Viability of NBA Advanced Metrics in Predicting Wins

By Brian Zamkotowicz

January 2019

Springboard Career Track Capstone Project #1

Section 1: Introduction

Starting with Bill James in the mid 1970's examining baseball, sports statistics have undergone a revolution. No longer is the argument "Who's the best of all time?" something to just be argued about amongst friends at the local pub. It is a question that is attacked with complex mathematics, and advanced statistics developed specifically for that purpose. The use of these statistics stretches well beyond the sports fan. In recent years (again starting with baseball), some of this information has even led to changes in on the field strategy (shifting infielders to one side of the field in baseball for instance). Nowhere is the impact of advanced statistics felt more keenly than in the front office of professional sports teams. The influence of statistics is so prevalent, in fact, that a new term, "Moneyball", was invented to refer general managers using advanced statistical metrics to build winning teams in a cost effective way.

Problem Description

The following project seeks to utilize readily available data from the National Basketball Association (NBA) in the hopes of gaining insight into the statistics that may have predictive power for determining wins and losses in an upcoming season. From a business perspective this information could be useful to NBA general managers in the construction of their teams. If the relationships between certain statistics and a positive change in wins in the coming season can be ascertained, then the predictive value of those statistics can be used as basis for off-season player personnel decisions. In summary, I would like to determine which statistics are most relevant in predicting wins, since this information is relevant in connection with making business decisions regarding team construction.

Section 2: Approach

Section 2:1 Data Wrangling and Analysis

The first thing necessary before proceeding with the analysis was to find relevant data. Because basketball, and sports in general, are very statistics-oriented much of the information was readily available on the internet. In this case, the website basketballreference.com provided a wealth of information. In addition to well-known statistics like points, rebounds, assists, locks and steals, which even a casual basketball fan is thoroughly familiar with, they also provide a number of advanced metrics. Because these advanced metrics are intended to provide a clearer picture of how players perform relative to one another, I chose to focus on several of these advanced statistics as a jumping off point.

In choosing the stats to examine, I decided to focus on PER, WORP, VORP, and Box Plus Minus. I chose these particular stats for a number of reasons. Each of those stats is supposed to effectively gauge players against one another--in other words, the higher the number the better the player is considered to be. Also each of these statistics has been used for a number of years, and are therefore at least somewhat familiar to some sports fans. Each of these advanced metrics try to combine more basic statistics by weighting them in a way that would help to statistically characterize the various levels of performance of players.

One metric examined is PER (Player Efficiency Rating),developed by ESPN's John Hollinger [http://www.espn.com/nba/columns/story?columnist=hollinger_john&id=2850240]. It seeks to determine how effective a player is in minutes spent on the court in relation to other players, with league average being 15. One potential issue with this metric is that there is no way to separate players who play well in limited minutes (often older players) from players who apply that same efficiency over a large portion of the game. WORP translates to wins above replacement player. In theory if Player X has a WORP of 3, he should contribute to his team winning 3 more games than if they had a league average player at his position instead of him. VORP, or Value Over Replacement Player, also seeks to compare players to the league average players in a meaningful way. In fact these statistics have some overlap in the way they are calculated. Finally, I chose to examine Box Plus Minus (BPM). Box Plus Minus attempts to quantify a player's contribution per 100 possessions. A box plus minus of +5 means that a team with player X on the floor instead of a league average player, would should score 5 extra points over 100 possessions. More information on BPM can be found here (https://www.basketball-reference.com/about/bpm.html). While WORP and VORP are metrics used in other sports as well (specifically coming from baseball and then being adapted to basketball) Box Plus Minus is a basketball-specific statistic and definitely the one I was most anxious to explore.

Due to the preponderance of sports information on the internet, finding a data set to begin working with was relatively simple. The following spreadsheet found on basketballreference.com contained all of the advanced metrics I had chosen, as well as many other pieces of useful information: BPM Excel Spreadsheet. Once the data was found it was necessary to organize it in a way that would be useful for my specific examination. The first issue became obvious when doing a simple count of team names. I had chosen to explore team data from the period of 2011 to 2015, and in this time several teams had moved or changed names. Specifically, there were issues with the Nets, Hornets and Pelicans. These were resolved by appropriately grouping the new teams back to their original monikers. I then extracted the categories I planned to work with into a new Pandas dataframe. At this point I also chose to make an adjustment to the BPM statistic. Since BPM was a per 100 possession measure, and I thought it was important to look at how many minutes (per season) the player contributed at that level, I created a new stat BPM_A (Box Plus Minus Adjusted) to account for playing time.

The next decision dealt with rookies. College statistics do not translate well into the NBA. In truth, a lifetime of statistical analysis could probably be dedicated determining how college players will perform in the NBA, but that is outside the scope of this project. I chose to address this problem by adding a new entry for each rookie in the year before they entered the league with the same statistics they generated in their rookie year. In this way I made it appear that each rookie basically performed as expected, since this was not a variable that could be addressed in the offseason as part of our business problem anyway. Rookies essentially became a sort of control group within the examination.

