Machine Learning Examination of NBA Defense

Alex Block

Advisors: Chris Fernandes and Nick Webb

June 12, 2014

Abstract

The project described in this paper will explain a machine learning approach for identifying the characteristic attributes of defensive effectiveness in the NBA. The means of numerically defining defensive effectiveness is a statistic known as "Defensive Efficiency," which is a measure of points scored by the opponent normalized by 100 possessions. Attribute sets of decreasing size were used to predict "Defensive Efficiency" to discover the factors that are unique to good and bad defenses in the NBA. The conclustions outline strategies and areas to focus on to optimize a team's defense.

Contents

List of Figures

1 Introduction

This project attempted to determine the attributes that are most characteristic of defensive effectiveness in the NBA. Each of these attributes is represented by a statistic, which is broken down into its smallest available form. This will allow for the analysis to discover the attributes that are direct causes of a good or bad defense rather than attributes that are merely indicators of defensive success.

The project uses a machine learning approach to model the data as a means of completing this task. A standard model typically requires a calculated correlation that is greater than 0.7 to be deemed statistically significant. For a model to be considered for further analysis, it must display that a significant correlation exists between the attributes and predicted value. The models in this project all perform above this benchmark.

The primary goal is to determine the most influential attributes, but this process has the potential to reveal numerous other truths as well. However, the output and design of the models merely suggest possible conclusions, but in no situation do they provide definitive proof. Each assumption must be analyzed in consideration with preexisting basketball knowledge to determine its validity.

This report is organized in a similar format to the path of discovery. By following the process outlined in this project, the reader should be able to recreate the project with the same results.

2 Background and Related Work

When Michael Lewis published *Moneyball: The Art of Winning an Unfair Game* in 2003, he revolutionized the way that Major League Baseball, the media, and the average fan viewed and understood the game of baseball. Traditional statistics were challenged, obscure statistics became mainstream, and a sabermetric approach to player evaluation became commonplace.

The desire for analytical innovation quickly extended beyond baseball and into other sports. One year after Moneyball, statistician Dean Oliver published *Basketball on Paper* and declared, "When basketball starts playing Moneyball, this is the book they will use." Oliver states that there are four factors to succes in basketball: Shooting Percentage, Forcing Turnovers, Offensive Rebounding Percentage, and Free Throw

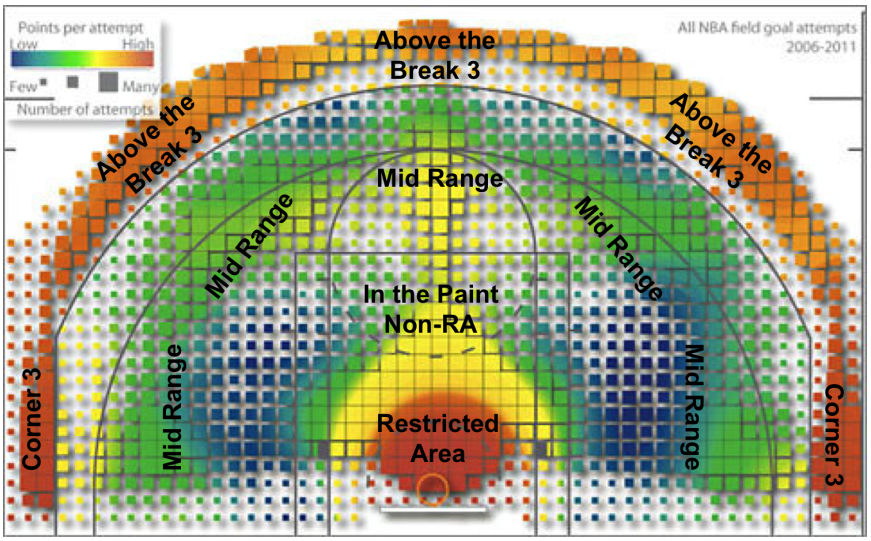

Figure 1: A heat map of every NBA Field Goal Attempt from 2006-2011. The color of the square indicates the Points per Attempt from that location and the size of the square in each zone indicates the number of Field Goal Attemps from that Location. The zonal location labels were added manually for reference.

