What Makes a "Good" Basketball Team?

Joshua Moss
Jiaming Shen
George Zhou

Introduction

Sports analytics is a growing area within the field of Big Data. Recently, sports teams have been hiring data analysts to help coaches optimize their winning strategies. Hence, we decided to apply our data mining knowledge to basketball. The goal of our project is to determine the most important factors that contribute to a basketball team's win percentage. We would also like to determine the optimal method to predict the win percentage of a team based on their in-game statistics. Our data comes from the NBA's official website, and it contains data on every team from the 1996-97 season to the 2016-17 season. In total, our compiled dataset had 22 predictors and 652 observations. Below are the names and definitions of the starting variables. All of the variables except percentage variables are in units of average per game.

`WinPercentage` - Number of Wins Over Number of Losses Times 100.
`MIN` - Minutes Played
`FGM` - Field Goals Made
`FGA` - Field Goals Attempted
`FGPercent` - Field Goal Percentage
`ThreePtMade` - 3 Point Field Goals Made
`ThreePtA` - 3 Point Field Goals Attempted
`ThreePtPercent` - 3 Point Field Goals Percentage
`FTM` - Free Throws Made
`FTA` – Free Throws Attempted
`FTPercent` – Free Throw Percentage
`OREB` – Offensive Rebounds
`DREB` – Defensive Rebounds
`REB` – Rebounds
`AST` – Assists
`TOV` – Turnovers
`STL` – Steals
`BLK` – Blocks
`BLKA` – Blocked Field Goal Attempts
`PF` – Personal Fouls Committed
`PFD` – Personal Fouls Drawn
`PTS` – Points

In order to turn this into a classification problem, we also created a new variable called GoodTeam. We assigned it a value of "Y" if the team's `WinPercentage` exceeded 50% and "N" otherwise. This GoodTeam variable will be the response variable that we attempt to predict.

About 80% of the data or 521 observations were randomly chosen as training data, while the rest were used as testing data.

Initial Challenges and Limitations

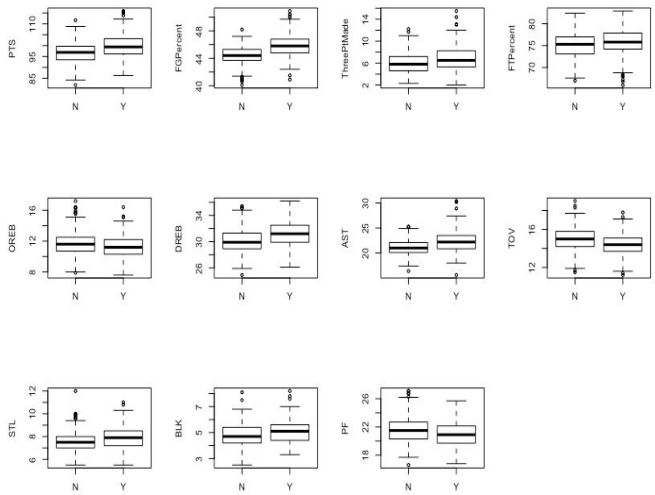
Before we could conduct a data analysis, there were problems with our dataset that had to be taken into account.

1. *Multicollinearity*. Some of the variables were perfect linear combinations of one another. For example, `FG% = FGA/FGM`. Therefore, we had to remove some of the variables to meet the assumptions of logistic regression. Out of our original predictors, we had to remove `FGA, FGM, ThreePtA, ThreePtMade. FTA, FTM, BLKA`.
2. *Moderately Correlated Predictors*. Ideally, when performing classification methods, we would want complete independence for our final predictors. In our project, however, there exists some predictors that were moderately correlated with one another. For example, a 3-Pointer in basketball is defined as a type of a field goal. This means a team's field goal percentage is inevitably influenced by the team's 3-point percentage. Despite this problem, we decided to not remove such predictors since the correlation coefficient between any predictor did not exceed .5. Removing more predictors would further A team's 3-Point Percentage can differentiate themselves from other NBA teams, define their overall playstyle, and better predict the classification.
3. *Relevance*. Not all of the predictors were actually relevant to the classification problem. Minutes played (MIN) is an exogenous variable and thus are roughly constant for each team. Therefore, we removed it before beginning the analysis.
4. *Changes Over Time.* Between 1996 and 2017, the NBA modified the way some statistics were calculated. For example, before 2005, the average PFD statistic was only 0.1, but now it is over 25.0. Therefore, for consistency, we had to remove this variable from the analysis. Also, NBA game strategies have changed over time: teams are now less focused on defense (blocks and steals) and attempt more 3-point field goals.