Once I was confident that I had workable data I set to work on aggregating it into the form needed. Since I planned to examine changes in WORP, VORP, BPM_A and PER of teams from year to year I aggregated the data by team and year. I then wrote a function that could figure the difference in each stat (or any category) from the previous year and add it as a new column of the dataframe. This was also useful for compiling the change in wins from the previous year (hereafter referred to as Win_Delta.) I chose to explore both the relationship between team totals in WORP, VORP, BPM_A and PER, and team wins as well as the offseason changes in those statistics and how they related to Win_Delta.

I should also note that while it was not part of my initial data wrangling, I later realized that the 2012 season was shortened by labor dispute that led to a player lockout. This led to unusually low win totals in 2012 and unusually high Win_Deltas in 2013. I chose to deal with this by normalizing all the data as if it came from 82 game seasons. Since the metrics being examined were already in either per game, per minute, or per possession form, the only things that needed to be adjusted were Wins and Win_Delta. I found the correct multiplier and used each teams 2012 win percentage to gauge how many wins that would have equated to had 2012 been an 82 game season.

With appropriately clean data on NBA season 2012 through 2015, that now included information on changes from the previous season, it was time to start exploring the information contained within the data set.

Section 2:2 Storytelling and Statistical inference

Once the data was in order, I could try to establish a relationship between player statistical measures and wins in the next season. I chose to initially approach the problem by examining every team's offseason change in each of the metric, and how it related to change in wins. I initially chose to look at changes rather than outright wins because I believed from a business perspective, it might be easier to control change in wins. My initial intuition was that number of wins in the previous year would heavily influence number of wins in the coming year, perhaps even more so than the metrics being examined. The relationship between each statistical metric and change in wins is pictured in the regression plots in Figure 1 below.

A quick look at the scatterplots shows that a there is a clear positive correlation between Win Delta and 3 of the statistics, WORP, VORP, and the Adjusted Box Plus Minus Score. Surprisingly, PER, an established and well-respected metric shows a slightly negative relationship to change in wins amongst NBA teams.

