effective Field Goal Percentage	Opponent's effective field goal percentage
Opp_TOV%	Opponent's Turnover Percentage	Opponent's turnover percentage
Opp_FT/FGA	Opponent's Free Throws per Field Goal Attempt	Opponent's free throws per field goal attempt



For example, the above graphs show just how drastically different the Warriors points averages were at different weeks in 2019, mostly explained by the injuries to their star players preventing them from playing. Not only that, but slumps and streaks are real in the NBA, and we wanted to try to account for these different factors. Teams have stretches where they perform better or worse and these trends may not be captured by long term or season wide averages.

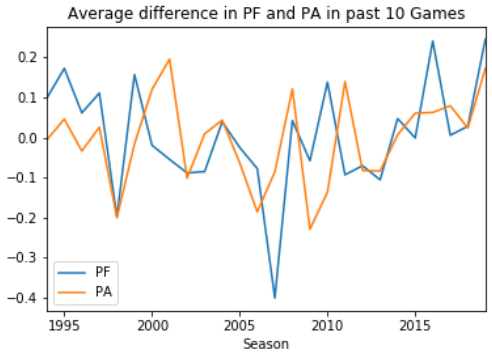
Because of this, we transformed the data to make predictions based on average stats of the past 10 games for each team similar to a rolling average. This method is not perfect but allows capturing temporary trends better than season averages or averages across multiple years. Doing this, however, limited the games we could predict to only games past the 10th game of a season for each particular team.

2 PREDICTIVE TASK

With our data, we could make two kinds of meaningful predictions, either simply predicting the winner of each game or predicting the score differential of the game. In this report, however, we focused on simply predicting the winner of the game, making it a binary classification problem.

Obviously for our model, we wanted to have an accuracy of at least over 50%, which would be obtained by a random predictor or predicting all wins or all losses. As such, the random predictor we created had an accuracy of 0.497. However, another model to compare to would be a simple classifier based only on win percentages.

As mentioned before, for the advanced statistics noted in Table 1 and Table 2, we reformatted the data so that for each game, each of these stats had the average in that category from the past 10 games. Then, we calculated the averages for those statistics of the opponents past 10 games as well and took the difference between the two teams for a final version of each statistic. Additionally, as we wanted to obtain a metric that wouldn't be significantly skewed by data points being from different years, we analyzed the trends of statistics per year.



In this example, as difference in PF and PA does not have a significant trend over the years from 1993-2018, we could assume that the two were not statistics biased by year and should work on any data points in the set. However, a thing to note is by entirely changing data points to be based on the average of the previous 10 games and based on a different, the feature data points would essentially become anonymized from their team, due to them not following any specific trend of any team. Thus our predictive task would be able to predict whether a team would win or not purely based on the statistics of the past results between two teams. Thus team name would not be related or needed past generation of the dataset.

Some simple transformation needed to be applied to certain variables as well, such as encoding whether not the game was home or away to 0 or 1. Another feature we created was the number of days since each team last played by filtering the data and subtracting dates between the target game and the previous game played. We believed this had impact the outcome of games due to the tiredness of players.

Again, we took the difference between the team and the opponents days since the last game to account for extended breaks in play such as the All Star break. Overall, we used 23 different features based on advanced statistics, home or away, and difference in games since last played between the teams. After making these features, we split the data into training, validation, and test sets on a 70/15/15 split.

Table 3: C vs Validation Accuracy

C	Validation Accuracy
0.001	0.6672
0.01	0.6740
0.1	0.6718
1	0.6708
10	0.6712
100	0.6715

3 MODEL

While trying to create the best possible model, we tried many different methods and models as well as adjusted the features along the way. First we started off with something basic, predicting wins when the team had a better win percentage in the last 10 games than the opponent. This yielded an accuracy of just 62.63%. Even with adjustments such as an extra condition of predicting win when the difference in W/L was zero but the difference in PF was higher, accuracy only rose to 62.95%.

After creating this baseline model, we started using logistic regression, deciding that this would be appropriate for predicting a binary classification based on the format of the features. This is also preferable to a simpler method such as naive Bayes because many of the statistics are not independent such as those related to shooting and scoring. Logistic regression gets rid of the double counting problem since some features may have the same kind of effect on the prediction.