These limitations led us to keep 11 predictors prior to fitting models: `PTS, FGPercent, ThreePtPercent, FTPercent, OREB, DREB, AST, TOV, STL, BLK, and PF`.

Data Visualization

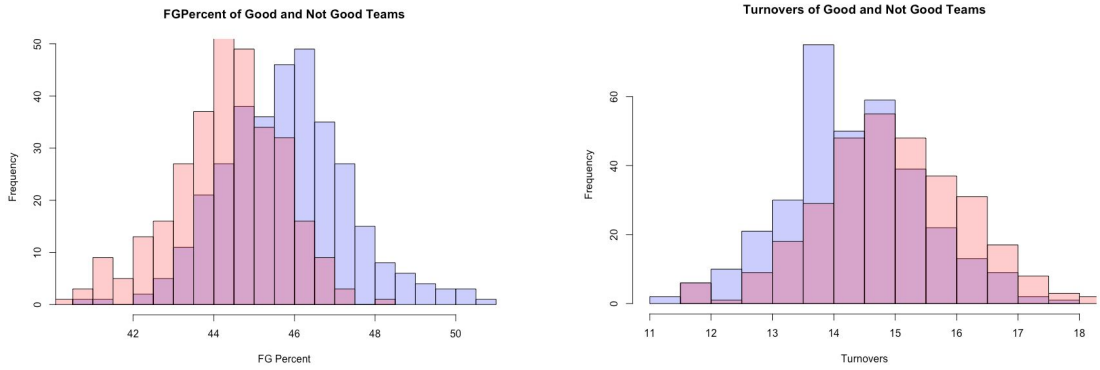
Box-plots of our 11 Predictors



The diagram to the left shows boxplots of the 11 predictors after removing the problematic variables. When examining the diagram, we notice very slight differences between the boxplots of the classes for certain predictors. No boxplot, however, substantially differs between each class. However, we notice that Good Teams usually have a higher field goal percentage than Not Good Teams, and Not Good

Teams have higher number turnovers than Good Teams.

When comparing the overlapping histograms below, we can see the Good Teams have a higher frequency of teams with high Field Goal Percentages while Not Good Teams have a higher frequency of teams with high turnovers.