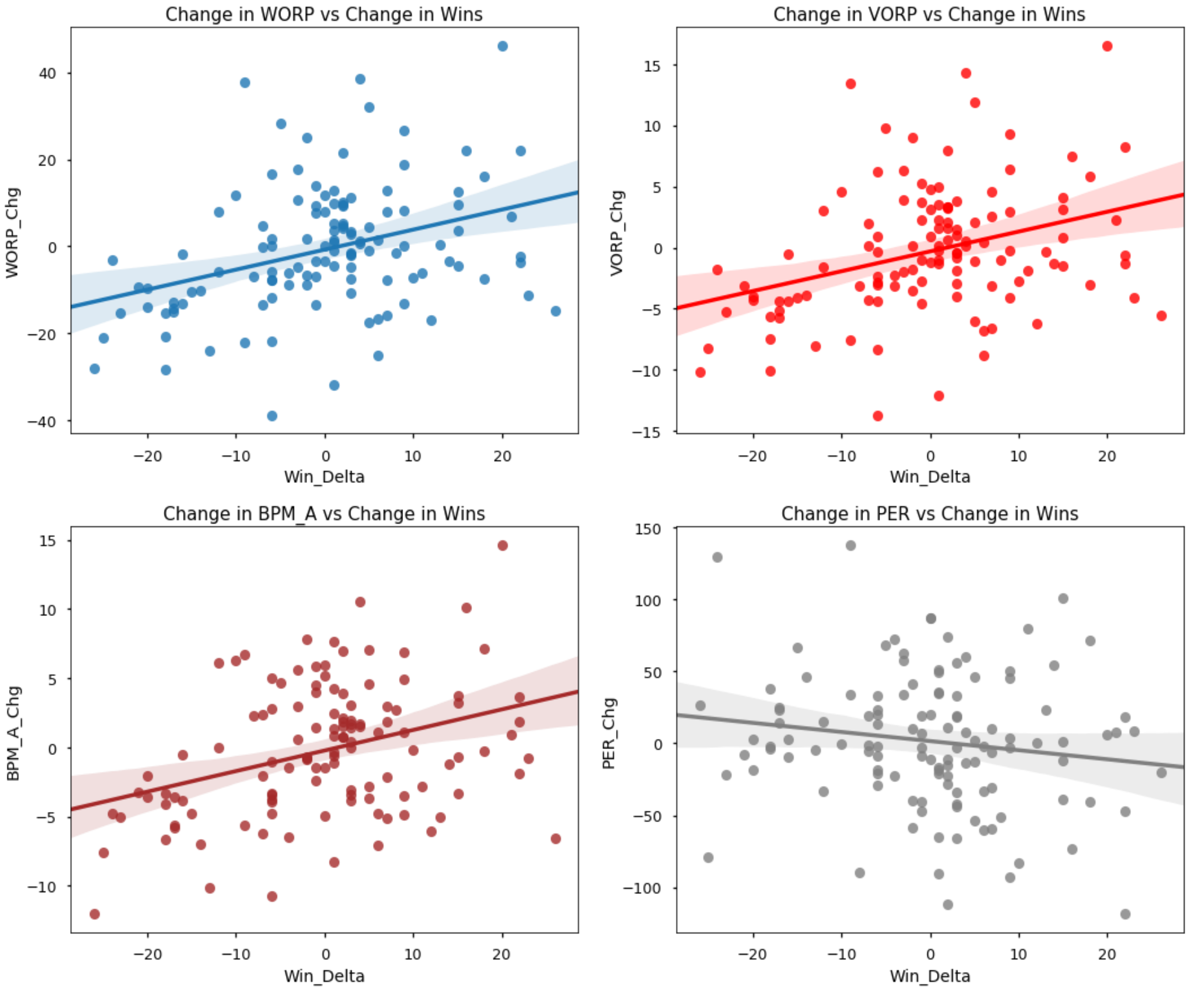


Figure 1

Since I was exploring how statistical factors affected wins in the NBA, I thought it was important to first learn a bit about how wins and change in wins were distributed in the league.

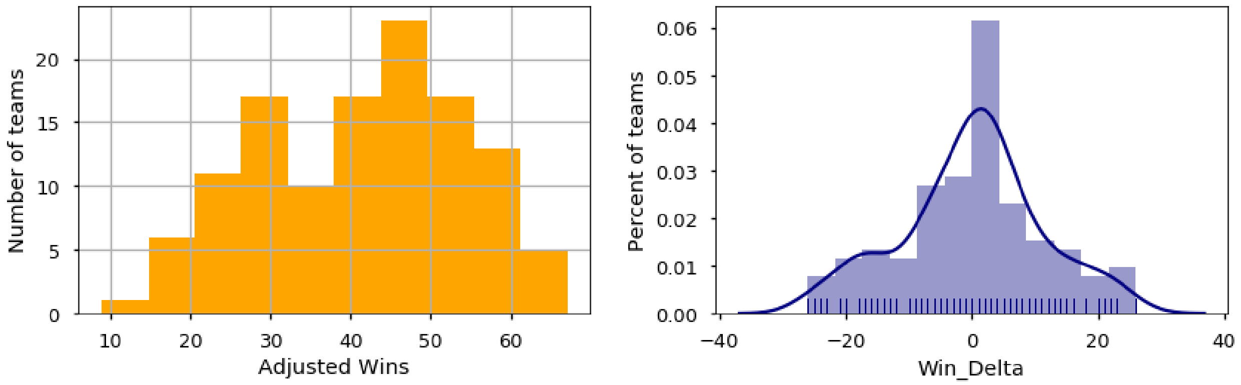


Figure 2

In figure 2, the first graphic shows that most NBA teams win around half (42) of their games. The second graphic indicates that the vast majority of teams win about the same number of games in the next year as they did in the previous year, and that large Win_Deltas are not the norm.

It is also worth noting that perhaps the biggest factor in Win_Delta may have been wins in the previous year. A regression model shows that NBA teams tend to regress back towards the mean of winning approximately half of their games. Teams with low win totals tended to improve, while teams with high win totals had difficulty maintaining that success.

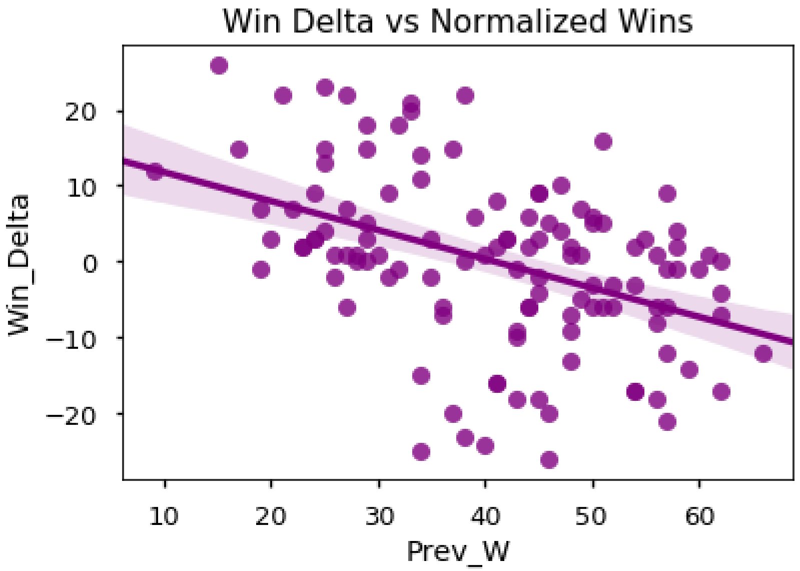


Figure 3

I also explored the relationship between age and the advanced stats in question, in the hopes that perhaps age could be a criteria for team construction, or at least be of predictive use. The results were surprising to me. I expected statistical prowess to peak and then drop off in a bell curves fashion, but this was not the case. Some of the oldest players in the league seemed to be some of the best statistical performers. I then examined the yearly change in those metrics by age, again thinking that I would see a peak and roughly bell shaped curve (Figure 4).

