The NBA is a lucrative business. Revenue generated from the sport has seen continued growth year-after-year due to the star caliber players who take the court every night. However; despite being a part of a billion-dollar business, players, coaches and front office employees of all 30 NBA teams have one goal in mind and that is to win. But to win, general managers need to ensure that the team they have constructed is built to do so.

Predicting NBA Team Performance based on Player Statistics

Neelan Muthurajah - 500795247

Predicting NBA Team Performance based on Player Statistics

Neelan Muthurajah
Student Number: 500795247

The NBA is a lucrative business. Revenue generated from the sport has seen continued growth year-after-year due to the star caliber players who take the court every night. However; despite being a part of a billion-dollar business, players, coaches and front office employees of all 30 NBA teams have one goal in mind and that is to win. But to win, general managers need to ensure that the team they have constructed is built to do so.

Sports statistics has gained momentum in recent years, especially within the NBA. Statisticians have taken typical box score statistics and have compiled advanced formulas to truly capture a player's overall value to their team. These statistics pay tribute to the areas of the game not captured by typical box scores and are therefore valuable when evaluating a team's roster. For example, John Hollinger's PER serves as an all in one basketball rating which attempts to simplify a player's contribution into one number. Others include Box Plus Minus- which evaluates NBA players quality and contribution to their team, Player's Plus Minus to measure a player's impact on a game and win shares to estimate the number of wins a player produces for his team.

For the final capstone project, I have decided to develop a linear regression model that utilizes common/advanced NBA statistics of players to predict the regular season performance of all 30 NBA teams for the 2016-2017 season. It's no secret that players win games. The better your performance as a player, the greater your chances of winning basketball games. My goal is to determine if there is a statistical relationship between a player's overall skill level (predictor variables) and their team's performance (target variable) and to predict/forecast a team's regular season record based on their player's statistics.

To better my understanding of common and advanced NBA statistics, I reviewed various websites online. For example, **Basketball-Reference.com** not only had data readily available since the 1970 NBA season but also provided information on different advanced statistical metrics such as PER, win shares, VORP and BPM.

In addition, I also consulted online research papers to better understand the methodology used in assessing the relationship between player statistics and team performance. I have not been the only individual to ask this question as the papers outlined below have used years of NBA statistical analysis to predict the performance of teams.

One paper that I came across called **"Predicting NBA Games Using Neural Networks"** from the **Journal of Quantitative Analysis in Sports** examined the use of neural networks as tool for predicting the success of basketball teams in the NBA. Statistics for 620 games were collected and used to train a variety of neural networks such as feed forward, radial basis and generalized regression neural networks. The article when on to explain the subset of features most significant to the predictive power of the model ultimately yielding a 6% increase in accuracy in predicting team wins compared to other NBA experts in the league[1].

Another paper I consulted **"Predicting Regular Season Results of NBA Teams Based on Regression Analysis" –** outlined the process in predicting season results of NBA teams based on one advanced NBA metric. However, unlike the previous article mentioned, multiple linear regression analysis was used to predict the performance of NBA teams based on Team PER[2]. In addition, another paper- **"Selection of Significant NBA Statistics for the Prediction of Wins"** used overall common NBA metrics to predict season wins. Six predictors yielded the highest R^2 value on the training dataset: field goal percentage, rebounds per game, turnovers per game, opponents field goal percentage and opponents rebounds per game[3].

Lastly, I considered the reporting guidelines for experimental replications based on the analysis's described above. The paper titled **"Towards Reporting Guidelines for Experimental Replications: A Proposal"** by Jeffrey C. Carver outlines the proposed requirements when replicating a study. Considering I will replicate the approach from the previously mentioned articles, it is important to mention the original studies, provide information about the replication, compare the results I've obtained to the original studies and draw conclusions based on further insights that would not have been evident from either study individually.

The goal of my capstone project is to develop a predictive model. The best methodology for this would be to utilize linear regression to predict a dependent variable based on independent predictor variables. Classification algorithms such as decision trees, k-means clusters or KNN nearest neighbors wouldn't be

[1] Source: www.perducosports.com/media/NBA_Article.pdf.
[2] Source: www.stat.berkeley.edu/~aldous/Research/Ugrad/Stanley_Yang%20_Thesis.pdf.
[3] Source: sites.stat.psu.edu/~nsa1/stat511.perm/Projects/NBAWinsP2001.pdf

the proper statistical technique to use here as the overall question to be answered doesn't require finding similar groups in an unlabeled dataset.

The dataset that I leveraged for my model are player statistics from the 2009 to 2016 NBA season. I have compiled these player statistics from **Basketball-Reference.com**. The website provided csv extracts which allowed for easy data retrieval. I also extracted player plus minus data from **ESPN.com** and appended it to the Basketball-Reference dataset using season and player name for the inner join. The compiled dataset contains the following attributes for each player.