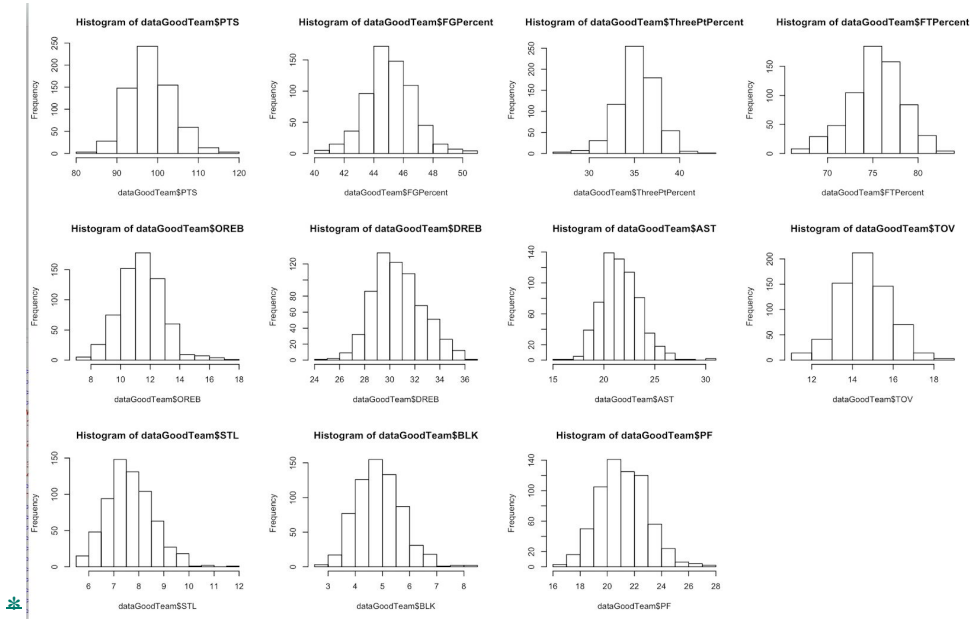


Overlapping Histograms of Field Goal Percentages and Turnovers for Good and Not Good Teams

Blue - GoodTeam = 'Y'

Red - GoodTeam = 'N'

Next in our data exploration, we analyzed the distribution of the data with histograms help us decide which methods would best classify our response.

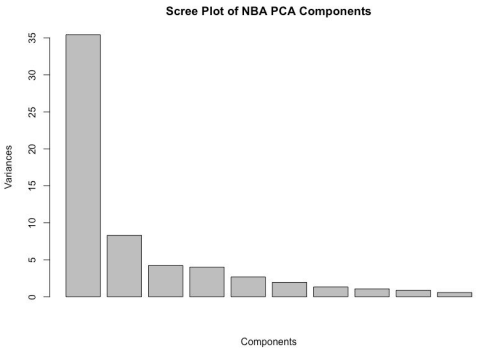


Histograms for Predictors

As seen in the diagram above, our predictors are all roughly normally distributed. With this in mind, our team was able to introduce methods that work well with the normal assumption, such as LDA and QDA classification.

```
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     5.9525 2.8820 2.05562 1.99898 1.6363 1.39649 1.15411 1.03279 0.93844 0.75699 0.64456
Proportion of Variance 0.5822 0.1365 0.06944 0.06566 0.0440 0.03205 0.02189 0.01753 0.01447 0.00942 0.00683
Cumulative Proportion  0.5822 0.7187 0.78816 0.85382 0.8978 0.92987 0.95176 0.96929 0.98376 0.99317 1.00000
```

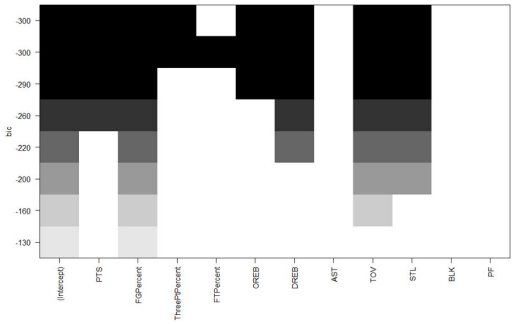
```
                                PC1         PC2
dataGoodTeam$PTS          -0.889361708 -0.24768137
dataGoodTeam$FGPercent    -0.190606718 -0.03496720
dataGoodTeam$ThreePtPercent -0.188415641  0.15691294
dataGoodTeam$FTPercent    -0.206230944  0.91993849
dataGoodTeam$OREB          0.104648842 -0.17862362
dataGoodTeam$DREB         -0.205584937 -0.03826933
dataGoodTeam$AST          -0.174159385 -0.08381255
dataGoodTeam$TOV           0.060118016 -0.12196514
dataGoodTeam$STL          -0.003998103 -0.08027830
dataGoodTeam$BLK          -0.002094913 -0.04567468
dataGoodTeam$PF            0.086548190 -0.05216293
```



In addition, we used principal components analysis (PCA) to grant us insight on the variability within our data. The first principal component is very highly correlated with the average number of points scored. It follows then that variables that statistics related to scoring (such as field goal and free throw percentage) are also moderately correlated with this PC.

Variable Selection

Although we were able to cut down variables from 22 to 11, we conducted feature selection in order to avoid overfitting and build a model that was more easily interpretable. We used BIC as the model selection criterion.