Attempts. Oliver believes that these factors provide the key to success on both offense and defense and therefore basketball as a whole. Since my project is a study on defense, I should expect to find parallels with Oliver's findings.

Kirk Goldsberry is another statistician who has done research in basketball. He is the inventor of ShotScore and CourtVision along with other statistical programs relating to basketball. Goldsberry is also the creator of Figure 1(Although I added some text to emphasize zones on the court. Many of the conclusions I expected to draw about defense were formulated or supported by this image.

My project will take Dean Oliver's research a step further and attempt to break down his factors into smaller subsets of information. I will use the knowledge gleaned from Goldsberry and the plethora of statistics that are available in this day and age to prove, disprove, and expand upon the existing research in basketball defense.

```
=== Summary ===

Correlation coefficient                  0.7155
Mean absolute error                      1.8759
Root mean squared error                  2.3953
Relative absolute error                 79.2152 %
Root relative squared error             79.2139 %
Total Number of Instances               30
```

Figure 2: Sample summary output from Weka. This summary section will be outputted for each classifier with different values.

3 Project Design

3.1 Weka

Weka is a collection of machine learning algorithms that can be applied to a dataset. These algorithms are capable of classification, numeric prediction, or both. In this project, I will use Weka to test the available machine learning algorithms on multiple combinations of my data. The goal will be to find an algorithm or algorithms that can numerically predict the effectiveness of each defense in my dataset. This requires selecting a numeric value that can represent defense for the algorithms to predict. I will use Defensive Efficiency as my value to predict; my reasoning for choosing this value, and an explanation of its calculation can be seen in Section 3.2.

Weka output provides information on the structure of each generated model as well as a summary section that gives information on the success or failure of the model in predicting the actual numeric value. An example of the summary section can be seen in Figure 2.

3.2 Numerically Defining Defense

There is no official statistic for the numeric value of a basketball team's defense. Typically, one of the following statistics is used as a proxy to numerically rank defenses or offenses in order, however, each statistic is separately flawed.

Points Per Game: The most widely used statistic to compare defenses but also the most flawed. For a comparison of defensive effectiveness, this statistic assumes that each team has approximately the same number of possessions of the ball each game. A statistic known as PACE defines the average number of possessions that a team has each game; for the current 2013-2014 NBA season, PACE ranges from 92.4 to 102.3. Since NBA teams average approximately one point per possession, this could lead to inconsistencies between the statistical representation and actual effectiveness of defense.

Opponent Field Goal Percentage: Another commonly used defensive comparator. This is defined by the number of shots made divided by shots attempted. This statistic weights every shot equally and therefore does not account for the additional point that is earned from a Three Point Shots or the points earned from Free Throws.

$$\text{Field Goal Percentage} = \frac{\text{Field Goals Made}}{\text{Field Goals Attempted}} \times 100 \tag{1}$$

Effective Field Goal Percentage: This modified version of Field Goal Percentage accounts for the additional point for a Three Pointer but still does not account for Free Throws.

$$\text{Effective Field Goal Percentage} = \frac{(\text{Field Goals Made}) + 0.5 \times (\text{Three Pointers Made})}{\text{Field Goals Attempted}} \tag{2}$$

True Shooting Percentage: A more advanced version of Effective Field Goal Percentage. This accounts for both Three Pointers and Free Throws. Unfortunately, the constant that is multiplied by Free Throw Attempts is an estimation rather than a truly calculated value. Additionally, True Shooting Percentage is not readily found online for teams.

$$\text{True Shooting Percentage} = \frac{\text{Points Scored}}{2 \times [(\text{Field Goals Attempted}) + 0.44 \times (\text{Free Throws Attempted})]} \tag{3}$$

Defensive Efficiency: Normalizes the number of points scored over 100 possessions. Since it is points-based, it accounts for Three Pointers and Free Throws. Additionally, since a possession does not end until

the other team gets the ball, this statistic also is able to incorporate the importance of rebounding.

$$\text{Defensive Efficiency} = \frac{\text{Total Points Scored}}{\text{Number of Possessions}} \times 100 \tag{4}$$

For this project, I will use Defensive Efficiency as a means of numerically valuing defense since it incorporates the most aspects of defense. This statistic will be the value that the machine learning algorithms will attempt to predict. Note that small values of Defensive Efficiency indicate better defense and large values correlate with a less effective defense.

