# Data Mining for NBA Shot Predictions

Matthew Byam, Matthew Spirio, Emma Pullen

Data 450-111: Data Mining & Predictive Analytics

May 3 2023

**Abstract**

Sports continue to be quantified more and more everyday as technology continues to advance. With the rise of sports office's actually using analytics to build their teams and game plan against opponents, the more numbers that they have access to, the better. Basketball is an extremely popular sport that is just beginning to enter into the realm of advanced analytics. With this new magnitude of numbers that are becoming available, we can begin to create an approach that can help basketball teams improve.

In order to assist these basketball teams in the future, we will analyze the varying factors that contribute to making different shots (location and point value) in basketball. We will train and test the data since there is such a large amount. A stepwise linear regression will be run so we can determine which variables are most important in terms of successfully making a shot. We will perform a KNN classification on the shots that were made versus the shots that were missed. Following the KNN classification, we will create a decision tree that takes into account all of the variables and whether or not the shot was successfully made along with calculating the performance metrics from the confusion matrix. By seeing how often shots are made based on distance, shot clock, and closest defenders, among other metrics, we can determine the optimal environment needed for shots to be made for the average player.

**Introduction**

The use of data analytics has continued to rise in sports, even more so in recent years. At every single level of sports, from the top professional leagues down to lower amateur contests, everybody wants to find any kind of edge possible to beat their opponent. As advanced analytics have become more prevalent in sports, it raises questions about the extent to which coaches and

athletes alike believe in their role in success, as well as why some teams rely on them more than others.

*Moneyball,* a book published in 2003 by Michael Lewis, follows the Oakland Athletics of the MLB, and is considered to be the grandfather of modern analytics, depicting the first real time a major sports franchise went against the grain in accepting analytics. The book describes how the Athletics were able to find deficiencies in the free agent market, and build their team heavily on players that excelled at undervalued statistics, such as on-base-percentage. The team went on to be very successful, and while they were unable to win a championship, they laid a blueprint that was followed by the rest of the league. In 2017, Brown et al. recorded a 10-year retrospective look at *Moneyball*, found that post-*Moneyball* free agents were compensated more when they excelled in these new statistical areas, and that the impact that was had was greatly felt throughout the league. However a further investigation, done five years after, found that while there certainly was more acceptance, not every team in the league had fully committed (Duquette et al., 2019).

This skepticism among athletes and coaches was posed by Davenport (2014) when he theorized that most of those who get heavily involved in professional sports rely instead on their experience and playing expertise instead of computers who are outputting numerical results. This idea is supported in a similar study by Baghurst, et al. (2021) in which they studied how basketball teams performed when their coaches designed plays and created their offensive and defensive sets based on their pre-concieved notions of their players skills, instead of analytical data that showcased exactly where their players were strongest on the court. The result was that most coaches were unable to accurately find their players strengths and weaknesses, and couldn't create successful winning basketball better than a computer using analytics. Additionally,

coaches operate using bias when they are not using analytics, which is something that can be completely removed when conducting studies using data. This is the starting point and basis for our studies and exactly what we aim to try and achieve using NBA shot data.

**Dataset**

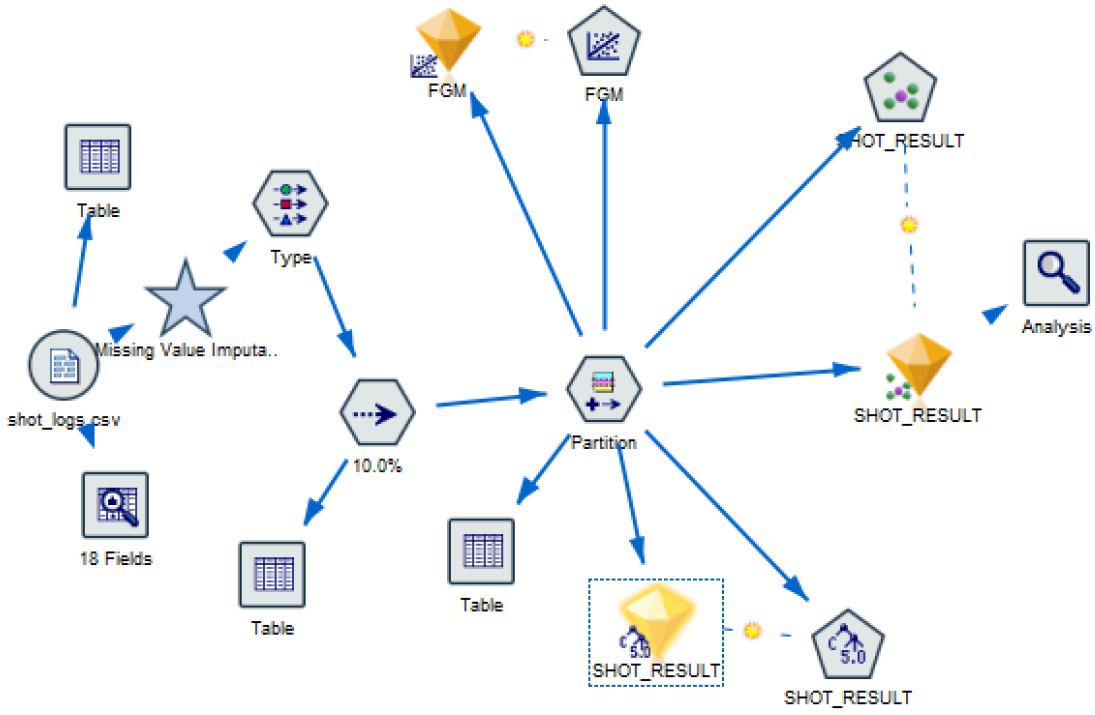
In this report, we will be utilizing the NBA shot logs from the 2014-2015 season. Our data set contains 128,070 values that includes the game ID, the date of the matchup, location of the matchup, win or loss, final margin, shot number, period, the game clock, shot clock, dribbles, touch time, shot distance, point value of the attempted shot, shot result (made or missed), the closest defender's name, the corresponding ID of the closest defender, the distance from the shooter to the closest defender, field goals made,  points scored, the shooter's name, and the shooter's corresponding ID.