- Season
- Player
- Position
- Age
- Team
- Games Played
- Games Started

Per Game/ Per 36 Minutes/Per 100 Possessions/ Totals Per Season from 2009 to 2016:

- Minutes Played
- Field Goals per game
- Fields Goals Attempted
- Field Goal Percentage
- 3 Pointers Made
- 3 Pointers Attempted
- 3 Point Percentage
- 2 Pointers Made
- 2 Pointers Attempted
- 2 Pointer Percentage
- Effective Field Goal Percentage
- Free Throws Made
- Free Throws Attempted
- Free Throw Percentage
- Offensive Rebounds
- Defensive Rebounds
- Total Rebounds
- Assists
- Steals
- Blocks
- Turnovers
- Points Scored

Advanced Statistics from 2009 to 2016:

- PER- Player Efficiency Rating
- True shooting percentage
- 3 Point Attempt Rate
- Free Throw Attempt Rate
- Offensive Rebound Percentage
- Defensive Rebound Percentage
- Total Rebound Percentage
- Assist Percentage
- Steal Percentage
- Block Percentage
- Turnover Percentage
- Usage Percentage
- Offensive Win Shares
- Defensive Win Shares
- Win Shares
- Win Shares per 48 minutes
- Offensive Box Plus Minus
- Defensive Box Plus Minus
- Box Plus Minus
- Value Over Replacement Player (VORP)
- Player Plus Minus
- Player Plus Minus Per Game
- Player Plus Minus Per 36 minutes

Glossary:

PER -- Player Efficiency Rating
A measure of per-minute production standardized such that the league average is 15.

TS% -- True Shooting Percentage
A measure of shooting efficiency that takes into account 2-point field goals, 3-point field goals, and free throws.

3PAr -- 3-Point Attempt Rate
Percentage of FG Attempts from 3-Point Range

FTr -- Free Throw Attempt Rate
Number of FT Attempts Per FG Attempt

ORB% -- Offensive Rebound Percentage
An estimate of the percentage of available offensive rebounds a player grabbed while he was on the floor.

DRB% -- Defensive Rebound Percentage
An estimate of the percentage of available defensive rebounds a player grabbed while he was on the floor.

TRB% -- Total Rebound Percentage
An estimate of the percentage of available rebounds a player grabbed while he was on the floor.

AST% -- Assist Percentage
An estimate of the percentage of teammate field goals a player assisted while he was on the floor.

STL% -- Steal Percentage
An estimate of the percentage of opponent possessions that end with a steal by the player while he was on the floor.

BLK% -- Block Percentage
An estimate of the percentage of opponent two-point field goal attempts blocked by the player while he was on the floor.

TOV% -- Turnover Percentage
An estimate of turnovers committed per 100 plays.

USG% -- Usage Percentage
An estimate of the percentage of team plays used by a player while he was on the floor.

OWS -- Offensive Win Shares
An estimate of the number of wins contributed by a player due to his offense.

DWS -- Defensive Win Shares
An estimate of the number of wins contributed by a player due to his defense.

WS -- Win Shares
An estimate of the number of wins contributed by a player.

WS/48 -- Win Shares Per 48 Minutes
An estimate of the number of wins contributed by a player per 48 minutes (league average is approximately .100)

OBPM -- Offensive Box Plus/Minus
A box score estimate of the offensive points per 100 possessions a player contributed above a league-average player, translated to an average team.

DBPM -- Defensive Box Plus/Minus
A box score estimate of the defensive points per 100 possessions a player contributed above a league-average player, translated to an average team.

BPM -- Box Plus/Minus
A box score estimate of the points per 100 possessions a player contributed above a league-average player, translated to an average team.

VORP -- Value over Replacement Player
A box score estimate of the points per 100 TEAM possessions that a player contributed above a replacement-level (-2.0) player, translated to an average team and prorated to an 82-game season.

Multiply by 2.70 to convert to wins over replacement.

Player Plus Minus

A metric that looks at how teams perform with a certain player on the court, how they perform with a certain player off the court, and calculates the overall impact that player has on team success.

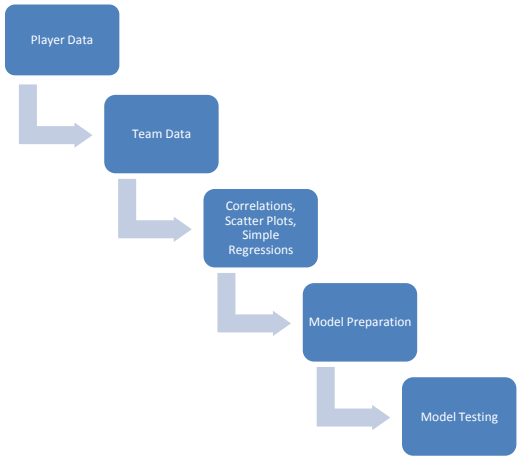
There are 102 different attributes in my dataset, however not all were used in the final dataset to build the regression model. This is because some common NBA statistics were repeated due to different rates. For example, points were recorded for different time periods which included points per game, points per 36 minutes, points per 100 possessions and total points. I decided to only incorporate common NBA statistics per 36 minutes to ensure all common metrics were standardized. Advanced NBA statistics were already standardized considering they are based on either a per minute or per 100 possession rate.