4 Project Implementation

4.1 Data Collection

All the needed data for this project was gathered from the official statistical website of the NBA (www.stats.nba.com). The focus of this project is identifying the key attributes of team defense so each data point consists of the statistics of a specific team for a specific year. The available data spans from the 1996-1997 season to the current 2013-2014 season and I will therefore have 532 total data points $[(30 \times 10) + (8 \times 29) = 532$. The NBA expanded from 29 to 30 teams in 2004].

Although there is data available in each year for both the regular season and the playoffs, the data gathered only includes the regular season. The playoff data will be skewed as teams play the same opponent multiple times in a row. This means that the opposition's offense will greatly impact the effectiveness of the team's defense; the quality of opponent is less influential during the regular season when teams play more games and play the same opponent less frequently.

Additionally, all data gathered will be normalized per possession or per game. Fortunately, the statistics website allows for this option when gathering this data so it requires no additional calculations. This step will allow for the inclusion of the data from this season (2013-2014) - which was incomplete at the time of data gathering - and from the 2011 season which was shortened to 66 games (A normal season has 82 games).

The complete data needed for this project is located in a number of different tables on different pages of the NBAs statistics site. An example of one table that contains information on the opponent's shooting

Team	Restricted Area			In The Paint (Non-RA)			Mid-Range		
	FGM	FGA	FG%	FGM	FGA	FG%	FGM	FGA	FG%
Atlanta Hawks	15.1	24.7	61.1%	5.3	13.3	40.2%	7.4	18.4	40.5%
Boston Celtics	15.0	26.2	57.2%	4.2	11.6	35.9%	10.6	26.4	40.3%
Brooklyn Nets	13.4	22.4	60.0%	5.4	12.6	42.5%	8.3	20.5	40.7%
Charlotte Bobcats	15.2	26.2	58.1%	4.9	11.6	42.2%	9.7	26.5	36.5%
Chicago Bulls	15.4	27.0	56.9%	3.6	10.3	34.6%	9.5	25.4	37.5%

Figure 3: Example of a data table located on www.stats.nba.com. The information in this table includes Field Goals Made, Field Goals Attempts, and Field Goal Percentage (Made/Attempts) from specific locations on the court. The court locations can be referenced in Figure 1.

effectiveness from different locations on the floor can be seen in Figure 3.

Since the data covers spans so many tables, it was necessary to write web-scraping scripts to gather the desired data more efficiently. The data on the statistics page is generated dynamically, which means that it does not appear in the source code for the webpage. The solution to this complication requires automating a browser so that the dynamically generated content is included in the source code. I used a tool called Selenium for web browser automation.

While many of the available statistics are not needed for this project, it was much simpler to collect every attribute and then manually remove the attribute columns that are not needed. These included attributes such as Wins, Losses, in addition to statistics that deal only with offense. While these attributes would almost certainly be removed during the attribute selection process described later on, there are potential consequences to including them. In some cases, irrelevant statistics can have negative impacts on machine learning algorithms. Additionally, numerous statistics appear in more than one table (sometimes even with a different column header) and it is not necessary to include the same statistic more than once; these were also removed manually.

At this stage, I was left with 64 attributes in my dataset.

4.2 Algorithm Evaluation

My next step was to run every available algorithm to determine which ones perform the best. The important value that I considered for performace was the correlation coefficient which appears in the Weka

output summary shown in Figure 2. This value ranges from -1 to +1 and is a measure of the correlation between the attributes and the value we are trying to predict. A value closer to +/-1 implies a stronger correlation and a value near 0 implies a weaker correlation.