**Report**

In the creation of this model, we utilized many of the skills and techniques that we learned throughout the semester so far, specifically major topics such as KNN Classification, Decision Trees, and Linear Regression. Initially when building the topology, we noticed that there were some NULL values in the data set. So, we included a Missing Value Imputation Node that replaced the missing values with the mean. Then, we included a type node to manage the data types of all of the fields and adjust flags as necessary. Again, the next node we used was a 10% sample node, which is necessary when working with such a large data set that includes over 120,000 records because it becomes difficult to train a model in a realistic amount of time with laptop level computer specs. This was not only necessary for training the model but it makes viewing and analyzing graphs and charts much easier. 10% of over 120,000 is still a large amount of data that will allow us to see and investigate a majority of the trends without

overloading visualizations to the point of being uninterpretable. From here, the last step we took to prepare the data for our model was by splitting the dataset into training and testing records. This is performed by the partition node, with a ratio of 70% Training data to 30% Testing data, and is a necessary step when building any predictive model because you always need to have different testing and training data so that you can evaluate the performance and accuracy of the model after it has been trained.

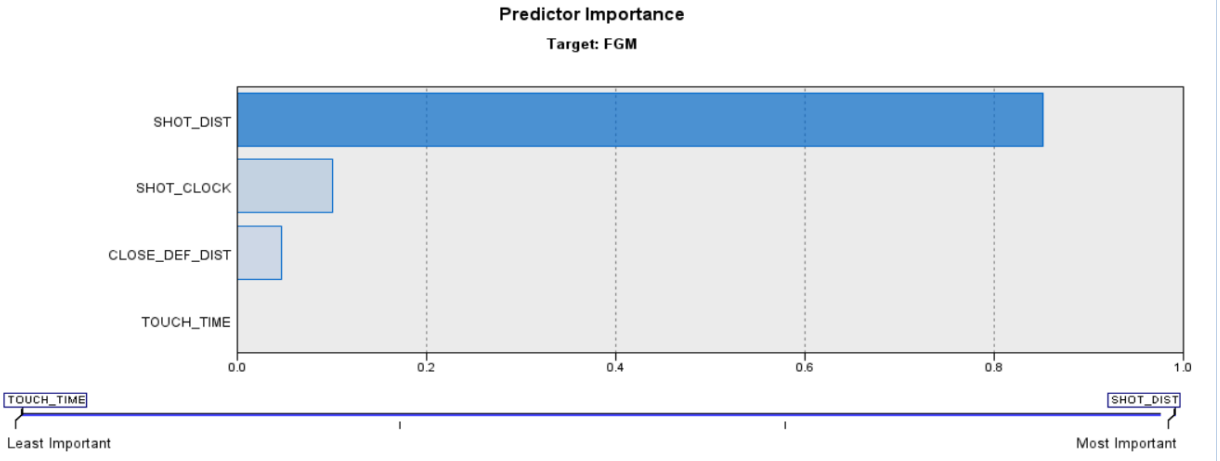
**Shot_Logs Full Topology**




After cleaning the data, we first wanted to use a linear regression model to determine which fields will play a larger role in predicting a FGM (Field Goal Made). We decided to start with a linear regression model because it provides us with basic generalizations and trends that were not too complex when first beginning to work with the data. Following the regression model we created, we were able to see that there was a distinct correlation between SHOT_DIST

and if the shot was made, which logically makes sense as well. For instance, if someone attempts to make a shot from a shorter distance, they are much more likely to successfully make the basket. We also could see that there were two other correlations to shot made that appeared in our model which were SHOT_CLOCK and CLOSE_DEF_DIST (Time left on the shot clock, and how close the nearest defender is to the shooter). From a basketball perspective, both of these correlations make perfect sense. As time decreases on the shot clock, the player may become flustered or not be in a great position to get a fair shot off. Additionally, if a defender is playing from a close distance, it would be difficult for the shooting player to produce a quality shot. That being said, there will always be greater expectations for the shot to be made if it is, say a two foot layup rather than a three point attempt from midrange since it is an objectively much simpler shot. This reinforces our findings that SHOT_DIST makes a much larger impact than the other factors.