Descriptive Statistics

There are 30 NBA teams, each consisting of around 12-15 players. In terms of stat leaders, Lebron James has played the most minutes in the past 8 seasons and has the highest average PER, BPM, VORP and win shares amongst all NBA players. A possible outlier in the dataset would be the PER rating of Sim Bhullar (43) who had the highest average PER over the past 8 seasons. Since PER is a per minute metric and Sim Bhullar has only played a total of 3 minutes in the NBA, his PER rating is significantly overstated. Therefore, minutes played as a percent of total minutes available to play was used to determine each player's "true" contribution to their team when determining overall team metrics based on players.

Approach

Box-Diagram:



Step 1: Extract Player Data for each Season-Create the Dataset

To predict team performance of all 30 NBA teams based on player statistics, I first needed to extract the necessary independent variables from Basketball-Reference.com. I compiled data from the 2009 to the 2016 NBA season (8 seasons in total) and extracted each player's metrics per season. These metrics included points, rebound, PER and win share to name a few. I also merged data from ESPN.com using season and player name as the inner join. The additional metric from ESPN.com was Player Plus Minus- a sports statistic used to measure a player's impact on the game, represented by the difference between their team's total scoring versus their opponent's when the player is in the game.

Step 2: Extract Teams Win Totals for each Season

Team win ratio (Wins/Total Games Played)) was used as a measure of team performance. Win ratio allowed for a fair comparison between different teams and was the most quantifiable metric for determining team performance. I also utilized NBA statistics of players to determine overall team metrics. For example, PER, a metric that measures a player's effectiveness with a single number, was outlined on a per player basis in the above dataset. To determine team statistics, *I sorted players on the roster of a specific team by minutes they played for that team over the entire season, 2) then multiplied individual statistics by minutes they played and 3) summed the product (player stat * minutes played)*. For example, Lebron James played 70% of available minutes for the Cavaliers during the 2016-2017 season. Therefore 70% of his PER went towards the overall team PER of the Cavaliers. I then summed PER based on these individual weightings of players to get an overall team statistic. A similar process was done with all other metrics.

Step 3: Correlations, Scatter Plots and Simple Regression

Once I calculated overall team statistics based on individual player metrics, I analyzed the relationship between the different variables using correlations, scatter plots and simple regression on a subset of data (Season data from 2009 to 2010). I did this to determine which predictor variables could be used in determining Win Ratio and to identify potential multicollinearity between different predictor variables. **Multicollinearity** refers to independent variables that are related with each other. If included in a model, it can lower the prediction power of that model by increasing the variance of the coefficient estimates and make such estimates sensitive to minor changes. The ideal case is for all independent variables to be correlated with the dependent variable but not with each other.

Step 4: Model Preparation

I used **cross validation** to assess the predictive performance of the model using test data sets and all independent variables. However, some independent variables or sets of independent variables are better at predicting the dependent variable than others. It is also not correct practice to include all independent variables in a model due to potential **overfitting**. Therefore, one crucial step in developing the regression model was **feature selection** (determining the ideal independent variables that explained the variation amongst the dependent variable). In this case determining which team statistic properly justified a team's win ratio. The **"Caret"** package and **"Boruta"** in R was used to determine which independent variables were most significant in the final regression model.

Step 5: Model Testing

Lastly, based on the results of step 4, I used the linear regression model to predict the win ratio of all 30 NBA teams based on the 2016-2017 regular season which served as the out of sample dataset. I then compared these results with the actual win ratio of teams to assess the accuracy of the regression model.

As mentioned in the Literature Review section, I have replicated two previous studies done on this topic. The first report analyzed the relationship between Team PER and Win Percentage. The second report utilized common basketball statistics to again predict Win Percentage. My goal is to further advance these findings by providing a better prediction model for Win Percentage using a combination of advanced and common basketball metrics. The results outlined below reinforce the key findings identified in the mentioned reports, however go one step further in identifying better predictor variables of Win Percentage.

Initial Findings

Correlation matrices are good indicators in identifying independent variables that have a strong relationship with the target variable. Considering the goal is to predict Win Ratio, it is very important to understand the relationships between predictor variables and Win Ratio. I decided to first assess these relationships using a subset of data. **Therefore, the below statistics strictly relate to the 2009-2010 NBA season.** Based on the correlation matrix for the 2009-2010 NBA season, three variables stand out amongst the other predictors. These include PER, BPM and Plus/Minus. The chart below uses Pearson correlation to compute this relationship.

Independent Variable	**Win Ratio**
PER_Advanced	0.84
Plus/Minus_Per36Min	0.94
BPM_Advanced	0.97

***Figure 1: Pearson Correlation between Independent variables and Win Ratio**

There is a very strong relationship amongst these three predictor variables and the target variable-Win Ratio. The summary statistics shown below reinforce this finding.