The two top performing algorithms for the full dataset were a Linear Regression classifier, which outputted a correlation coefficient of 0.9997, and a Multilayer Perceptron model (A type of Neural Network), which outputted a correlation 0.9993. As stated before, a correlation value greater than 0.7 indicates a strong correlation and these algorithms were able to perform far above that value. I continued to evaluate these two algorithms throughout the rest of the project.

4.3 Attribute Selection

Although the two models are extremely accurate in predicting Defensive Efficiency, there are still many superfluous attributes that are present in the model. The presence of these variables makes it difficult to analyze the output and also increase the amount of time needed to compute the models. The following sections describe the removal of attributes and provide justification for their removal.

At each stage the models were recalculate to ensure that they maintained their validity and the Correlation Coefficient remained at an acceptable level

4.3.1 Attribute is not a cause of Defensive Effectiveness

The general reason for removing these attributes is that their inclusion in the model would not provide any information on what makes a defense good or bad. One of these attributes is Opponent Field Goal Attempts. There is nothing substantial to be learned by the amount of shots your opponent takes during the game and it is not expected to be a significant factor in the prediction of defense anyway.

This section also includes the removal of variables that may actually help in the prediction of defense. This includes Field Goal Percentage, Effective Field Goal Percentage, and Points per Game, which were all considered as a means of numerically defining defense earlier on. These attributes are likely highly correlated with Defensive Efficiency as they are also indicators of defensive effectiveness. While these statistics indicate that a defense is good or bad, they do not give any information as to the why. The goal of this project is to

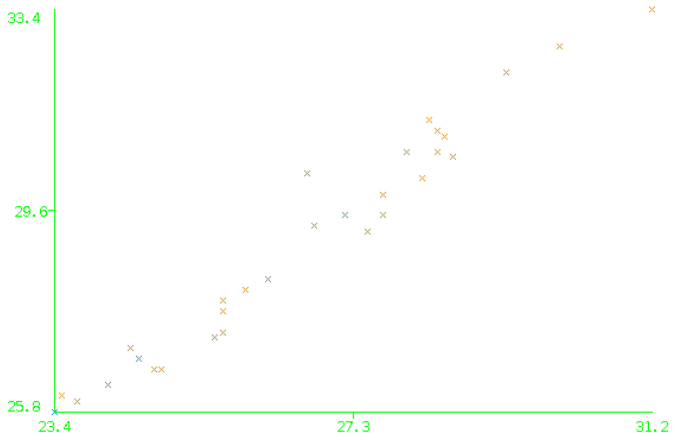

Figure 4: Visualization scatterplot for Restriced Area Field Goal Attempts and 0-5 Ft. Field Goal Attempts. The plot indicates correlation between the two variables.

find attributes that are characteristic of defensive effectiveness and not to find those that are indicative of defensive effectiveness. Therefore, these attributes are removed.

The number of attributes is reduced to 56 at this point. The two models now perform with the following Correlation Coefficients:

Linear Regression: 0.9954

Multilayer Perceptron: 0.9972

4.3.2 Multicollinearity

Multicollinearity is a situation in which two or more dependent variables (attributes) have a strong correlation with each other. As a result of multicollinearity, the effect of each attribute on the predicted value will be significantly reduced, although the Correlation Coefficient should not be significantly affected.

Weka provides a visualizer, which plots every attribute against all the other attributes. An example of multicollinearity observed in the visualizer can be seen in Figure 4. If the plot of the variables can be fit by a single line then there is collinearity between the attributes and one of them needs to be removed.

Some obvious cases of multicollinearity resulted from having the same statistic normalized by a different rate. This was the case with Turnovers per Game and Turnover Ratio (Normalized by 100 possessions), Free

Throw Attempts per Game and Free Throw Attempts Ratio (Normalized by 100 possessions), along with some other statistics. Since Defensive Efficiency is normalized by possessions, I chose to keep the variables in this case that were also normalized by number of possessions.