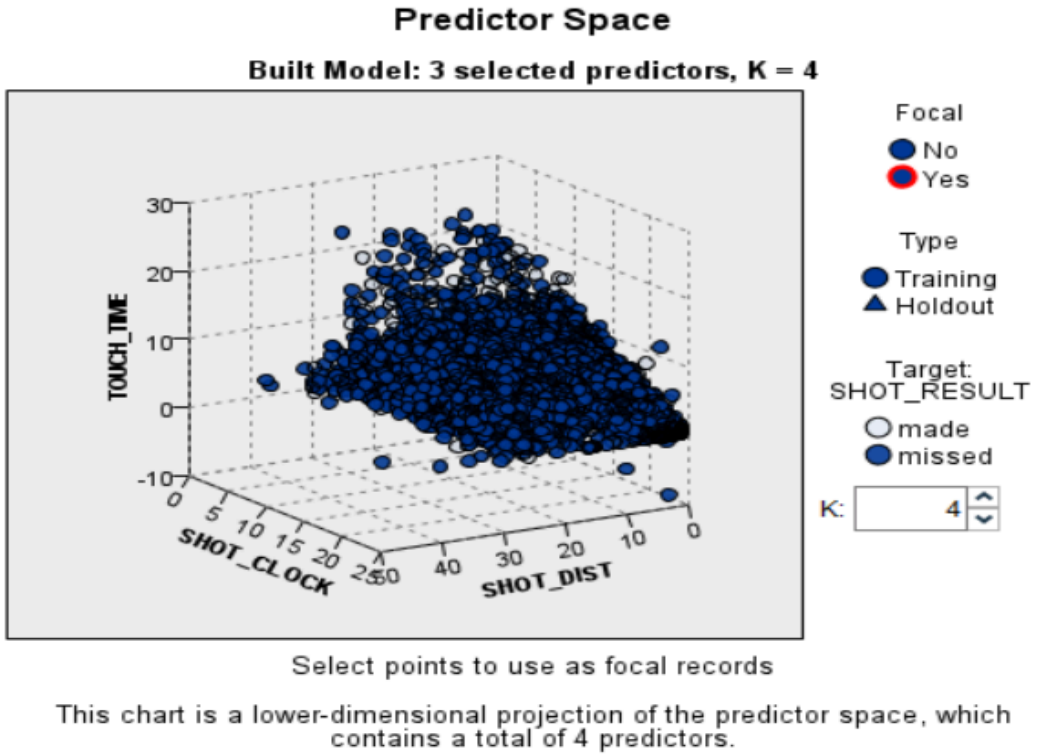
**Linear Regression - FGM, Predictor Importance**




**KNN**

To build our classification model, we utilized the KNN classification method. To do this, we decided to reduce the size of the data since there were so many records. We decided to take a

10% random sample of the 128,070 records to give us a smaller data set to work with. Luckily, the dataset already included the shot result (made or missed), so we did not need to reclassify any of the data. From there, we partitioned the data into training and testing sets (70% training and 30% testing). Following the partition, the KNN classification was run and due to the 10% sample it did not take a notably long time to run. Below is the Predictor Space of 3 selected Predictors.

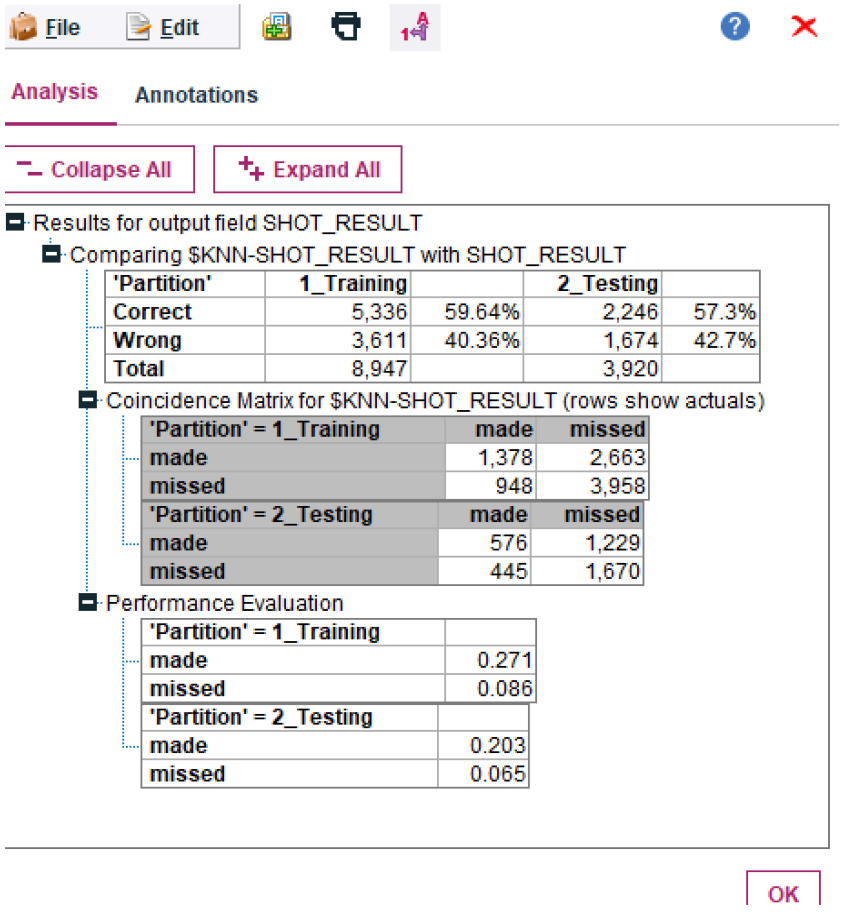



**Confusion Matrix**

In the figure below, we can see the performance of the classifiers. We see from the table that our training data was correct 59.64% of the time. Additionally, our testing data accuracy is reported 57.3% correct. This performance metric may appear extremely low and inaccurate at first glance, but is extremely consistent with the inconsistencies in a basketball player's performance. For example, a FGM% (field goal made percentage) of .500 (50%) or above is considered a very good percentage, although this criterion does not apply equally to all positions.