Team PER

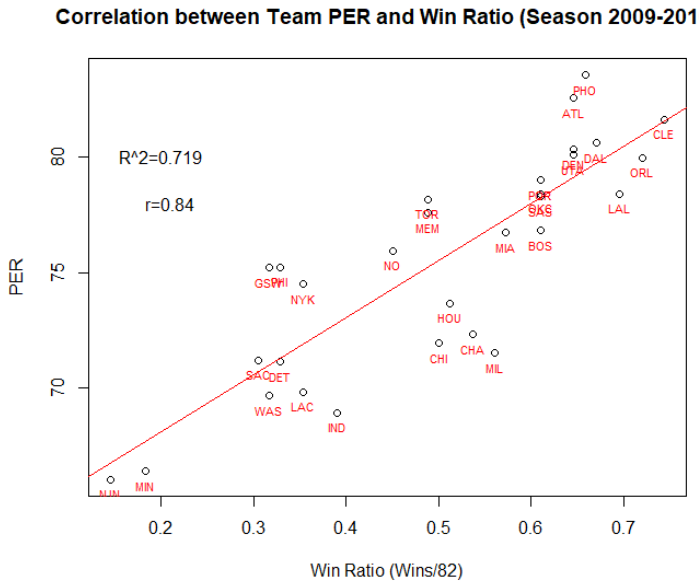


Figure 2: Scatterplot of Team Win Ratio versus Team PER for 2009 to 2010 NBA Season

We can clearly see a linear trend from the scatterplot, which indicate teams with higher PER are more likely to perform better than teams with lower PER. This finding also coincides with the report previously mentioned in the Literature Review section. The results of that report indicated a strong relationship between Team PER and Team Performance which can be observed here as well. Next, we look at the summary statistics of the linear model for additional information.

```
Call:
lm(formula = WinRatio ~ PER_Advanced, data = Season_2009to2010)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17435 -0.06742  0.01120  0.06633  0.17719

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.695919   0.259963  -6.524 4.53e-07 ***
PER_Advanced  0.029077   0.003436   8.463 3.34e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08792 on 28 degrees of freedom
Multiple R-squared:  0.719,     Adjusted R-squared:  0.7089
F-statistic: 71.63 on 1 and 28 DF,  p-value: 3.338e-09
```

Based on the summary statistics, the linear model yields an adjusted R^2 of 0.71. Again, this coincides with the original analysis presented in the Literature Review section of the report which specified an R^2 of 0.71 when comparing Team PER to Team Win Ratio[4]. The significant F-statistic from the Wald test further confirms that the linear model fits the data well and could potentially be used to make future

[4] Source: www.stat.berkeley.edu/~aldous/Research/Ugrad/Stanley_Yang%20_Thesis.pdf.

forecasts. In addition, the linear model is statistically significant as both p-values are below 0.05. Therefore, the null hypothesis can be rejected and we can be certain that the coefficients associated with the variables of the regression equation are not equal to zero.

Similar analysis was done regarding the other two variables mentioned above.

Team Plus/Minus_Per36Min

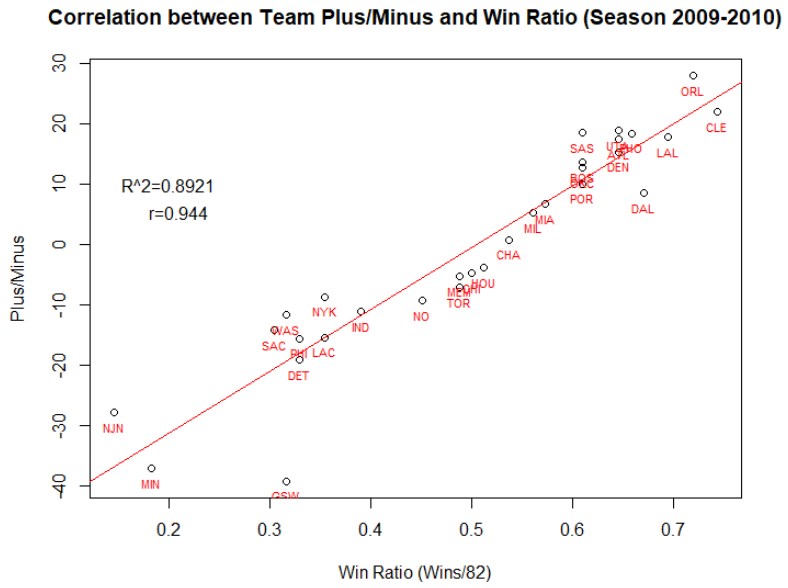


Figure 4: Scatterplot of Team Win Ratio versus Team Plus/Minus for 2009 to 2010 NBA

Like the scatterplot for Team PER, there is again a linear trend between Team Plus/Minus and Win Ratio. Next, we look at statistical summary of the model.

```
Call:
lm(formula = WinRatio ~ `Plus/Minus_Per36Min`, data = Season_2009to2010)

Residuals:
      Min        1Q    Median        3Q       Max
-0.115713 -0.022331 -0.001314  0.028972  0.154564

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.5044041  0.0099490   50.70  < 2e-16 ***
`Plus/Minus_Per36Min` 0.0087167  0.0005728   15.22 4.56e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05447 on 28 degrees of freedom
Multiple R-squared:  0.8921,    Adjusted R-squared:  0.8883
F-statistic: 231.6 on 1 and 28 DF,  p-value: 4.555e-15
```

Based on the summary statistics, the linear model yields an adjusted R^2 of 0.89, making it an even stronger indicator of Win Ratio than Team PER. The significant F-statistic from the Wald test further confirms that the linear model fits the data well and could potentially be used to make future forecasts.