This result was more along the lines of what was expected but still showed some anomalous spikes. Further examination of the data showed that there were so few players at advanced ages that one exceptional performance could really affect the data. This was shown by the age distribution of players in the sample (Figure 5).

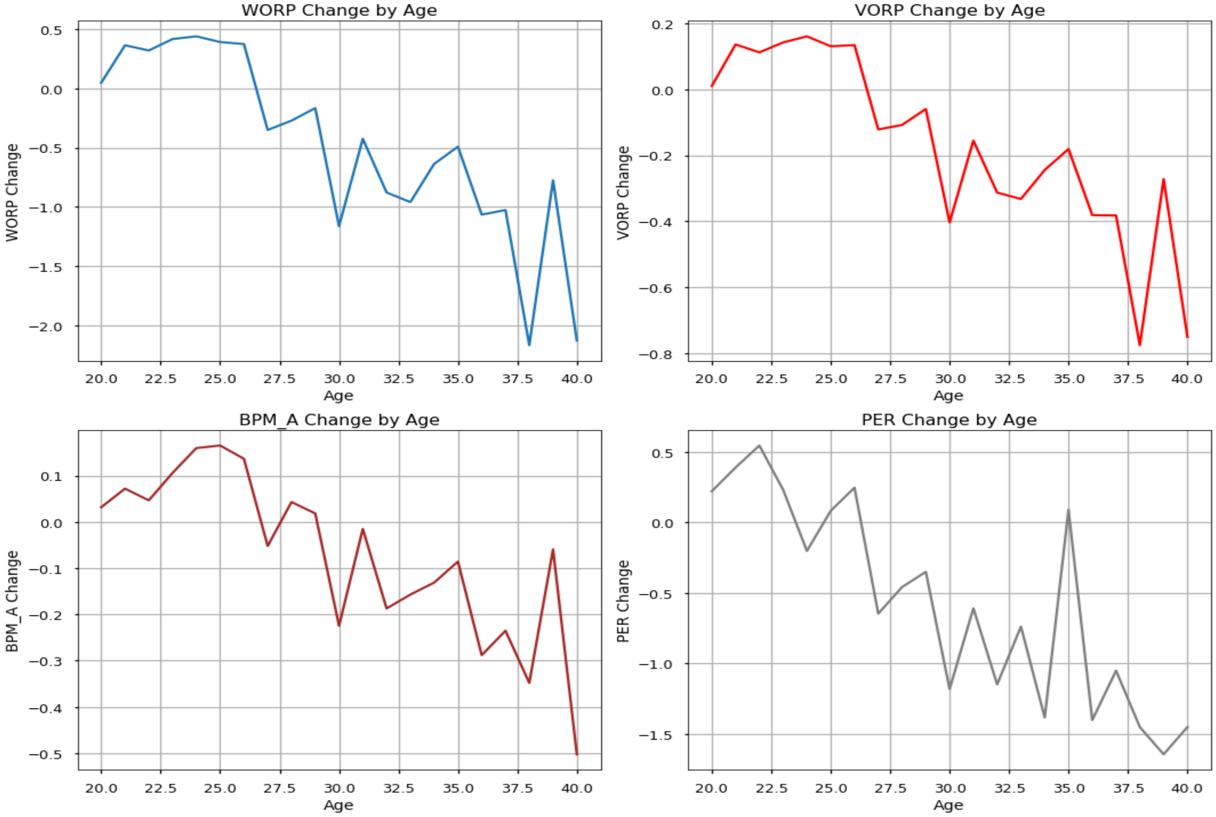


Figure 4

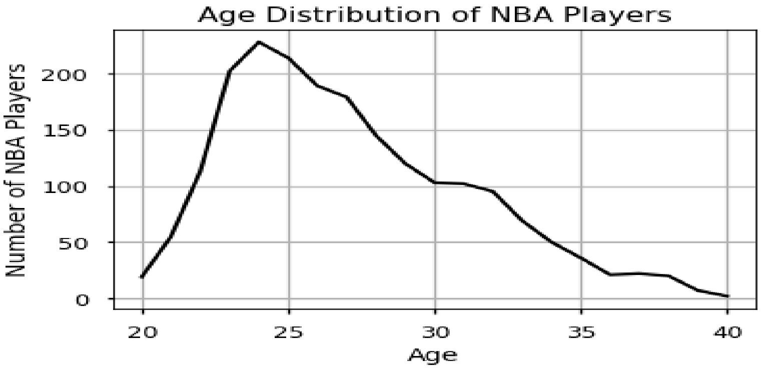


Figure 5

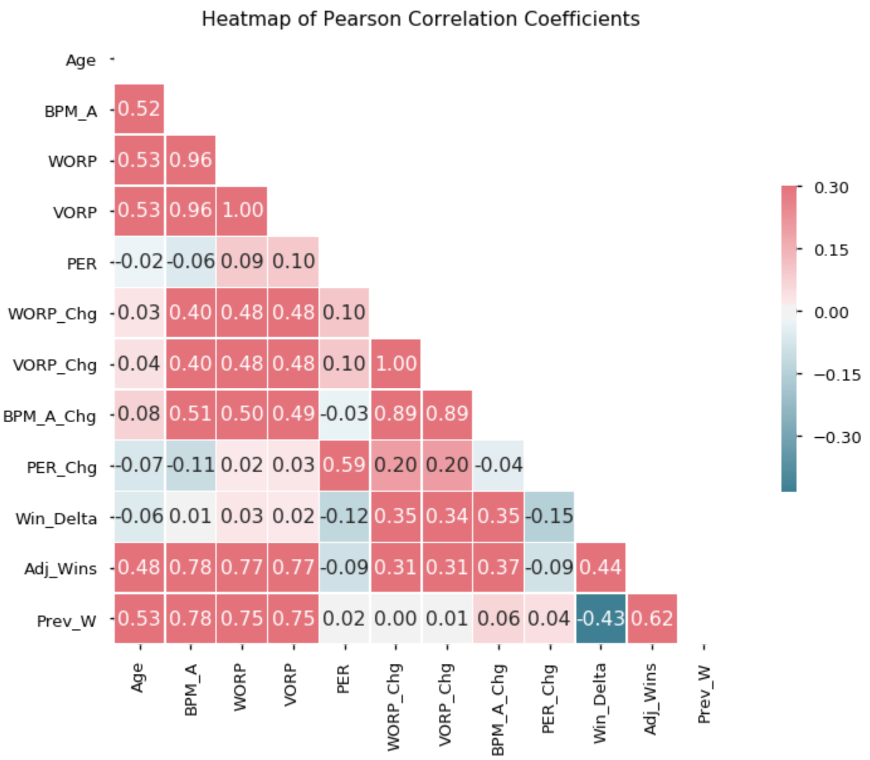
After a visual examination of the data, I was able to dig a little deeper into some of the information gained through that analysis. I previously had discovered through a regression plot that WORP, VORP, and BPM_A appeared to be correlated to Win_Delta. An examination of their Pearson Correlation Coefficients and associated p-values further explored that relationship:

WORP_Pearson = (0.353, 7.283 e-05)