The highest reported FG% in the 2016-2017 season was 0.714 by DeAndre Jordan who plays for the Clippers. The average FG% of the top 50 leaders for the 2016-2017 season was 0.509. There are approximately 450 players in the NBA and the top 50 of those players only have a FG% of about 50%. The highest reported FG% for a team was 49.4% by the Golden State Warriors. Therefore, a 57.3% prediction percentage is in fact quite good considering the variability in player's abilities, their positions, and the overall low percentage that they successfully complete a shot. The accuracy percentage is also quite a bit above the average player's shooting percentage, which is incredibly useful from a numbers standpoint. Since we have a confusion matrix from this node, we are now able to calculate other performance metrics such as accuracy, recall, precision, FP rate, and the specificity rate.

**KNN SHOT_RESULT Analysis**




| 'Partition' | 1_Training | | 2_Testing | |
|---|---|---|---|---|
| Correct | 5,336 | 59.64% | 2,246 | 57.3% |
| Wrong | 3,611 | 40.36% | 1,674 | 42.7% |
| Total | 8,947 | | 3,920 | |

| 'Partition' = 1_Training | made | missed |
|---|---|---|
| made | 1,378 | 2,663 |
| missed | 948 | 3,958 |
| **'Partition' = 2_Testing** | **made** | **missed** |
| made | 576 | 1,229 |
| missed | 445 | 1,670 |

| 'Partition' = 1_Training | |
|---|---|
| made | 0.271 |
| missed | 0.086 |
| **'Partition' = 2_Testing** | |
| made | 0.203 |
| missed | 0.065 |

**Performance Metrics Calculations:**

To calculate the Accuracy, we use the equation (TP+TN)/(TP+TN+FP+FN) which correlates to (576+1670)/(576+1229+445+1670), which equals 57.3%, which we can see is calculated by modeler as well (Testing=Correct). Since the testing accuracy is a bit less than the training data accuracy, this could be a sign that our model is overfitted.

Next, we can calculate the Recall which uses the equation TP/(TP+FN) which corresponds to 576/(576+1229) which equals 31.9%. The recall percentage shows us that we captured 31.9% of shots made.

Additionally, we are able to calculate the Precision utilizing the equation TP/(TP+FP). Precision is the fraction of instances predicted as positives that are correctly predicted as positives. For our model, the precision is given by 576/(576+445) which equals 56.42%.

The FP rate is given by fraction of false positives over all positives (FP/(TN+FP)). For our model, that equals 1670/(1670+445) or 78.96%.

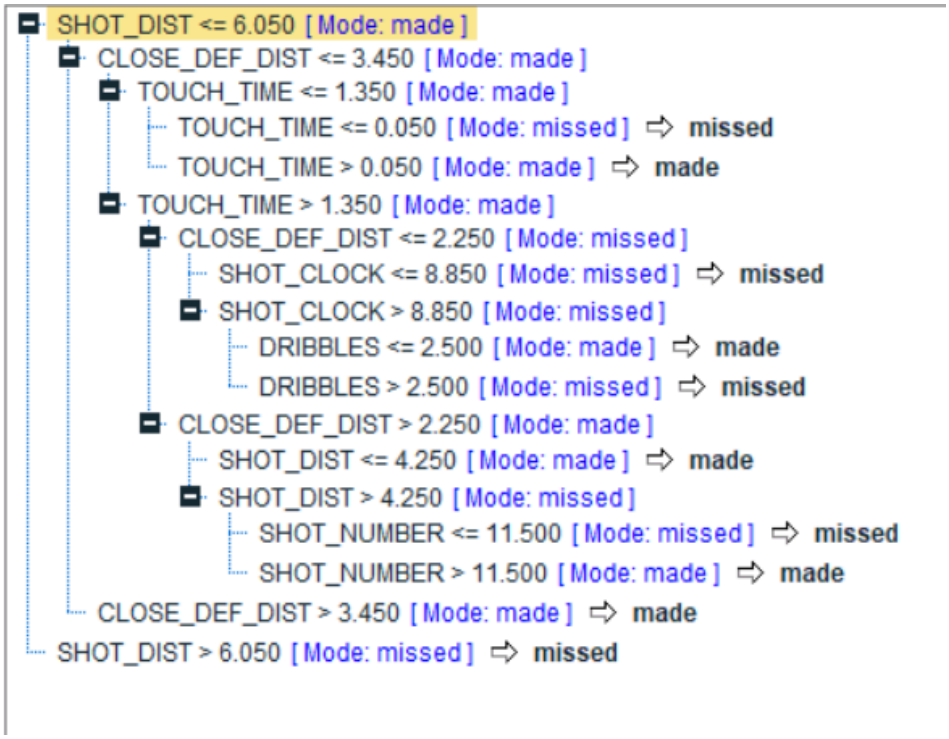
Lastly, Specificity can be given by subtracting the FP rate above from 1. Specificity provides the proportion of actual negatives which were predicted as a true negative. For our model, the specificity is 21.04% .

