**Predicting NCAA Shots**

Chris Watkins

May 15, 2021

# Introduction

* 1. **Background**

College sports provide a starting point for young athletes on their way to professional careers. One of the major leagues in this category is the NCAA Men’s Basketball League. Player performance in a game is dependent on many factors; some tangible and some not. However, there are some concrete statistics we can utilize to make some informed predictions.

**1.2 Problem**

We want to determine if a shot will go in based on the circumstances surrounding the shot – not necessarily factoring in the player’s skill. We aim to predict the likelihood of a given shot scoring based on only factors that are a circumstance of the game.

**1.3 Interest**

This analysis could be of interest to sports analysts, coaches, and players for use in improving shot selection and running defensive strategies to limit high percentage shots. It could also be of interest to fans for purposes of fantasy basketball and betting.

# Data Sourcing and Cleaning

* 1. **Data Source**

The entirety of the data is available as a public data source in Google’s BigQuery platform which can be found [here](https://cloud.google.com/bigquery/public-data). The specific table used “mbb\_pbp\_sr” which contains play by play event data for each match. For our usage, we select the most recent year of data available and use a subset of one million rows of event data.

* 1. **Cleaning Data**

The data was already in a relatively workable state. The data was filtered for only the event types pertaining to shot attempts, and then selected our target year of 2017.

The data in a categorical format, containing such features as “shot type” which can have a value such as “layup” or “jump shot”. Thus, we need to create dummy variables for every feature we want to be used in our algorithm.

Another datapoint that was transformed was the (x,y) coordinates of the shot. The distance from the net was calculated and used in place of these variables.

Since the data was brought in through an SQL query, there is no need to grab features we do not plan to use. For this reason, we select only features we think will impact the prediction:

* tournament
* tournament\_type
* team\_basket\*
* away\_name
* home\_name
* period
* elapsed\_time\_sec
* team\_name
* event\_coord\_x
* event\_coord\_y
* event\_type
* shot\_made
* shot\_type\*
* shot\_subtype\*
* points\_scored
* three\_point\_shot

\* later converted to dummy variable

**3.3 Data Exploration**

First, we can take a look at all of the made shots mapped out:

A picture containing chart

Description automatically generated

Figure 1. Location of made shots.

There are very few shot attempts from just inside the three-point line as there is high risk and low return.

We check to see if there are any immediate and strong correlations in the features:

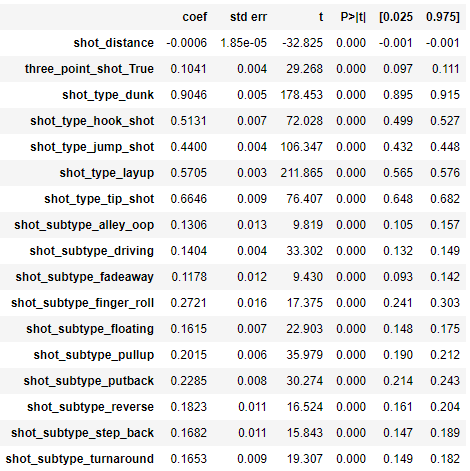


Figure 2. Summary of coefficient and P values.

As expected, some shot types such as “dunk” and “tip shot” are strongly correlated with a shot going in. We also observe that the distance from the net has less correlation than expected.

# Application of Models

* 1. **Benchmarking - Logistic Regression**

Logistic Regression was used as a benchmark for other models throughout testing. Upon first run, having not factoring the shot distance in, the maximum accuracy achieved was 57.45%.

After calculating and adding shot distance into the model, we were able to increase the accuracy to 63.64% using 500,000 rows.

In the final run, using 1 million rows, we achieved a precision of **64.47%**

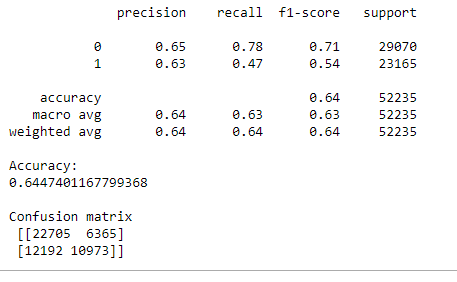


Figure 3. Accuracy Metrics for Logistic Regression Model

* 1. **Worst Performing - Random Forest Classifier**

This model was implemented as it is known to be accurate for classification problems on large sets of categorical data with high dimensionality.

With 1 million rows, we were able to achieve a precision of **59.20%**

It was observed that this model had a much higher false positive rate than the others, which is evidenced in the confusion matrix.

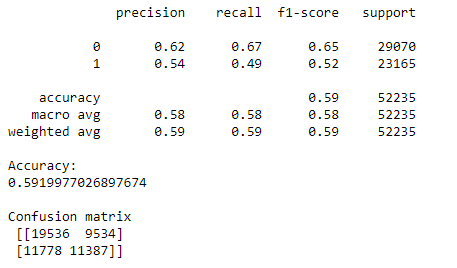


Figure 4. Accuracy Metrics for Random Forest Classifier

* 1. **Summary of Model Performance**

Overall, logistic regression remained the best performing model of the group.

This is a result of it having the lowest false negative rate of all the models, which proved to be the biggest source of inaccuracy in general.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Logistic Regression** | **Random Forest Classifier** | **Neural Network** | **Gradient Boosting** | **K-Nearest Neighbours** | **Support Vector**  **Classifier** |
| **Accuracy** | 64.47% | 59.20% | 64.43% | 63.79% | 64.33% |  |

# Conclusion

In summary, we were able to produce a model that is able to predict the if a basketball shot in the NCAA 2017 season will go in or not with an accuracy of 64.47%. To do this, we used one million rows of event data from NCAA games to build a classification algorithm that can tell us if the shot will go in or not based on such things as shot type, distance from the net, and time left in the game.

This could help coaches and players identify key shooting areas where they are most likely to have success. It also can be applicable to fans and hobbyists interested in sports statistics.

# Opportunities for Improvement

We were able to increase the accuracy by over 10% from our initial diagnostic run – however there is still a lot of inputs that would help to improve the accuracy. One noticeable gray area in this dataset is any statistical information on the shooter – this is intentional in the current state, as it is more of an analysis of shots based on position and situation within the game as opposed to a measure of the players skill. However, if we wanted to increase the predictive power, incorporating player statistics (namely age, shooting percentage, position) would likely have a large positive impact on the model’s accuracy.