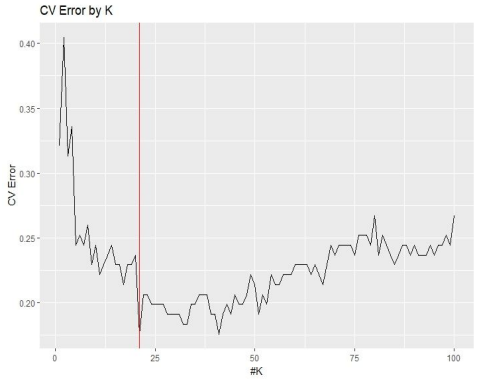


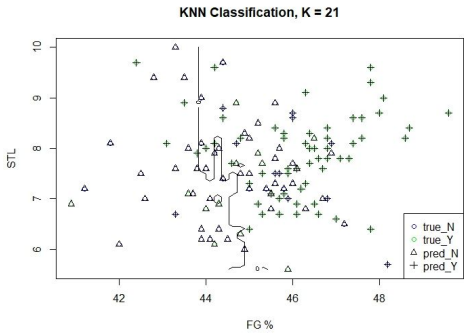
The model with the lowest BIC value contained the predictors `PTS, FGPercent, ThreePtPercent, OREB, DREB, TOV, STL`. Variables `FTPercent, AST, BLK, PF` were removed.

Classification Methods

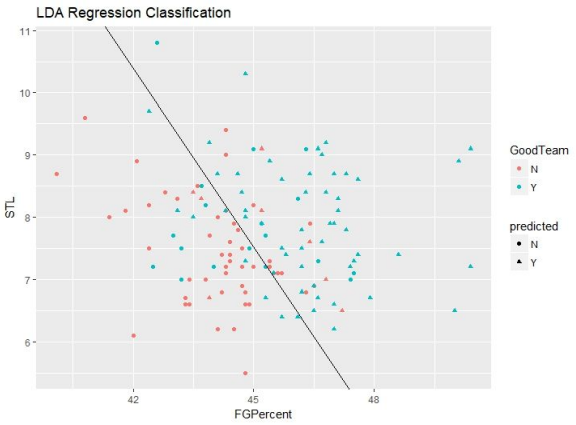
1. *KNN* - The first method we employed was k-nearest neighbors classification. Because this algorithm does not assume a particular functional form or distribution of the predictors, it is an

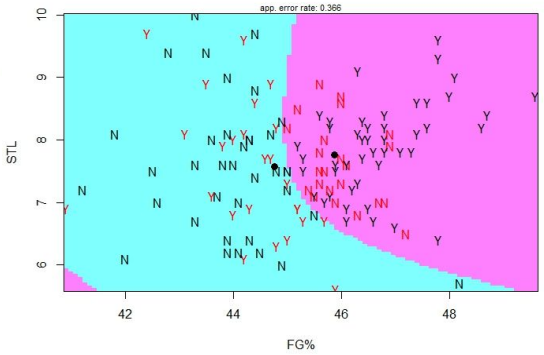
apt method to get an initial feel for the data. First, we attempted the algorithm with all of the original 11 predictors, selecting the optimal K where test error achieves a minimum.
This yields a K value of 21, which is quite a flexible model relative to the other values of K we investigated. This achieves a test error of 17.56%, which is about 23 of 131 observations. After feature selection, we reattempted the KNN algorithm with the BIC-selected predictors since KNN tends to perform better in lower dimensions. The minimum test error was 19.38% achieved at K=38, which is a slightly more biased model. It appears that by removing some variables, we lost too much information and thus were not able to improve upon the full model attempt.