**Decision Tree construction**

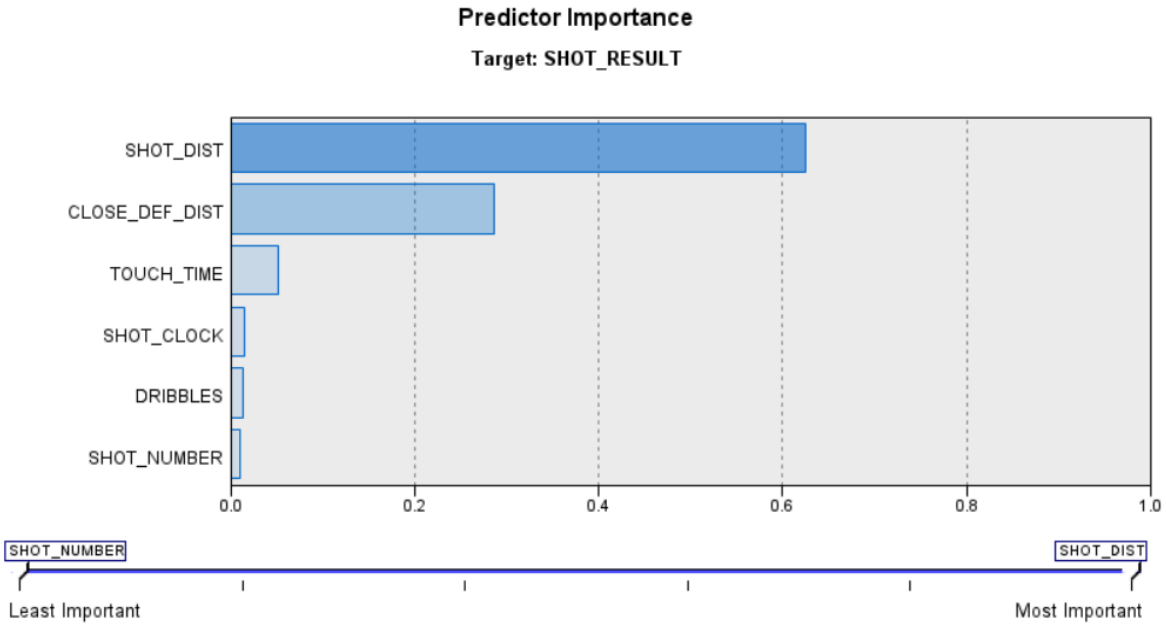
Following our KNN classification, we decided to create a C5.0 decision tree. Decision Trees can be used to characterize the predictors based on the rules derived from the tree. In our case, we will create a decision tree to describe the features that characterize if the shot was made or missed. Now, we are able to create our C5.0 decision tree. Below is a model view of our decision tree followed by predictor importance, and a picture of our full decision tree including

all nodes and branches. From the predictor importance chart, we can see that shot distance is by far the most important predictor with a value of about 0.65. As stated previously, it makes perfect sense that the distance from the hoop affects the difficulty of the shot and therefore would be a very crucial predictor. Following shot distance, the closest defender's distance is the next most important predictor with a value of  approximately 0.3. Thus, shot distance and the distance to the closest defender accounts for about 95% of the predictions. The rest of the predictions are accounted for by touch time, shot clock, dribbles, and shot number.