Multicollinearity can also result when one attribute is a composite of other attributes. One example of this is the locational shooting data that was collected. For each available location, I collected the Field Goals Made, Field Goals Attempted, and Field Goal Percentage, which is calculated by Field Goals Made divided by Field Goals Attempted. In this case, there is no reason to include all three of the statistics and so I chose to remove all of the Field Goal Made statistics for each zone since I viewed the number of attempts allowed and the shooting percentage as being more explanative of defensive style and effectiveness.

Other cases of multicollinearity were less obvious. One such example was Defensive Rebound Percentage and Opponent Second Chance Points. Second Chance Points are a sum of all the points scored after your opponent misses a shot and then gets the offensive rebound. In this case, I considered Defensive Rebounding Percentage to be the root cause of Second Chance Points since the team must first get the rebound in order to receive Second Chance Points. Therefore, I removed Second Chance Points from the model. I used similar logic to remove Opponent Points in the Paint in favor of Opponent Restricted Area Field Goals Made.

4.3.3 Locational Data Overlap

My full dataset included locational shooting data that was broken up in two different ways. One way of dividing this data was to break up the data into 5-foot segments calculated by the distance from the basket. The other was to break up the data into zones of the court, which included the Restricted Area, In the Paint (Non-Restricted Area), Mid-Range, Left Corner Three, Right Corner Three, and Above the Break Three. The location of these sections can be seen in Figure 1.

The two locational strategies both sum together to represent the same total data. Therefore, these sections have a fair amount of overlap across the two areas with some representing almost the exact same area. These sections with direct overlap are another example of multicollinearity. One example is the Restricted Area zone (which represents an area 0-4 feet from the basket) and the 0-5 foot distance segment, the In the Paint (Non-Restricted Area) zone and the 5-9 foot segment, and the Mid-Range zone and the 15-19 foot segment. For each of these overlapping areas, I removed each locational attribute individually

and tracked the change in the Correlation Coefficient. In all three of these cases, the removal of the zonal representation had a larger effect on the correlation than the segmental data. This indicates that the zonal representation is more important to the model, providing reason to remove the three segments above.

The number of attributes is reduced to 28 at this point. The two models now perform with the following Correlation Coefficients:

Linear Regression: 0.9886

Multilayer Perceptron: 0.9961

The fact that the segmental location data was unimportant compared to the zones that it overlapped with suggested that other segmental data was also not needed. Removal of the Field Goal Attempts and Field Goal Percentage data from 10-14 feet, 20-24 feet, and 25-29 feet reduced the number of attributes to 22. The minor effects on the Correlation Coefficients of the two models can be seen below and suggest that these attributes are not needed.

Linear Regression: 0.9863

Multilayer Perceptron: 0.9958

Note: At this point it becomes clear due to the more rapid decline in the Corelation Coefficient that the Linear Regression model is more reliant on the number of attributes than the Multilayer Perceptron model is.

4.3.4 Basketball Reasons

Before this point, attribute removal was done very mechanically. Basketball knowledge played a role in an understanding of what the statistic meant, but no statistic was removed specifically because I did not think it was important to the success of the model. In this section, I removed statistics that I did not believe to be important one by one and tracked the value of the Correlation Coefficient for each removal. The individual results can be seen in Table 1.

The number of attributes is reduced to 16 attributes at this point. The two models now perform with the following Correlation Coefficients:

Linear Regression: 0.9818

Multilayer Perceptron: 0.9952

Interestingly, I tried to remove Opponent Free Throw Percentage, but was surprised by the effect on the Correlation Coefficients of both models. I could not justify removing it because it contributes more to the correlation than attributes which I want to keep.