Team BPM

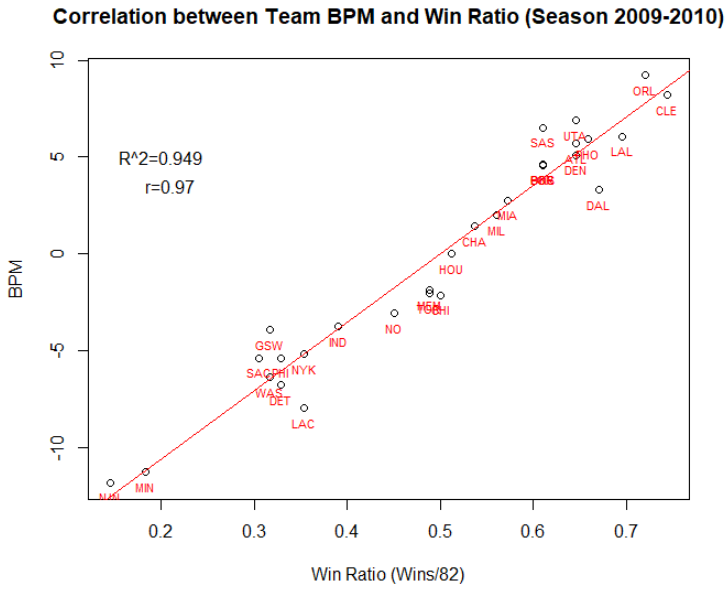


Figure 3: Scatterplot of Team Win Ratio versus Team BPM for 2009 to 2010 NBA Season

Like the scatterplot for Team PER and Team Plus/Minus, there is once again a linear trend between Team BPM and Win Ratio. However, the linearity is much more evident in this plot. Next, we look at statistical summary of the model.

```
Call:
lm(formula = WinRatio ~ BPM_Advanced, data = Season_2009to2010)

Residuals:
      Min        1Q    Median        3Q       Max
-0.077579 -0.013995 -0.004559  0.020544  0.082077

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.500141   0.006841   73.11   <2e-16 ***
BPM_Advanced 0.026820   0.001175   22.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03747 on 28 degrees of freedom
Multiple R-squared:  0.949,     Adjusted R-squared:  0.9471
F-statistic: 520.6 on 1 and 28 DF,  p-value: < 2.2e-16
```

Based on the summary statistics, the linear model yields a strong adjusted R^2 of 0.9471, indicating the model fits the data well. The significant F-statistic from the Wald test (F-Statistic significantly greater than 1) further confirms that the linear model fits the data well and that there is a relationship between BPM and Win Ratio. In addition, the results of this test indicate that Team BPM has a stronger relationship with Win Ratio than either Team PER or Team Plus/Minus.

Clearly, advanced NBA statistics serve as a better indicator of team performance than common NBA statistics. This is because advanced NBA statistics consider many common metrics to fully capture a player's skill level. Although the literature review report indicated a strong relationship between Team PER and Win Ratio, Team BPM and Team Plus/Minus seem to display an even stronger relationship to team performance.

Assessing for Multicollinearity

I used the caret package in identifying highly correlated variables. Using the code snippet below, I identified 13 variables that had a Pearson correlation of 0.75 or higher and removed them from the dataset.

```
correlationMatrix<-cor(TeamStats_Correlation)
highlycorrelated<-findCorrelation(correlationMatrix,cutoff = 0.75,verbose = F)
highlycorrelated<-sort(highlycorrelated)
names(TeamStats_Correlation)[highlycorrelated]
```

The code above indicated that the following variables were correlated with one another.

```
> names(TeamStats_Correlation)[highlycorrelated]
 [1] "FG_Per36Mins"          "FGA_Per36Mins"         "ThreeP_Per36Mins"     "TwoP_Per36Mins"       "FT_Per36Mins"          "DRB_Per36Mins"
 [7] "TRB_Per36Mins"         "PTS_Per36Mins"         "PER_Advanced"         "BPM_Advanced"         "VORP_Advanced"         "Plus/Minus_Per36Min"
[13] "TRBPercent_Advanced"
```

I decided to keep variables not correlated with each other rather than remove all 13. As a result, the following variables were kept in the dataset: FGA_Per36Min, ThreeP_Per36Min, TwoP_Per36Min, TRB_Per36Min, PTS_Per36Min, PER_Advanced, BPM_Advanced and TRBPercent_Advanced. In addition, I decided to remove Win Share variables because Win Share only attributes the number of games that a team had already won to individual players. Therefore, it is incorrect to use the win share of previous seasons to predict next year's result. It is however a good metric to look at when wanting to gauge the actual contribution a player had on their respective team.