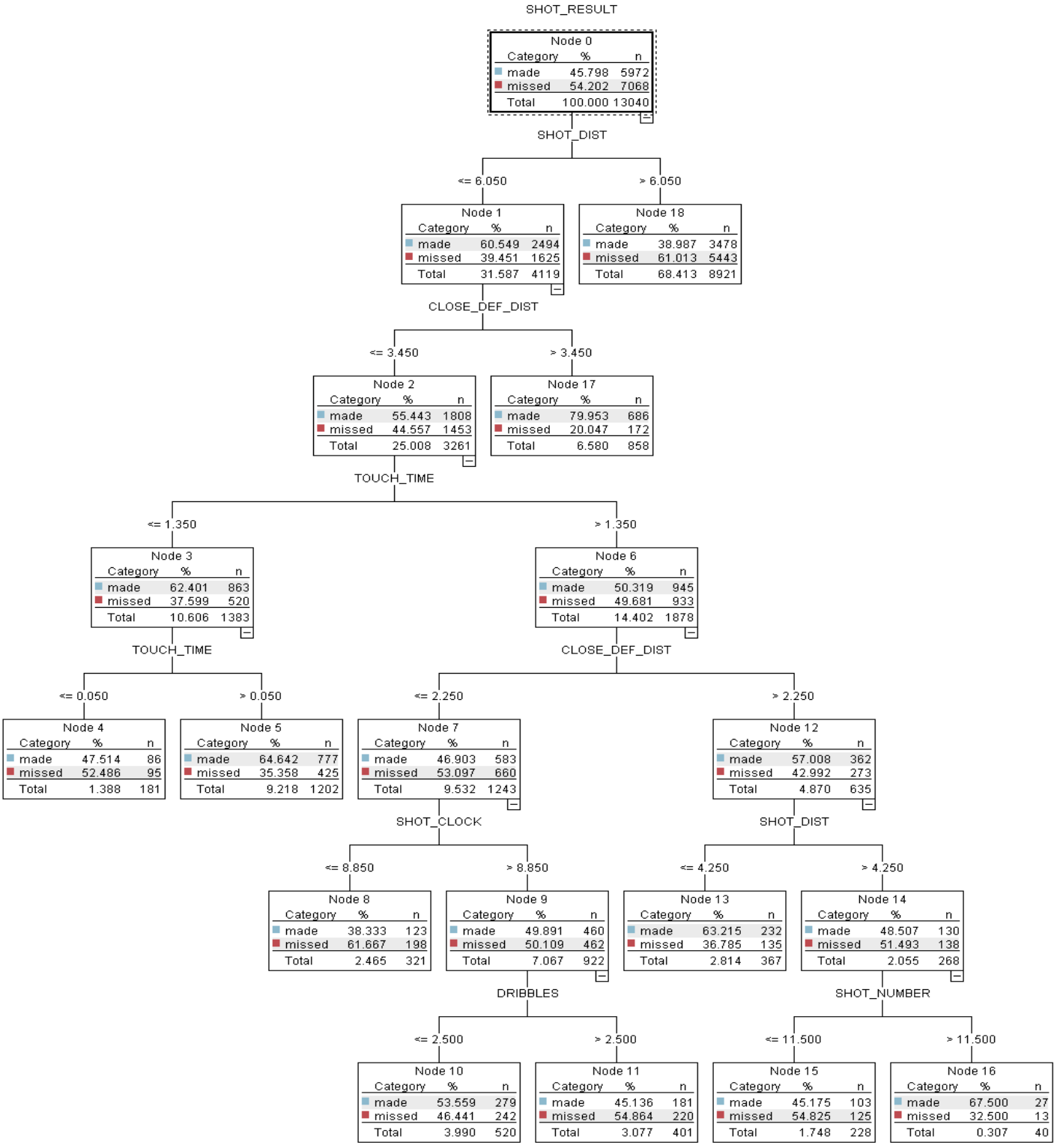
**Decision Tree Model View**



## Decision Tree Predictor Importance



# Full Decision Tree



**Decision tree Rules & statistical metrics (confidence / support)**

Based on the tree we created we can create rules that correspond to if a shot was made or missed with an accompanying confidence and support. If the shot distance is less than or equal to 6.050 feet, the closest defender is within 3.450 feet or less, and touch time is less than or equal to 0.050 seconds, then the shot is missed (Support: 95/13040; Confidence: 95/180). If the shot distance is less than or equal to 6.050 feet, the closest defender is within 3.450 feet or less, and touch time is greater than 0.050 seconds but less than 1.350, then the shot is made (Support:777/13040; Confidence: 777/1202). If the shot distance is less than or equal to 6.050 feet, touch time is greater than 1.350 seconds, the closest defender is within 2.250 feet, and the shot clock has less than 8.850 seconds left, the shot will be missed (Support: 198/13040; Confidence: 198/321). If the shot distance is less than or equal to 6.050 feet, touch time is greater than 1.350 seconds, the closest defender is within 2.250 feet, the shot clock has more than 8.850 seconds left, and the amount of dribbles are less than or equal to 2.5, the shot will be made (Support: 279/13040; Confidence: 242/520). If the shot distance is less than or equal to 6.050 feet, touch time is greater than 1.350 seconds, the closest defender is within 2.250 feet, the shot clock has more than 8.850 seconds left, and the amount of dribbles are more than 2.5, the shot will be missed (Support: 220/13040; Confidence: 220/401). If the shot distance is less than or equal to 4.250 feet, touch time is greater than 1.350 seconds, the closest defender is greater than 2.250 feet away but less than 3.450 feet, the shot will be made (Support: 232/13040; Confidence: 232/367). If the shot distance is greater than 4.25 feet but less than 6.050 feet, touch time is greater than 1.350 seconds, the closest defender is greater than 2.250 feet away but less than 3.450 feet away, and the shot number is less than or equal to 11.5, the shot will be missed (Support: 125/13040; Confidence: 125/228). If the shot distance is greater than 4.25 feet but less

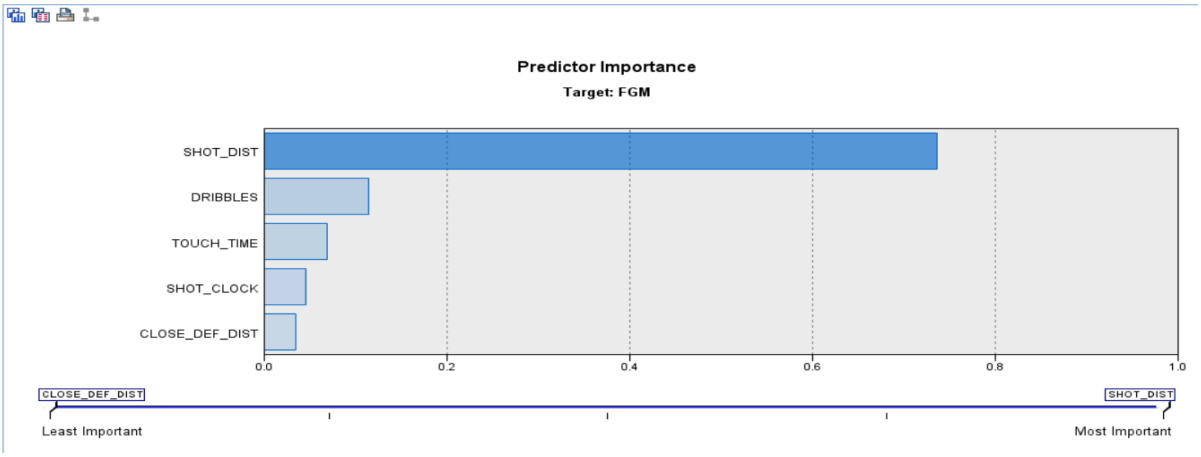
than 6.050 feet, touch time is greater than 1.350 seconds, the closest defender is greater than 2.250 feet away, and the shot number is greater than 11.5, the shot will be made (Support: 27/13040; Confidence: 27/40). Additionally, if the shot distance is less than or equal to 6.050 feet and the closest defender is greater than 3.450 feet away, the shot will be made (Support: 686/13040; Confidence: 686/858). Lastly, if the shot is from greater than 6.050 feet, it will be missed (Support: 5443/13040; Confidence: 5443/8921).