VORP_Pearson = (0.344, 0.00011)
BPM_A_Pearson = (0.352, 8.047 e-05)
PER_Pearson = (-0.151, 0.09930)

I then broke the league into thirds based on wins to see if I could determine statistical significance between these stats and win delta. On initial testing I was able to reject the null hypothesis that WORP did not have a statistically significant relationship to Win_Delta, but tests of the other statistics came in above my α of 0.05, meaning the null hypothesis could not rejected.

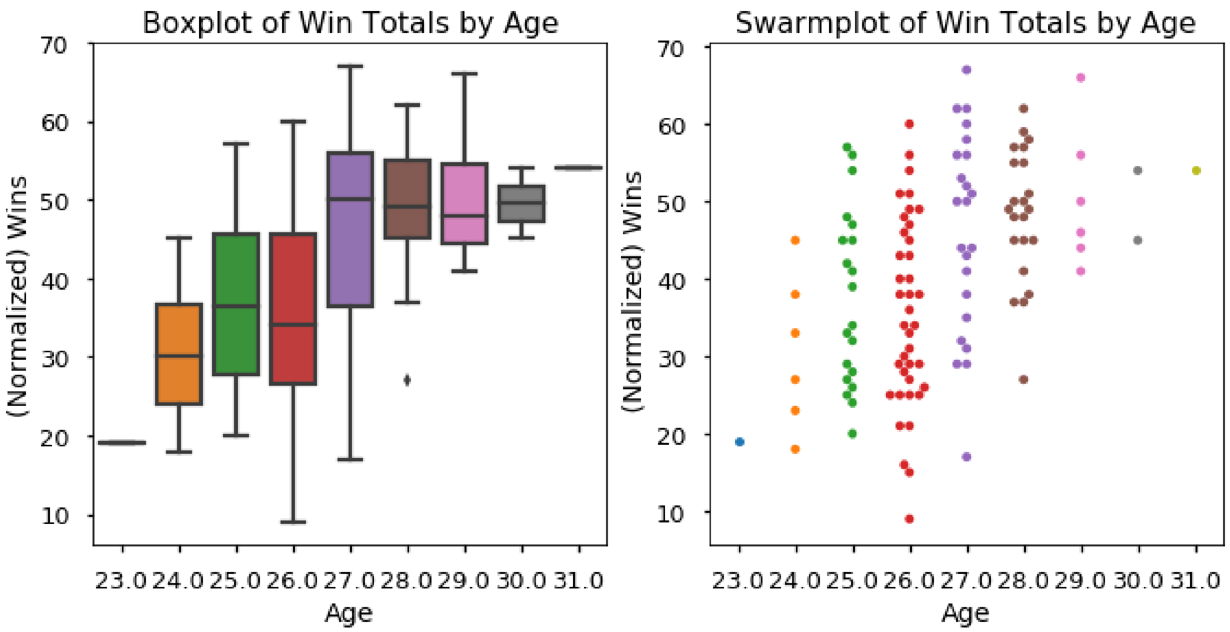
I believed that the result may have come from dividing the league into thirds, leaving me with groups that were too similar to each other. I divided the league into quarters and retried the test. This time I was able to invalidate the null hypothesis for WORP, VORP, and BPM_A. PER once again did not show a statistically significant relationship.