2. *Linear and Quadratic Discriminant Analysis*
Based on our histogram of the predictors in our data exploration, the NBA data appeared to satisfy the normal assumption. Therefore, LDA and QDA seemed like a promising way to classify our data. The classes are evenly distributed (50% are "Y" teams and 50% are "N")





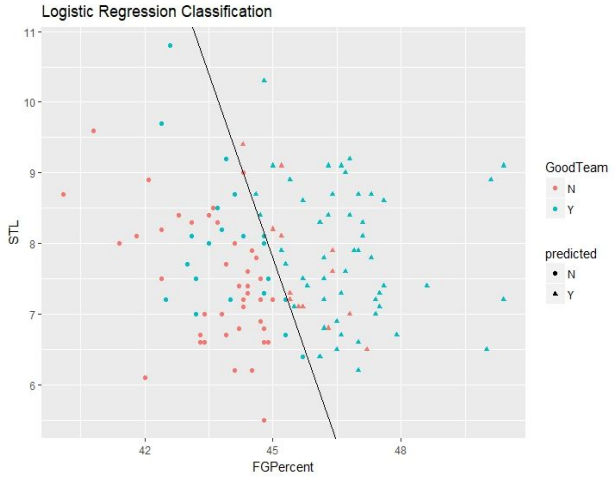
3. *Logistic Regression*

Based on the performance of LDA, it appeared that the class-separating boundary was linear. Therefore, we decided to fit a logistic regression using the BIC-selected predictors. Hence, the model is

$$log\ \frac{P(GoodTeam='Y'|X)}{1-P(GoodTeam='Y'|X)} = \beta_0 + \beta_1 PTS + \beta_2 FGPercent + \beta_3 ThreePtPercent + \beta_4 OREB + \beta_5 DREB + \beta_6 TOV + \beta_7 STL$$

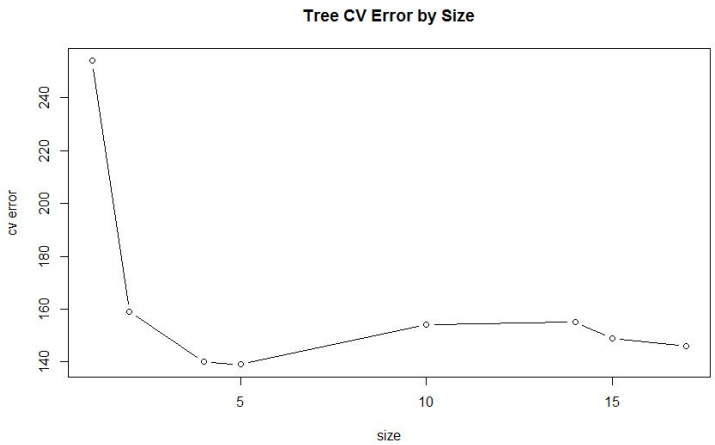
(Intercept)	PTS	FGPercent	ThreePtPercent	OREB	DREB	TOV	STL
-74.3622099	-0.3112669	1.3376897	0.3520576	0.6717814	0.8803200	-0.9849786	1.5399331

One of the benefits of logistic regression is that the coefficients are directly interpretable. For example, a one percentage point increase in FGPercent increases the odds of being a classified as a good team by a factor of $e^{1.338} = 3.810$. An increase in average turnovers per game by one cuts down the odds of being classified as a good team by $e^{-.9849} = .373$. Interestingly, the coefficient attached to `PTS` is negative, which implies counterintuitively that lower scoring teams win more. Below is the boundary suggested by the model.



4. *Trees*

Using the tree method with pruning provides a graph that explicitly shows which variables are the most important. We used cross-validation to prune the original tree. Five branches gave the lowest cross validation error. In the tree graph, if a team has a field goal percentage greater than 45.65%, they are considered a good team. If the team has a lower field goal percentage, then it depends on other factors, such as steals and turnovers. The most important predictors (from most important to least important) are `FGPercent`, `STL`, and `TOV`. The testing error was 26.0%, so relatively it



is not a reliable predicting method (but easily interpretable).

Interpretation of Results and Conclusion

Method	Test Error
KNN w/ 11 predictors	17.6%
KNN w/ 7 predictors	19.4%
LDA	16.8%
QDA	19.8%
Logistic	20.6%
Pruned Tree	26.0%

The results of our tests helped us answer the two central questions of our project and more. Based on the high performance of LDA, the normal assumption and equal class-specific covariance assumption holds up for our data. Our findings that LDA performed the best also suggests that the decision boundary between the two classes is linear. Thus we would recommend owners, coaches, and anyone who loves the sport of basketball to use the relative simple LDA method to model their favorite team's data to predict the team's final performance.