**Individual Player Prediction**

After determining the most important indicators of whether a shot goes in or not, we wanted to take things a step further and actually compare two players on the same team who play similar roles, but have slightly different skill sets, to see if we could find the best way to utilize them in an offense. For this study we selected a pair of All-Star teammates and multiple time champions, Tony Parker and Manu Ginobili of the San Antonio Spurs.

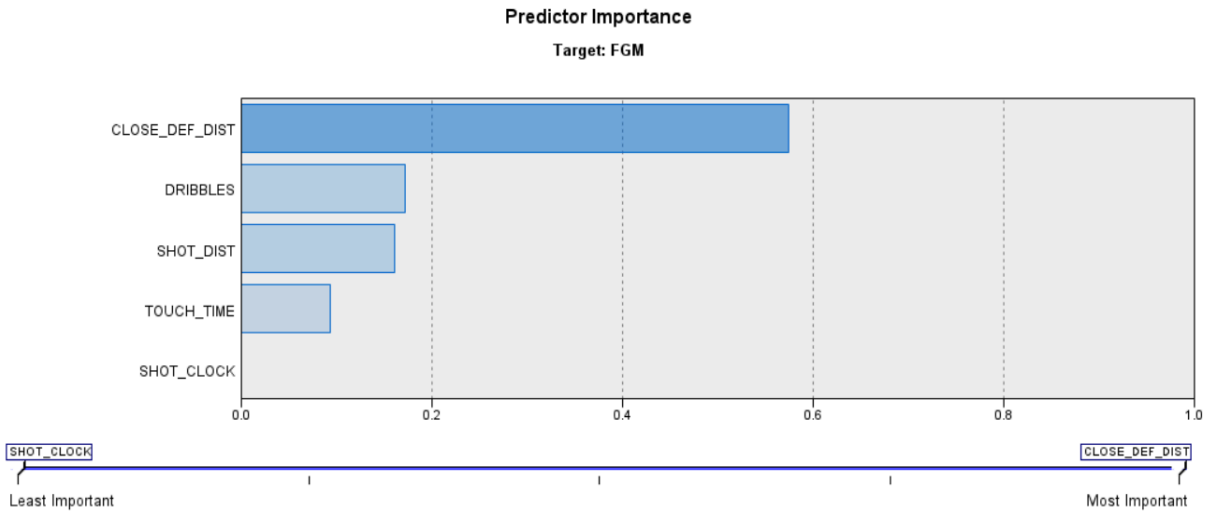
In order to see their individual metrics and how each indicator affects their shot making ability, linear regression onto the field goal being made, was used on each of their shots separately. When doing so, using predictor importance, we can see which statistics are most crucial for each player's shot making.

**Tony Parker Predictor Importance**



By looking at Parker's predictor importance we can see that the most important aspect for Parker is shot distance. What this means is that getting him as close to the basket as possible is ideal in order for him to make the most shots. While this is to be expected, when comparing with his teammate Manu Ginobili, we can see that this is not always the case.

**Manu Ginobili Predictor Importance**



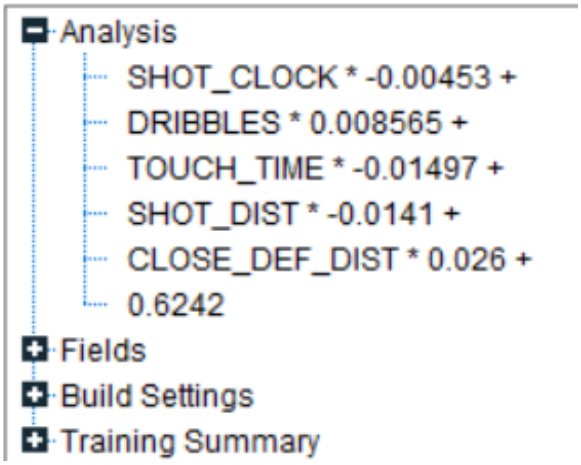

With Ginobili, the most important indicator is how close the opposing defender is, meaning that if he can get open looks away from the defense, he becomes a way better shooter and more valuable asset to the team. This is different for Parker, who is less affected by opposing defenders.