Figure 6



I also created a heatmap (Figure 6) showing the Pearson correlation coefficient of the stats contained in my dataframe to further explore the relationship between each variable. Since average team age and wins appeared to show a positive correlation on the heatmap, I decided to take my exploration of age a bit further. I thought that there might be a peak age that NBA general managers could aim for when building a team, and once again I felt that the distribution of wins by age might be bell curved (Figure 7). I was once again surprised not to see any drop off amongst older teams. I did, however, discover that teams with average ages under 27 tend to lose more games than they win.

Figure 7



Section 3: Machine Learning

3:1 Baseline Modeling

The next phase of exploration was to actually create a working model that could use the metrics I had explored throughout the earlier parts of the project (specifically WORP, VORP, BPM_A, PER, and now Average Age of the team as well as the team's previous season win total). I used the Statsmodel package to build models for both change in the statistics versus Win_Delta and the outright statistics versus team wins.

The model that predicted wins achieved an R-squared score of 0.56 as compared to the Win_Delta which only achieved a 0.39. At this point it was becoming clear that the model used to predict outright wins was having significantly more success. I then scatter plotted each model to see if one was obviously better than the other. Figure 8 below shows a much tighter grouping meaning that the wins model was actually making more accurate predictions than the Win_Delta model. I also ran a leverage plot to look for outliers influencing the result (Figure 9). This is how I discovered the 2012 issue. Item #16 was a Charlotte Team with a very low win total and the #97 was a Knicks team with an unusually high Win_delta. Once I had determined which set of inputs created a more effective model, I wanted to see if by experimenting with different types of models, and also by tuning hyperparameters, I could improve the predictive accuracy of the linear regression