Cross Validation

It is important to rigorously test a model's performance as much as possible. Therefore k-fold cross validation works well considering the model is built on different subsets of training data and predicted on the remaining k-1 portion. Again, using the caret package I ran cross validation on the dataset after removing correlated variables. I decided to use cross-validation on season data from 2009 to 2015 instead of including the entire dataset. This is because I wanted to utilize season data from 2016-2017 as the true test set. The results of the cross validation are below.

```
#Feature Importance
train_control<-trainControl(method = "cv",number = 10,savePredictions = TRUE)
model<-train(WinRatio~.,data = TeamStats_Model,trControl=train_control,method="lm")
check<-model$pred
```

```
> head(check)
       pred   obs rowIndex intercept Resample
1 0.5024138 0.537       31      TRUE   Fold01
2 0.6660349 0.695       36      TRUE   Fold01
3 0.3691020 0.366       38      TRUE   Fold01
4 0.5170285 0.500       53      TRUE   Fold01
5 0.5862187 0.585       55      TRUE   Fold01
6 0.3255708 0.293       56      TRUE   Fold01
>
```

Figure 4: Predicted Vs. Observed Results of Cross Validation

```
> print(model)
Linear Regression

210 samples
 27 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 190, 187, 189, 190, 189, 188, ...
Resampling results:

  RMSE        Rsquared   MAE
  0.04251517  0.9308175  0.03418339

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line the data points are. Based on the results of the error estimates both MAE and RMSE are relatively low indicating that the model, with all predictor variables included, does a good job in predicting Win Ratio when comparing actuals to predicted values of the 10 test data sets. A low RMSE/MAE indicates better fit and demonstrates that the model accurately predicts Win Ratio based on test data.

Feature Selection

A fundamental problem of regression modeling is approximating the relationship between an input variable and an output variable based on training data. Most of the time, the output variable is not determined by all variables but by a select few. By including all variables, one major problem may be evoked by the irrelevant features in the learning process. This is known as overfitting. An overfit model can cause regression coefficients, p-values and R-squared values to be misleading. In addition, an overfit model becomes tailored to fit all the random noise in the specific sample data which means an overfit model may not likely fit new data. Therefore, it is highly important to select the correct features for a predictive model without potentially overfitting it.

Using the caret package, I identified a few important variables that explained the variation in Win Ratio. Below is the snippet of code that I used for the feature selection process.

```
> control<-rfeControl(functions=rfFuncs,method="cv",number=10)
> rfe.train<-rfe(TeamStats_Model[,1:27],TeamStats_Model[,28],sizes = c(1,27),rfeControl = control)
> rfe.train

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables    RMSE Rsquared     MAE   RMSESD RsquaredSD    MAESD Selected
         1 0.04595   0.9237 0.03628 0.004573    0.02067 0.004701        *
        27 0.04625   0.9292 0.03567 0.008634    0.01669 0.007865

The top 1 variables (out of 1):
   BPM_Advanced

>
> plot(rfe.train,type=c("g","o"),cex=1.0,col=1:11)
> predictors(rfe.train)
[1] "BPM_Advanced"
```

In addition, I also used the Boruta package to assess important predictor variables.

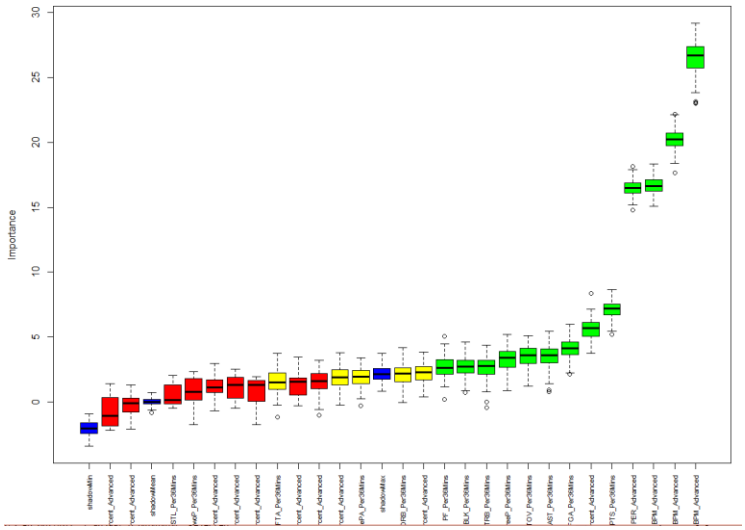


Figure 5: Important Features based on Boruta Package

The results of the Boruta package also showed that BPM_Advanced was the most important variable in the prediction model. However, it did provide 4 other important variables. These included DBPM_Advanced, OBPM_Advanced, PER_Advanced and PTS_Per36Min.

Based on the results of the caret package, BPM_Advanced was the most important variable. Results of the linear regression model using this metric is outlined below.

```
#Based on Feature Selection from Caret
model1<-train(WinRatio~BPM_Advanced,data=TeamStats_Model,trControl=train_control,method="lm")


Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
      Min        1Q    Median        3Q       Max
-0.113704 -0.024410 -0.002456  0.027598  0.093100

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4999812  0.0027331  182.94   <2e-16 ***
BPM_Advanced 0.0263196  0.0004695   56.06   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03961 on 208 degrees of freedom
Multiple R-squared:  0.9379,    Adjusted R-squared:  0.9376
F-statistic:  3143 on 1 and 208 DF,  p-value: < 2.2e-16
```

BPM_Advanced explains much of the variation in Win Ratio. In addition, the model is statistically significant (low p-value and high F-statistic) with an R-squared of 0.9379.

Based on the results of the Boruta package, the most important variables were BPM_Advanced, DBPM_Advanced, OBPM_Advanced, PER_Advanced and PTS_Per36Min. Results of the model using these metrics are outlined below.