Now when applying this practically, in a general sense, it appears that Parker would be a player better suited to have the ball in his hands, and try to drive to the basket where he is way more efficient, instead of being a shooter who waits to receive the ball. He could be used to try and get inside and draw defenders towards him, as he is less affected by such things, and get his

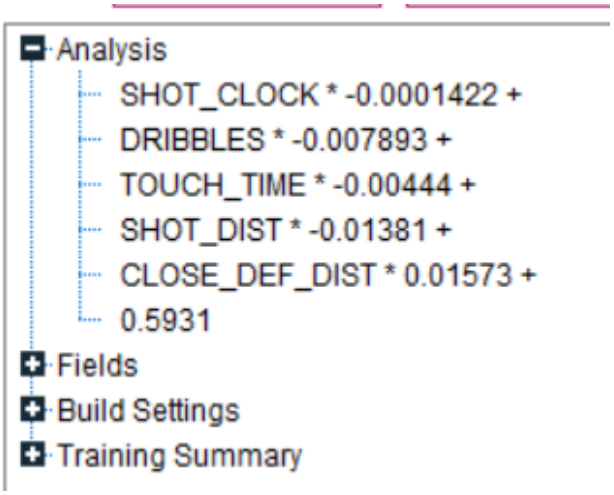
teammates, such as Ginobili who is way better when he is wide open, shots away from the defense.

This idea is furthered when looking at the actual coefficients of their individual linear regressions to see the numerical impact of their statistical indicators.

**Tony Parker Linear Regression**




**Manu Ginobili Linear Regression**




When comparing these two, the biggest fact of note is that while shot clock, touch time, shot distance and closer defender distance affect each player in similar ways, Parker's shot making ability is actually improved when he dribbles more (indicated by a positive coefficient) whereas for Ginobili, it slightly decreases his ability (seen by the negative coefficient. This further adds to the premise that Parker should be dribbling more and more and attacking the basket, whereas Ginobili should try to limit his dribbles and instead look for open catch-and-shoot opportunities.

While this is an analysis done on a small scale involving just two players for the sake of time and memory, it could be applied to an entire team and even opposing teams to figure out each player's strengths and weaknesses.

**Conclusion**

After spending an extensive amount of time selecting, cleaning, modeling, and analyzing data from Shot_logs.csv we are able to understand the basic data analytics behind the game of basketball. Utilizing many different methods of analysis we were able to see how specific variables such as shot distance, defender distance, amount of time remaining on a shot clock, and specific player attributes can help a team put their players in ideal situations suited to their individual strengths. Through Linear Regression we were able to determine a basic outline of how important each variable will be in building our future models with KNN and Decision Trees. After analyzing the results from our KNN model we saw how a predictive model actually performed and we saw that both training and testing had roughly an accuracy of 55-56%. This seems low for a model with so much data, but it makes sense in the context of the situation, because in basketball even under the best circumstances with the very best players, they will still miss shots occasionally.

To further refine our understanding of the data, we built a C5.0 Decision Tree, which provides us with a table of rules that the model follows when determining if the shot was made or missed in which we can manually calculate the accompanying Support and Confidence. The Predictor Importance model showed that almost 95% of the decision on if the shot is made or missed is determined by shot distance and the defender distance. Wanting to understand the implications of this information, we decided to take it a step further and apply what we know now on individual player comparisons. We looked at two well known former players, Manu Ginobili and Tony Parker, to determine how our model would evaluate their shots with respect to their specific statistics. When evaluating the results, it was very interesting because we could see clear trends in the individual player's style and how to optimize the use of player's strengths in a

game setting. Indicators showed how Parker was a more optimal shooter when close to the basket and was not influenced as much by defender proximity, while Ginobili was a better open look shooter. Knowing this information, if you were the San Antonio Spurs head coach, you can strategize to try and place your players in these situations that would be beneficial for the team. Overall, we were able to successfully dissect an interesting dataset to discover impactful trends, which is exactly what is happening currently in professional sports. Data analytics such as the tools utilized throughout this report have been allowing sports teams to become more efficient in selecting players for their rosters who will fit their team's needs as well as putting players in their optimal situations to succeed.

References

*2016-17 NBA Player Stats: Totals*. (n.d.). Basketball-Reference.com. Retrieved April 27, 2023, from https://www.basketball-reference.com/leagues/NBA_2017_totals.html#totals_stats::pts

Baghurst, T., Lackman, J., Drewson, S., Spittler, P., Turcott, R., Smith, M., Illescas-Marquez, G., & Boolani, A. (2021). A hot mess: Basketball coaches' perceptions of ability versus actual performances of their athletes. *AUC Kinanthropologica*, *57*(1), 11–25. https://doi.org/10.14712/23366052.2021.3

Brown, D. T., Link, C. R., & Rubin, S. L. (2017). Moneyball after 10 years: How have major league baseball salaries adjusted? *Journal of Sports Economics*, *18*(8), 771–786. https://doi-org.marist.idm.oclc.org/10.1177/1527002515609665

Davenport, T. H. (2014). Analytics in sports: The new science of winning. *International Institute for Analytics*, *2*, 1-28.

Duquette, C. M., Cebula, R. J., Mixon, F. G. (2019). Major league baseball's *Moneyball* at age 15: A re-appraisal. *Applied Economics*, *51*(52), 5694-5700. https://doi.org/10.1080/00036846.2019.1617399