In order to optimize this approach, we implemented a regularization pipeline, testing the effect of different regularization coefficients [0.001:100] on the accuracy on the validation set. This was done to try to avoid over-fitting. The pipeline produced the following results in Table 3.

The logistic regression model with C=0.01 ended up having the best performance overall on the validation set but we also tried using other models as well. Both ridge regression and SVM were less successful than logistic regression with

Table 4: PCA Analysis

C	n	Validation Accuracy
0.001	5	0.6474534161490684
0.001	10	0.6478260869565218
0.001	15	0.6704347826086956
0.001	20	0.6704347826086956
0.01	5	0.6474534161490684
0.01	10	0.6480745341614906
0.01	15	0.6727950310559005
0.01	20	0.6732919254658385
0.1	5	0.6474534161490684
0.1	10	0.6480745341614906
0.1	15	0.6722981366459627
0.1	20	0.670807453416149
1	5	0.6474534161490684
1	10	0.6480745341614906
1	15	0.6726708074534162
1	20	0.6704347826086956
10	5	0.6474534161490684
10	10	0.6480745341614906
10	15	0.6729192546583851
10	20	0.6704347826086956
100	5	0.6474534161490684
100	10	0.6480745341614906
100	15	0.6729192546583851
100	20	0.6704347826086956

accuracies around 0.62-0.64. This is to be expected especially since support vector machines are typically more effective in higher dimensional spaces while this data set doesn't nearly have that many dimensions.

In addition, because the dataset is not linearly separable, the advantage of svm is slightly diminished. Another method we tried to use in combination with logistic regression was running PCA to try to minimize the dimensions. This method, however, did not improve performance, indicating that there was no need to decrease the dimensions and that all of the features were at least slightly impactful. Finally, knowing that logistic regression performs better without duplicate features, we sought to remove some features that while not exactly the same, contained similar data.

In the end, removing eFG%, FT/FGA, and BLK% increased the accuracy on the validation set when C=0.01 to 0.6775. While most features were strong, there are still areas in which our model was week. While the 10 game rolling average helps account for recent trends, it cannot adjust for randomness in the schedule. In addition, days since last game certainly helps adjust for the tiredness of players but this and the average still cannot fully account for when injuries happen and when star players are in and out of the lineup. Making a feature that can capture this was simply out of the scope of time and resources we could put into the making of this model. In theory, we would also need statistics for every player on the team including impact metrics which calculate individual player's impact on most of these major statistics. Even lacking these, our model was still strong while using minimal resources and easy to implement for predicting future games in the modern era of basketball.

4 LITERATURE

The dataset we created uses data taken from https://basketball-reference.com, which is available to the public and has been used to analyze the NBA and create predictors in the past such as in the study done by Torres. However, as we desired to create a sample with a substantial number of data points, we decided to analyze games beginning from 1993 to 2019, a span of 25 years. In contrast, many of the studies that we looked at had either a much smaller sample due to a methodology of simple stratified random sampling (primarily statistically based studies, such as the Magel and Unruh study or the Jones study), or used a smaller time frame (the Torres study). For the most part, our dataset was created from scratch, though this does not exclude coincidental overlapping.

Another big difference between our study and the other studies that have been done in the past is our testing methodology. Normally, these predictors are tested on games that we do not know the outcome of yet (future games). However, since we had a limited time table to complete the study, we have only been able to use a testset that contains games in the past that we already know the outcome of. Because of this, the goals of the previous studies differed from ours. For instance, the study by Eric Scot Jones was specifically geared towards predicting results of the 2016 NBA Playoffs. On the other hand, we attempted to predict the outcome/spread of random games that were selected from our dataset of 35 years and roughly 61000 games. However, despite these differences, each study did have a take-away of some sort that was recognized and accounted for within our own work.

In the study written by Rhonda Magel and Samuel Unruh [2], the two desired to find which statistics in a basketball game could be most significant in causing a victory or defeat. A primary difference between our study and theirs is that their dataset was on college basketball, meaning NCAA games instead of NBA games. However, a key similarity between their model and ours was the decision to use independent variables comprised of differences between the statistics of two teams. Although their reason for doing so was not explicitly explained, we assume it was not due to trying to protect the data against time bias, but rather to have a single number to use for each statistic.