```
#Based on Feature Selection from Boruta (Top 5)
model2<-train(WinRatio~BPM_Advanced + DBPM_Advanced + OBPM_Advanced + PER_Advanced + PTS_Per36Mins,data=TeamStats_Model,trControl=train_control,method


Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
      Min        1Q    Median        3Q       Max
-0.106065 -0.023755  0.002596  0.024998  0.091610

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.453000   0.038205  11.857  < 2e-16 ***
BPM_Advanced  -0.015339   0.035833  -0.428 0.669062
DBPM_Advanced  0.040845   0.035868   1.139 0.256139
OBPM_Advanced  0.038489   0.035959   1.070 0.285728
PER_Advanced   0.004574   0.001319   3.467 0.000641 ***
PTS_Per36Mins -0.003973   0.001108  -3.587 0.000419 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03856 on 204 degrees of freedom
Multiple R-squared:  0.9423,    Adjusted R-squared:  0.9409
F-statistic: 666.4 on 5 and 204 DF,  p-value: < 2.2e-16
```

Although this model is also statistically significant, it does not improve the predictive power of the previous model by much. **Therefore BPM_Advanced seems to be the better predictor of Win Ratio and will be used in predicting the 2016-2017 NBA season.**

Box Plus/Minus

Considering Box Plus Minus (BPM) served to be a better predictor of Win Ratio than other variables I decided to further investigate this metric.

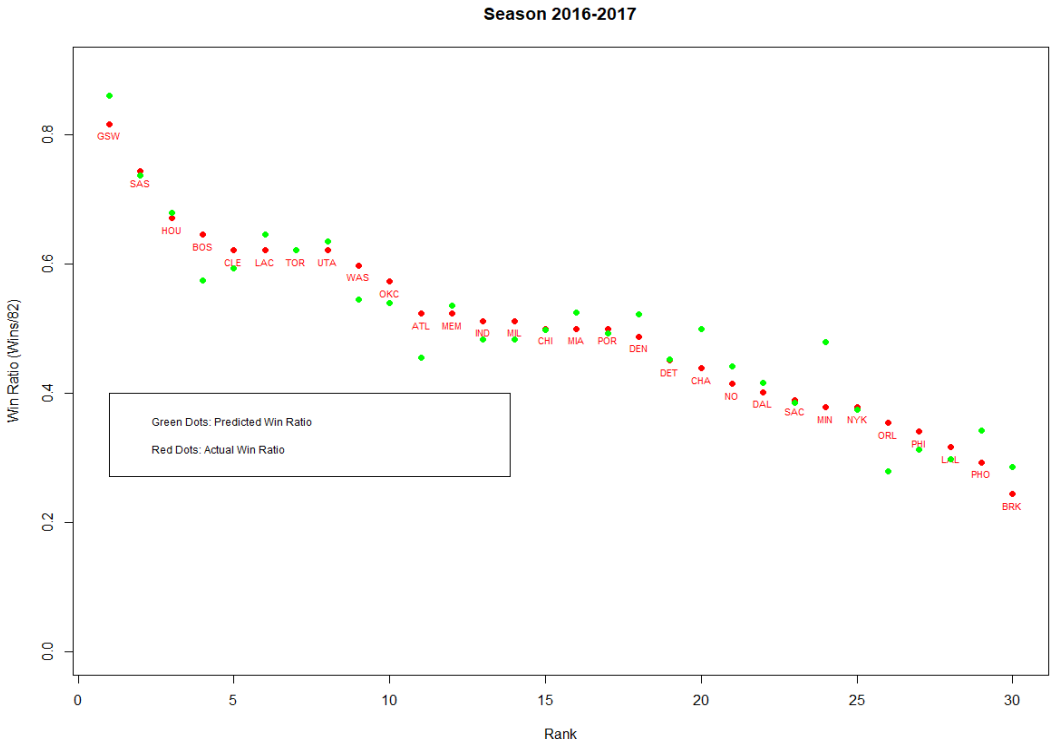
Box Plus/Minus (BPM) is a box score based metric for evaluating basketball players quality and contribution to the team. BPM relies on a player's box score information to estimate a player's performance relative to league average. BPM is a per 100 possession statistic, +5 means a player is 5 points better than average whereas -5 is bad. It considers a variety of common basketball metrics to compute and overall stat for a player, like PER.

The final BPM equation is:

```
Raw BPM = a*ReMPG + b*ORB% + c*DRB% + d*STL% + e*BLK% + f*AST% -
g*USG%*TO% +
h*USG%*(1-TO%)*[2*(TS% - TmTS%) + i*AST% + j*(3PAr - Lg3PAr) -
k] + l*sqrt(AST%*TRB%)
```

Evaluation on Test Data

Although I utilized cross validation to train the predictive model, I decided to leave one season out of the validation to test out of sample data. The graph below illustrates a team's actual win ratio versus the predictive win ratio based on Box Plus/Minus data for the 2016-2017 NBA season.