Table 1: Incremental Removal of Attributes for "Basketball Reasons"

Attribute Removed	Linear Regression	Multilayer Perceptron
Before	.9863	0.9958
Opponent Assists	0.9863	0.9953
PACE	0.9824	0.0.9949
Opponent Points Off Turnovers	0.9811	0.995
Blocks	0.9815	0.9944
Offensive Turnover Ratio	0.9815	0.995
Opponent Fast Break Points	0.9818	0.9952

4.4 Final Attribute Set

At this point I have the set of attributes that I will analyze in the following sections. The 16 attributes are ranked in order of importance. The process by which I decide importance is detailed in Section 5.2.

1. Opponent Turnover Ratio
2. Restricted Area Field Goal
3. Defensive Rebound
4. Mid-Range Field Goal
5. Opponent Free Throw Attempt Ratio
6. In The Paint (Non-RA) Field Goal
7. Above the Break 3 Field Goal
8. In The Paint (Non-RA) Field Goal Attempts
9. Mid-Range Field Goal Attempts

10. Restricted Area Field Goal Attempts
11. Opponent Free Throw
12. Left Corner 3 Field Goal
13. Right Corner 3 Field Goal
14. Above the Break 3 Field Goal Attempts
15. Right Corner 3 Field Goal Attempts
16. Left Corner 3 Field Goal Attempts

5 Analysis

5.1 Algorithms Used

In this section I will briefly explain the algorithms that were used and why I believe they were able to create effective models of my data.

5.1.1 Linear Regression

A simple linear regression with one dependent variable and one independent variable (attribute) finds the straight line that best fits the data (minimizes the error). If a second independent variable is included, the algorithm finds the flat plane that best fits the data. As more variables are added, the structure of the model is similar although considerably more difficult to visualize. Each variable is assigned a weight that is multiplied with its value; these are all summed together with a constant to produce the output value. The structure of this model makes the output relatively easy to analyze.

The success of this model suggests individual pieces can sum together to represent a defense. This suggests that the importance of each attribute is not reliant in any way on the value of other attributes. However, the fact that the performance of this model declined at a faster rate than the Multilayer Perceptron might indicate that interaction of attributes does play a small role.

Since my attributes are not normalized values, we cannot compare the weights of the attribute to each other, although the sign of the weight can be used in analysis. Below are some observations of the weights and conclusions. Keep in mind that a lower value for Defensive Efficiency indicates a better defense.

- The Field Goal Percentage for each zone have positive weights.
  - This shows that a lower percentage shot by your opponent from any zone leads to a higher Defensive Efficiency of your team. Basically, better defenses force their opponent to shoot a poor percentage from every zone compared to worse defenses.
  - This conclusion is not astonishing, but it is certainly correct. An accurate model should confirm things that are already known.
- Opponent Turnover Ratio and Defensive Rebound Percentage have negative weights.
  - Forcing turnovers is a sign of a good defense.
  - Rebounding a high percentage of your opponent's missed shots is a sign of good defense.
- The Restricted Area and the Corner Three Field Goal Attempts have positive weights, the other zones have negative weights.
  - These are the areas that you don't want to let your opponent shoot from. Forcing them to shoot from the other areas will make you a better defense.

5.1.2 Multilayer Perceptron

A Single Layer Perceptron has a very similar structure to a Linear Regression model. Attributes are given a weight that is multiplied with the value of the attribute and then summed together to find the output value. The difference lies in the training of the model. Each weight is given an initial arbitrary value and then the error is calculated for the model. The model then incrementally adjusts the weights to reduce the error. After many iterations, the model is able to accurately predict the output.

A Multilayer Perceptron works similarly but with the addition of a hidden layer which allows for interaction of attributes. The strong performance of this model (and better performance than the Linear

Regression) indicates that there is likely some non-linear computations involved in predicting Defensive Efficiency. However, there is also a danger of overtraining the data, since after enough iterations the model can learn to account for every data point individually without being able to predict data points that are not in the training set. Using a 10-fold cross validation typically discourages overtraining.

Unfortunately, the structure of the Multilayer Perceptron makes it nearly impossible to analyze the output.