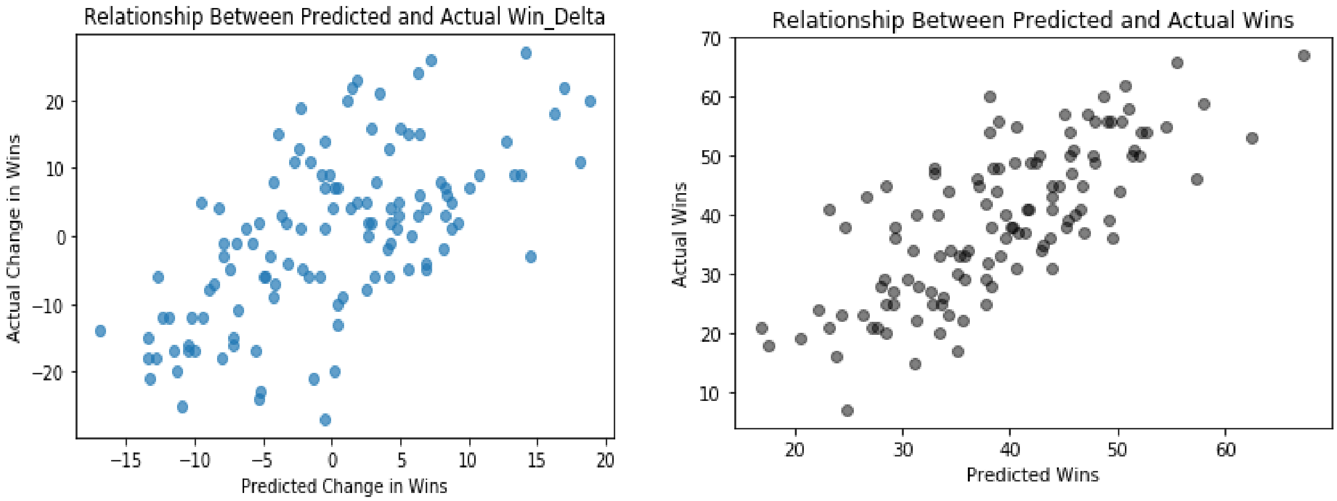


Figure 8

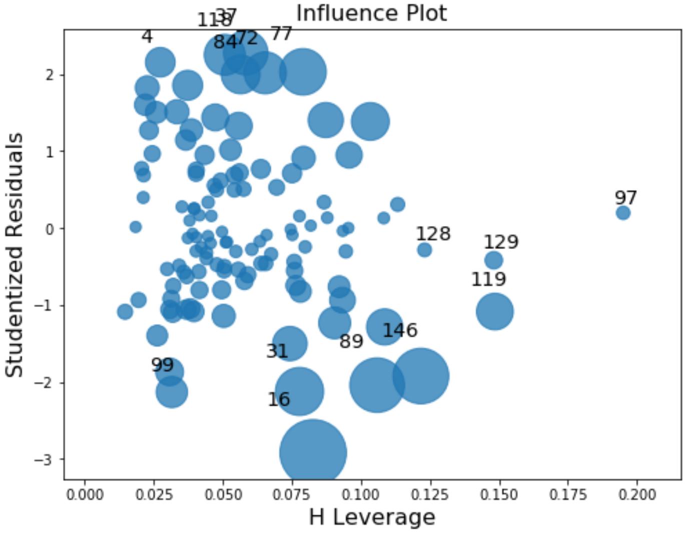


Figure 9

3:2 Extended Analysis

After making the necessary corrections and creating a normalized column for wins, I proceeded searching for the best model. I applied the search to both the wins, and Win_Delta model, and the search initially included Linear Regression, Lasso Regression, Ridge Regression, Linear SVR, Lasso CV and Ridge CV models. The results are below in Figure 10.

Figure 10

Model Performance

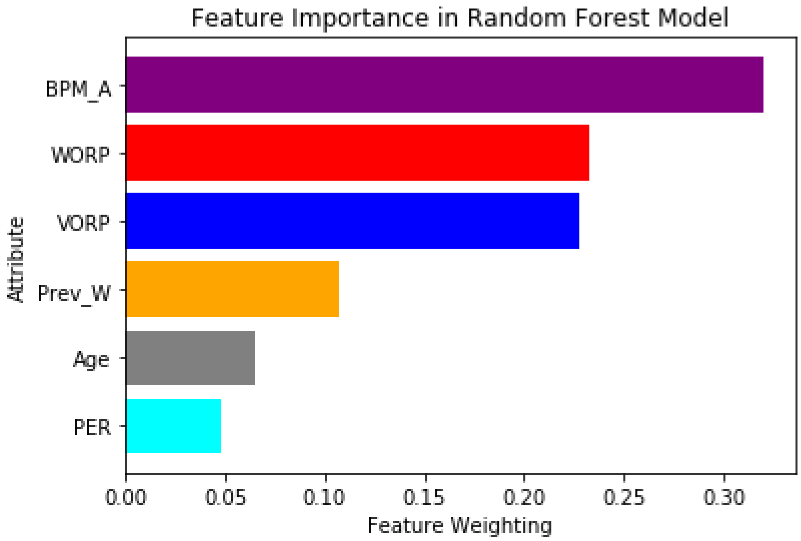
Prediction	*Model Type*	*MSE*	*Variance*	*Lower Bound*	*Upper Bound*	*90% Range Width*
Total Wins	*Linear Regression*	*93.47*	*0.47*	*-11.2*	*18.5*	*29.7*
	Lasso Regression	*87.02*	*0.5*	*-10.2*	*18*	*28.2*
	Ridge Regression	*92.99*	*0.47*	*-11.1*	*18.5*	*29.6*
	Linear SVR	*97.21*	*0.44*	*-6.9*	*21*	*27.9*
	Ridge Regression CV	*86.55*	*0.51*	*-9.7*	*18.2*	*27.9*
	Lasso Regression CV	*90.51*	*0.48*	*-10.4*	*18.3*	*28.7*
	Elastic Net	*88.98*	*0.49*	*-10.4*	*18.2*	*28.6*
	Random Forest	*85.38*	*0.52*	*-11.29*	*15.98*	*27.2*
Win Delta	*Linear Regression*	*110.64*	*0.27*	*-16.9*	*18*	*34.9*
	Lasso Regression	*108.54*	*0.28*	*-16.8*	*18*	*34.8*
	Ridge Regression	*103.54*	*0.31*	*-16.5*	*17.5*	*34*
	Linear SVR	*106.5*	*0.29*	*-16.2*	*17.7*	*33.9*
	Ridge Regression CV	*103.54*	*0.31*	*-16.5*	*17.5*	*34*
	Lasso Regression CV	*105.3*	*0.3*	*-16.7*	*18*	*34.7*

The models had significantly more success predicting wins than predicting change in wins. This was in line with what I had found in baseline modeling, where the model predicting wins had a higher r-squared score and a lower Akaike Information Criterion (AIC) score. Since a predicted change in wins number can be easily calculated by looking at predicted wins and previous season wins, the win model is a clear winner. As the wins model had performed the best in general, I tested 2 more models, Elastic Net and Random Forest on that set of data. While most of the win models were fairly close in terms of r-squared and mean squared error, the best of the group, with a variance score of 0.52 and an mse of 84.74 was the random forest model.

3:3 Feature Importance Analysis

Once the best model had been determined, it was important to understand how the model's conclusion had been reached. By analyzing feature importance we could determine which factors the model in question had used to make predict team wins in the coming season. Since Statsmodel's Random Forest model has a pre-built feature importance function, it was simple to examine (Figure 11).