```
> modelvalues<-data.frame(obs = Season_2016$WinRatio, pred=Season_2016$Predicted_WinRatio)
> defaultSummary(modelvalues)
      RMSE   Rsquared        MAE
0.03940661 0.91513843 0.03041562
```

The red dots in the above graph indicate a team's actual win ratio for the 2016-2017 NBA season whereas the green dots represent the predictive win ratio based on the BPM model. We can see that there is variation regarding actual and predictive win ratio. However, it is still evident that Box Plus/Minus predicted Win Ratio accurately based on the results of RMSE, MAE and R-squared. Like the cross-validation evaluation above, both RMSE and MAE are low indicating that predictive win ratio is close to actual win ratio. Also, the high R-squared value reinforces the fact that Box Plus/Minus explains much of the variation in Win Ratio.

In addition, if we look at the top 8 teams in the league based on actual win ratio, the model correctly identifies the top 8 teams, however in differing order. Therefore, we can be confident that the model accurately predicts win ratio based on BPM of players.

Actual Win Ratio	Predictive Win Ratio
GSW	GSW
SAS	SAS
HOU	HOU
BOS	LAC
CLE	UTA
LAC	TOR
TOR	CLE
UTA	BOS

Figure 6: Top 10 Teams of the 2016-2017 NBA Season based on Actual Win Ratio Vs. Predictive Win Ratio

Conclusions

In conclusion, BPM is a valid predictor of Win Ratio. It was significant in terms of the p-value and F-statistic and provided an R^2 of 0.93. The replicated study shown above yielded similar results to the articles mentioned in the Literature Review section. However, my study proved that there existed an even stronger predictor of Win Ratio than either study explained on their own.

The first study, **"Selection of Significant NBA Statistics for the Prediction of Wins"**, identified that the most important predictors of team wins were common NBA metrics which included field goal percentage, rebounds per game and turnovers per game along with the same statistics for its opponents. Unlike my study, the data collected was from the 1988-1997 NBA season[5]. In addition, information on team opponents were used as well, making this study's approach slightly different than mine. When fit to the training set, the model yielded an R^2 of 0.85, and once fit to the test set the R^2 was even higher at 0.91[6]. However, using feature selection my study indicated that these metrics weren't important compared to the other advanced NBA statistics in my dataset. This showed that advanced NBA metrics may be a better predictor of Win Ratio than common ones. Also, this study lacked the use of advanced NBA statistics. The next study solved for this issue as it used Team PER to predict team performance.

The second study, "**Predicting Regular Season Results of NBA Teams Based on Regression Analysis"**, used one advanced NBA statistic to predict Win Ratio. The methodology was like mine as player statistics were aggregated into overall team metrics to predict Win Ratio. In addition, similar seasons of player data were used to build the model. However, cross validation was used in my study whereas a training/test data split was used in this study. Results of this study indicated a strong relationship

[5] Source: sites.stat.psu.edu/~nsa1/stat511.perm/Projects/NBAWinsP2001.pdf
[6] Source: sites.stat.psu.edu/~nsa1/stat511.perm/Projects/NBAWinsP2001.pdf

between Team PER and team performance[7]. The statistical summary of the model on training data had an R^2 of 0.70 while additionally showing a linear relationship via scatterplots[8]. Consistent results were obtained in my study as indicated above. However, what this study lacked was the use of other advanced metrics such as Plus/Minus, BPM, and VORP.

The main objective of this study was to utilize NBA statistics to identify key variables that could be used in predicting Win Ratio. The previous reports laid out the foundation in terms of the approach but lacked the additional variables available to build a regression model. My study took into consideration a variety of different advanced and common metrics to predict Win Ratio. The results of my study indicated that BPM showed a stronger linear relationship to Win Ratio than either Team PER or common variables. Therefore, Team BPM can be considered a better predictor of Win Ratio and can serve to strengthen this research question. This insight can serve as an extension to the original studies considering it wouldn't be evident from either study individually.

Bibliography

Brownlee, Jason. "How To Estimate Model Accuracy in R Using The Caret Package."Machine Learning Mastery, 21 Sept. 2016, machinelearningmastery.com/how-to-estimate-model-accuracy-in-r-using-the-caret-package/.

Brownlee, Jason. "Feature Selection with the Caret R Package." Machine Learning Mastery, 21 Sept. 2016, machinelearningmastery.com/feature-selection-with-the-caret-r-package/.

Carver, Jeffrey. Towards Reporting Guidelines for Experimental Replications: A Proposal.

Gupta, Dishashree, and Kunal Jain. "How to Perform Feature Selection (Pick Imp. Variables) - Boruta in R?" Analytics Vidhya, 25 Mar. 2016, www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/.

Loeffelholz, Bernard , et al. Predicting NBA Games Using Neural Networks. www.perducosports.com/media/NBA_Article.pdf.

Wang, Stanley. Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics. www.stat.berkeley.edu/~aldous/Research/Ugrad/Stanley_Yang%20_Thesis.pdf.

Walters, Eli, and Brian Staudenmeyer. Selection of Significant NBA Statistics for the Prediction of Wins. sites.stat.psu.edu/~nsa1/stat511.perm/Projects/NBAWinsP2001.pdf.

[7] Source: www.stat.berkeley.edu/~aldous/Research/Ugrad/Stanley_Yang%20_Thesis.pdf.
[8]Source: www.stat.berkeley.edu/~aldous/Research/Ugrad/Stanley_Yang%20_Thesis.pdf.