Related to the Magel and Unruh study, the Eric Scot Jones study [1] sought to take the idea of the former and apply it to the NBA level, hoping to be able to predict the 2016 NBA Playoffs and Finals results. Jones' study at first glance seems to be very similar to ours, also using the idea of differences between two teams as the features of his model, but due to his objective of finding which team in a specific match-up would be predicted to win, his dataset was not nearly as general as ours. Additionally, because of Jones' goal to find results between specific match-ups of teams, he took the seasonal averages for the statistics of teams, which was not possible in our dataset due to transformed data. Though we designed our feature vector ourselves, this study had many overlaps in choice of features and helped solidify confidence in specific features, namely difference in assists and the difference in turnovers between teams.

In terms of other parts of model design, the Torres study [3] gave insight on models based on the last N games, similar to our previous 10 games model. The study found that the previous 8 games would provide the highest percentage, but due to the way we manipulated our dataset, we found it would be too difficult/expensive to do an analysis on the last N games, especially as we had a significantly larger number of features than the Torres study (the Torres only used win-loss percentage).

After adding in the features mentioned above, our model ended up having an accuracy of about 67.5%. This was the highest accuracy we were able to achieve out of all the different models. Compared to the other predictions such as that run by Jones, ours was slightly better, with Jones reaching an accuracy of about 62%. This is roughly the same as the study done by Torres, who got an accuracy of 67.9%.

5 RESULTS

Looking at our results, an improved accuracy of 67.75% on the validation set is about 17.% better than the random models and 5% better than the simple baseline model. When running the model on the test set, we got an accuracy of 66.1%. In terms of practicality, this accuracy is not overly impressive, but would suffice enough for betters to make a profit if they were betting based on our model. However, most betting relies on not only predicting the winner correctly, but also the point spread. Because of this, our model is not all that useful.

For this assignment, we focused mainly on improving the accuracy of predicting the winner but in doing so, still found interesting trends about the impact of statistics on the outcome of games based on our final set of features. In the end we used 19 features: 'Away', 'Days since last played', 'W/L', 'PF', 'PA', 'ORtg', 'DRtg', 'Pace', 'FTr', '3PAr', 'TS%', 'TRB%', 'AST%', 'STL%', 'TOV%', 'ORB%', 'Opp_eFG%', 'Opp_TOV%', 'DRB%', with all of these except for "Away", being calculated by the difference in average of each statistic between the team and the opponent in their last 10 games.

Conclusion

Of course basketball cannot be simplified to pure numbers and the result of a game will never be able to be predicted with a high degree of accuracy because otherwise there would be no point in watching the game. There are so many other factors involved in the outcome of a game that cannot be quantified, mainly due to the fact that the game is played my humans and humans change, their mood changes, their energy level changes, etc.

Still, this investigation yielded meaningful results. We found which features/statistics were important in predicting games and which were not. Free throw rates and block rates, turn out not to be impactful while turnover and steal rates are. Other than just for betting, using the results of these feature investigations also helps us determine which players are valuable to the team and can help general managers search for players which aid in areas need improving.

While this may not quite be a good enough model to bet and make money off of, the gambling companies already have extremely good models which are used to create the spreads in the first place. If anything the relative failure of models still goes to show how analytics will never be able to perfectly predict the future and how the "Any given Sunday" motto, while mainly used to reference football, applies to the NBA as well. A so called "bad" team always has a chance of upsetting a "good" team.

REFERENCES

[1] Eric Scot Jones. 2016. Predicting Outcomes of NBA Basketball Games. (April 2016). https://hdl.handle.net/10365/28084

[2] Rhonda Magel and Samuel Unruh. 2013. Determining Factors Influencing the Outcome of College Basketball Games. *Open Journal of Statistics* 3, 4 (July 2013), 225-230. https://doi.org/10.4236/ojs.2013.34026

[3] Renato Amorim Torres. 2013. Prediction of NBA Games Based on Machine Learning Methods. (Dec. 2013). https://homepages.cae.wisc.edu/~ece539/fall13/project/AmorimTorres_rpt.pdf