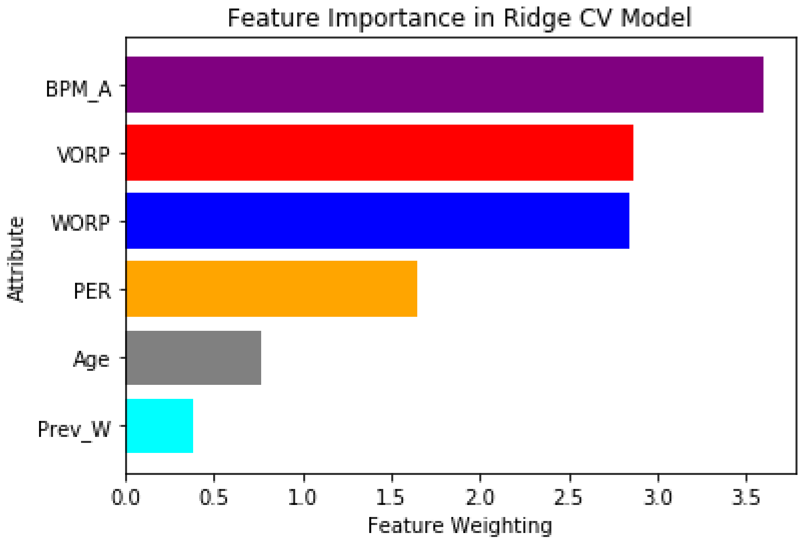
Figure 11



At this point in the process much useful information had been uncovered. First off it was easier to create a model to accurately predict team wins based on statistical metrics than it is to create one that uses changes in those metrics to predict change in wins. Secondly, the non-parametric Random Forest model was the most accurate model tested in terms of both

R-Squared and mean squared error. That model also had the the smallest range of variability from the predicted result (in looking at the 90% result range). In looking at the feature importance data from the random forest model, it is apparent that the most important statistic for the models prediction is Box Plus Minus (in this case adjusted to a per game basis). This comes as no surprise. Box Plus Minus is an established and well respected statistic designed specifically for basketball. WORP and VORP, which are calculated in similar fashion, provided similar levels of importance to the model. PER which I established early on as not showing strong correlation to wins was the least important factor in the random forest model. To satisfy my own intellectual curiosity, I also examined feature importance in the next most accurate model, the Ridge CV model (Figure 12).

Figure 12



In many respects the results were similar. Box Plus Minus was still the greatest determinant of future wins. While WORP and VORP are reversed from the previous model, they still both have similar importance, both to each other and to the model. PER, regardless of its seemingly negative correlation to future wins, was more important in this model, and wins from the previous season had the least effect on the model's outcome.

Section 4: Conclusions and Future Work

4:1 Conclusions

This analysis of NBA advanced metrics sought to examine the viability of those metrics (amongst other factors) as a predictor of future wins. By testing each factor individually I was able to prove that WORP, VORP and Box Plus Minus have statistically significant relationships to team wins. I was able to determine that wins could be predicticted more uniformly than change in wins (Win_Delta), and I was able to use linear regression to build a model that would predict future wins based on those metrics along with previous season wins and average age of the team.

4:2 Future Work

The model created as part of this project is a useful jumping off point, and could certainly provide an insightful basis for player personnel decisions. There are, however, some further explorations that could be done to potentially improve the model, or use the data in a slightly different way.

1. Player Injuries - One of the most unpredictable (and often frustrating) things about sports is player injuries. While I do not think it is possible to predict injuries, I think some exploration of past injuries could be used to improve the model. I think by looking at teams that lost top players for a great number of games and cross referencing them with teams that diverge from the prediction it my be possible to eliminate some of those data points as outliers and improve the model's performance.

2. Rookies - As I mentioned earlier in the paper, predicting rookie performance is a complex topic well outside the scope of the work in this paper. My choice to grade each rookie as though they performed "as expected," could potentially be improved upon though. I believe by examining past rookie performances, expected values could be assigned to rookies by draft position, and these could essentially be used as part of the predictive model.

3. Classification (Logistic Regression) - I believe the linear model was an effective solution to the problem, but logistic regression could have been used to approach the problem in a slightly different way. Instead of looking at a linear model and predicting a number of wins, a logistic regression could have been used to gauge the likelihood of an improved record in the coming season based on the combination of metrics. In addition to being of use to general managers, the logistic regression model could also be of use to gamblers betting the “over/under” on the number of games each team will win.

Section 5: Customer Recommendations

From a business perspective, the initial goal was to utilize the metrics in a way that could be useful to NBA general managers (GMs). The model created here does a number of things to that effect. The model can predict wins based on a combination of metrics. The client, an NBA general manager, could plug in available players’ (or combinations of available players) statistics to calculate wins for the next season before making decisions to sign or not sign those players. In addition, he could use the model to make decisions by comparing predicted wins associated with various combinations of players.

By plugging each one into the model (as part of the greater team) and determining their impact on predicted team wins, a GM could make informed decisions as to whether or not a player would be worth the resources required to sign him. On a more general note, I would recommend to the client, that when building a team, Box Plus Minus should be the statistic used to make decisions, with some emphasis placed on WORP and VORP as well. PER for the most part could be ignored when making personnel decisions.