Furthermore, our team also answered what the most important variables are in making a "Good" NBA team. From our findings, Field Goal Percentage (`FGPercent`) and Steals (`STL`) are the most important statistics in determining whether a team has an above 50% win rate.This is a reasonable and intuitive conclusion. If a team has a higher Field Goal Percent, a team makes a higher percentage of their shots which means they are more likely to score more points in a game. Many people would say that "defense wins games", and this is supported by this investigation in the form of steals. If a team has a higher rate of steals in a game, they are more likely to cause the opposing team to lose possession of the ball, giving themselves more chance to score points, while preventing the opposing team from scoring.

Looking Forward

If we spent more time with NBA basketball data, we would like to investigate some of the oddities that occurred during our research. For instance, we would like to examine why the coefficient attached to the `PTS` variable in logistic regression was negative. We hypothesize that this could be due to omitted variable bias. A possible variable that could grant us insight into this issue would be "opponent points scored". It could be the case that lower scoring teams who were classified as "good" teams were also holding opponents to very low point totals. If this hypothesis is true, it would be more intuitive that teams who score lower points are more likely

to be “good” teams. Additionally, it would give more credence to the saying that “defense wins games”.

Individual Contributions

The project was well-distributed among the three members. Josh focused on coding KNN, LDA, QDA, and Logistic Regression. George was in charge of cleaning the data and implementing the variable selection methods as well as other classification methods. Jiaming directed his efforts into the data visualization/exploration and the in-class presentation. The rest of the project, including interpretations, decoration, and other activities, were thoroughly discussed and completed together as a group.

Appendix: Variable Definitions

`WinPercentage` - Number of wins over number of losses times 100.
`MIN` - Minutes played
`FGM` - Field Goals Made, Number of shots actually made by a team
`FGA` - Field Goals Attempted, Number of shots taken by a team
`FGPercent` - Field Goal Percentage, Ratio of a team’s shots made divided by shots attempted
`ThreePtMade` - 3 Point Field Goals Made, Number of three point shots actually made by a team
`ThreePtA` - 3 Point Field Goals Attempted, Number of three point shots taken by a team
`ThreePtPercent` - 3 Point Field Goals Percentage, Ratio of a team’s 3 point shots made divided by shots attempted
`Free Throws (not a variable)` - an uncontested shot at a basket (worth one point) awarded to a player following a foul or other infringement.
`FTM` - Free Throws Made
`FTA` – Free Throws Attempted
`FTPercent` – Free Throw Percentage, Ratio of a team’s free throws made divided by free throws attempted
`REB` - a statistic awarded to team when their player retrieves the ball after a missed field goal or free throw
`OREB` – A rebound collected when the player is on the offense (trying to score the ball)
`DREB` – A rebound collected when the player is on the defense (trying to prevent the opposing team to score the ball)
`AST` – Assists, a statistic awarded to a player who passes the ball to a teammate in a way that leads to a scoring a shot
`TOV` – Turnovers, when a team loses possession of the ball to the opposing team before a player takes a shot at his team's basket.
`STL` – Steals, when a defensive player legally causes a turnover by his positive, aggressive action(s)
`BLK` – Blocks, statistic awarded to a defensive player who deflects an offensive player's field goal attempt, preventing it from going in the basket
`BLKA` – Blocked field goal attempts, Number of blocks attempted

`PF` – Personal fouls committed, when a player commits a violation that prevents the opposing player from moving, scoring, or performing another allowed activity
`PFD` – Personal Fouls Drawn, number of personal fouls committed on a team by opposing teams
`PTS` – Points, Number of points scored by a team

Works Cited

Our Data:
https://stats.nba.com/teams/traditional/?sort=W_PCT&dir=-1&Season=2017-18&SeasonType=Regular%20Season
Basketball Terms and Definitions:
https://www.sportingcharts.com/dictionary/nba/