5.1.3 Algorithm Comparison

The key to my project was finding the predictive attributes so the Correlation Coefficient is my most important comparator between these two algorithms. The Multilayer Perceptron outperforms the Linear Regression model with correlation scores of 0.9952 and 0.9818, respectively.

In order to ensure that the models show a consistent ability to predict values, I mapped a plot of the residuals over time. Both plots show consistent noise around 0, which further validates the models. Additionally, the Multilayer Perceptron plot seen in Figure 5 error ranges from -2 to +2, while the Linear Regression plot seen in 6 ranges from -6 to 6. This reinforces the Multilayer Perceptron as being the superior model.

5.2 Attribute Weights

Since the Multilayer Perceptron is the better performing model, the attribute weights will be based off of this model. The process of calculating the weights is described below and the weights can be seen in Table 2.

1. Individually remove each attribute from the model and track the change in the Correlation Coefficient.
2. Sum the individual changes to the Correlation Coefficient to get a total change.
3. For each attribute, divide the individual change by the total change and multiply by 100 to calculate its weight
    (a) The weights are a percentage out of 100 and represent the importance of the attribute.

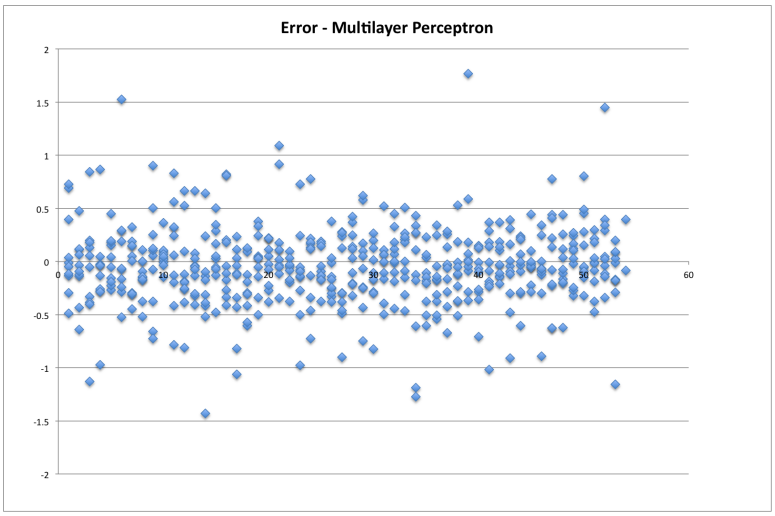

Figure 5: A plot of the residuals from the Multilayer Perceptron algorithm run on the final attribute set. The y axis represents the magnitude and direction of the error. Constant noise around 0 further confirms the validity of the model

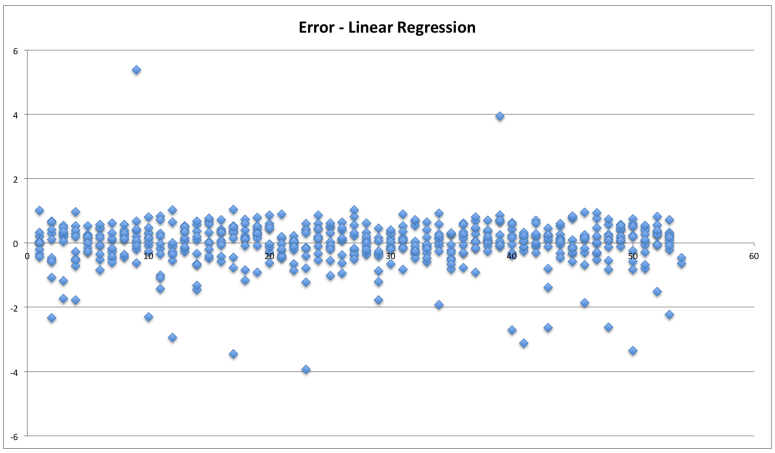

Figure 6: A plot of the residuals from the Linear Regression algorithm run on the final attribute set. The y axis represents the magnitude and direction of the error. Constant noise around 0 further confirms the validity of the model

Table 2: Attribute Weights

Attribute	Weight (%)
Left Corner 3 Field Goal Attempts	0.11
Right Corner 3 Field Goal Attempts	0.23
Above the Break 3 Field Goal Attempts	0.86
Right Corner 3 Field Goal Percentage	0.92
Left Corner 3 Field Goal Percentage	0.98
Opp. Free Throw Percentage	1.21
Restricted Area Field Goal Attempts	1.38
Mid-Range Field Goal Attempts	1.46
In The Paint (Non-RA) Field Goal Attempts	1.61
Above the Break 3 Field Goal Percentage	4.48
In The Paint (Non-RA) Field Goal Percentage	4.60
Opp. Free Throw Attempt Rate	6.60
Mid-Range Field Goal Percentage	8.81
Defensive Rebound Percentage	13.38
Restricted Area Field Goal Percentage	25.83
Opp. Turnover Ratio	27.55

5.2.1 Analyzing the Attribute Weights

The calculate weight of an attribute stand for the importance of that attribute in predicting Defensive Efficiency. Therefore, the weights represent the key differentiators in defense and not necessarily the most important aspects of defense. It is important to keep in mind that some attributes have a much higher variance than others; these attributes will inherently have a higher weight than those with a low variance.

In every case, the weight of the Field Goal Percentage from a zone is weighted higher than the weight of the Field Goal Attempts from that zone. This suggests that the variance in Attempts from each zone is somewhat insignificant, and it is far more important for a defense to force a difficult shot, no matter where it comes from.

The weights of the Field Goal Percentages in each zone are in the following order (most to least): Restricted Area, Mid-Range, In The Paint (Non-Restricted Area), Above the Break 3, Left Corner 3, and Right Corner 3. In fact, the weight of the Restricted Area zone is greater than that of the other zones combined. This lends credibility to the idea that the most important defenders on a team are the Centers and Power Forwards (The two players who are responsible for defending the Restricted Area).

6 Discussion and Future Work

My initial motivation for this project was to determine the attributes that are most characteristic of defensive effectiveness in the NBA. Instead, I completed a slightly different task by finding the key differentiators that separate NBA defenses. The difference is subtle but important. For example, if my findings were to be applied to my initial goal, it would suggest that defending the Three Pointer is not important. However, a team cannot simply allow their opponents to shoot open Three Point shots or their defense will be horrible and the will lose. Rather, my findings suggest that most teams defend the Three Pointer with a similar level of effectiveness, and the minor differences in these values do not decide by themselves whether a defense is good or bad.

In retrospect, I'm not certain that its possible to answer my initial goal with a purely statistical approach. However, I do believe that my current findings could be used to improve an NBA defense a small amount, and that this research is well-positioned for future work that could improve defense and personel choices a large amount.

I began to look at trying to calculate the importance of players relative to Defensive Efficiency but was unable to create a model that shows significance. I do think that there is a way to find correlation but that I simply haven't figured it out yet. Perhaps trying to find a correlation between players and one of the attributes that predict Defensive Efficiency would be a more completable task. Certainly, I would expect a relationship between Centers and the Restricted Area Field Goal Percentage. The new SportVu "Rim Protector" statistic could also be used to try and find a relationship.

Using my same approach to predict offense would lead to interesting conclusions. I believe that offenses are much more different from one another stylistically than defenses are. It would be interesting to see how the attributes, weights, and even algorithms would change when offense was the object of investigation instead of defense.

Ultimately, there are many lessons to be learned for any Machine Learning project. It's important to have expectations of what you expect, but its equally as important to not tailor the process to ensure that those conclusions are met. Finding these results organically ensures that the model is truer and not affected by these preconceived expectations. Additionally, it allows for beliefs to be disproved in the analysis stage,

leading to conclusions that are far more interesting.

Lastly, I would be curious to see someone do research into Opponent Free Throw Percentage. It seems to me to be an arbitray statistic, but my model found that it had significance. I would be interested in seeing if there is some concept of "smart" fouling that exists in the NBA or if there is some other explanation.