Problem Statement

Being such a popular sport, basketball attracts a great deal of attention from players and coaches worldwide. With so much competition in basketball within club, high school, college, and pro teams, players should always be finding ways to edge out their competitors, whether this means winning against opposing teams or beating out teammates in playing time. Because skill and athleticism play a huge role in determining the outcome of a game of basketball, basketball decision-making, often dubbed basketball IQ, can often be overlooked. A major aspect of this decision-making includes shot selection, or knowing when and how to take a shot in a game. As a result, placing a greater focus on shot selection can give coaches and players alike an advantage when facing whatever competitors they have within basketball.

Use Case

By inputting in the different context of a shot, such as shot distance, closest defender distance, and touch time and amount of dribbles before the shot, we can calculate a shot value. This information can be used by by coaches deciding on what specific plays to run to maximize the optimal shots being taken. A coach can record all the shots by his players in a game and see the most frequent context of a shot. A specific shot coaches can be looking for are those with a low touch time and 0 dribbles, indicating a catch and shoot, a distance of 24 feet, which is just outside the 3 point line, and a closest defender of 10 feet, indicating a fairly open shot with a high shot value. Then the coach can adjust his playbook to incorporate more plays that result in high value shots.

Dataset

"NBA Shot Logs", retrieved from https://www.kaggle.com/dansbecker/nba-shot-logs, is a 128k by 21 dataset on Kaggle that gives information about the context of 128 thousand shots taken in the NBA. This includes the player who took the shot, the amount of dribbles before the shot, touch time before the shot, shot distance, whether it was a 2 or 3, closest defender and his distance to the shooter, and whether the shot was made or not. This dataset contains 14 numeric columns and 7 non-numeric columns, although some of the numeric columns may be better used as a categorical column. 0 missing data, but shot_clock has na variables.

To judge shot selection we would be best off by trying to predict the shot value, or how much a shot is worth in the long run. At first glance, it may seem like the column stating whether the shot was made or not would be a useful column to predict. However, it would not completely encapsulate a shot efficiency due to the fact that some shots are worth more than others. Taking more 3 point shots will still produce more points than taking 2 point shots even though there is a lower probability of a 3 point shot being made. Therefore, we would have to look at both in conjunction to best determine a shot value, given by the column with resulting points of the shot taken.

Data Wrangling

After an initial glance at my dataset, I found three different columns that contained mistakes or missing values. The shot_clock column had some missing values. The touch_time column had negative values, and it is literally impossible for this to happen, so these values would have to be a mistake. Lastly, some of the observations in the points_type column did not match up to the shot_distance column. For example, some of the 3 point shots had a shot_distance that would be too close to the rim, while some of the 2 point shots had a shot_distance that would be too far from the rim.

Because of the fact that any observations containing only one of these issues can be pretty easily remedied, I opted to not drop all observations where any of the issues are present. However, if any observations contained more than one of these issues, then the whole observation would probably become too unusable for the final model. Therefore, I opted to drop all observations with at least two of these issues, which ended up being 80 observations.

The resulting data frame contains observations with only one of these issues, which can be fixed relatively easily. The first column I fixed was the touch_time values with negative values. To fix this, I simply changed all the negative values into a 0 (although technically, it is still impossible to shoot a basketball with a 0 touch time. But, as touch time is a continuous variable, it is impossible for any value to be exactly any number, as there would be far more decimal places than can be tracked. So this 0 would just represent a touch_time so short that it is practically 0).

Next, I chose to impute with the mean all the missing values in the shot_clock column. As the range of the shot_clock can only from 0 to 24, there should be no outliers, making mean a valid choice over the median. Also, touch_time is a continuous variable, so using the mode would not be ideal.

The final column that needed fixing was the shots_type column, as it had discrepancies with the shot_distance column. For the observations with discrepancies between the shot distance and the points type, it is unclear whether the discrepancy comes from a mistake in points type, or if it comes from a mistake in shot distance. I decided to change the points type to match to the shot distance because, given a shot distance in a certain range, you can find out exactly how many points the shot is worth (unless the distance is between 22 and 23.75, but those observations were not counted as errors).

As the column to be predicted is binary, there were no outliers that needed to be dealt with.

Initial findings

My cleaned dataset ended up having 9 columns that can be used to predict whether or not the shot was made. I started examining my data by first comparing the percentage of shots made for 2 pointers and 3 pointers. I used a barplot for it because it is an easy way to see the difference between the two kinds of shots. As expected, 2 point shots have a much higher shot percentage (at about 49%) compared to 3 point shots (at about 35%). A bar graph is helpful in this situation because it allows us to visually see that 3 point shots are made about 67% times as much as 2 point shots. With my dataset as large as it is, a t-test is not needed to know that this difference is not only statistically significant, but also practically significant.

Interestingly, despite the disparity in percentage of shots made, the 3 point shots would result in more points in the long run. These shots would result in an average of 3 * .35 = 1.05 points per shot, while the 2 point shots would result in an average of 2 * .49 = .98 points per shot.

Afterwards, I decided to create scatterplots for my variables that have many different values, as scatterplots with a few values would not make sense. These variables are my continuous variables, shot_clock, touch_time, shot_dist, and close_def_dist, and dribbles. I opted to not use game_clock as it is it generally would not affect a shot too much unless the game clock is very low.

With so many data points, it is hard to see every single point for both kinds of shot results, but it can be assumed that there are many points where the shot was not made within predominantly orange sections. The sections that are predominantly blue indicate shots with much more misses than makes, and these are the shots that players and coaches should generally avoid taking.

A few things to note are:

1. Dribbles and touch time before the shot seem to have a strong correlation, which makes sense as taking more dribbles naturally increases your touch time. Of note are the shots with no dribbles but a long touch time. These kinds of shots signify a player holding the ball for a while without moving, leading to stagnant offense and out of rhythm shots, so it makes sense that these shots are predominantly missed.

2. Shot_clock seems to very closely follow a normal distribution. Also, it seems that shots taken when the shot clock is close to 0 has a much lower chance of going in. This makes sense as these last second shots are typically very rushed.

3. Not surprisingly, the greater the shot distance, the lower the chance that the shot will go in. This is shown by the fact that the bar plot has a longer orange (made shots) bar with a lower distance and a longer blue bar with higher distances.

4. The shots that are taken where the distance of the closest defender is very large seem to have a very high shot percentage. This makes sense as these cases are usually when the opposing team turns the ball over so the other team is able to dribble down the court with no defenders nearby, leading to very open shots.

Afterwards, I proceeded to examine the rest of the variables, which are shot number and period. I proceeded to use a line plot for shot number and a bar plot for the period and found the following:

1. The shot number generally does not affect whether or not the shot was made up until maybe around shot 19. This could be due to the fact that there is so much less data for shots after a certain point. After shot 19, the percentage of shots made seems to vary more drastically, and even more so after shot 31.

2. The percentage of shot made does not vary too much within the first 4 quarters, though the 4th period seems to have lower shot percentage. The overtime periods (5,6, and 7) all have lower percentages, with periods 5 and 7 being much lower. Their total amount of shots taken at 911 and 43 represents a decent size. Furthermore, in an overtime period, players are much more tired compared to the first 4 quarters, so their shot percentages should naturally drop.

After doing some exploratory data analysis, I am diving deeper into the data for more information on what may have the biggest impact on shot percentage. To do so, I am plotting the variables from the scatterplots against shot percentage. Because the majority of these variables are continuous, I am going to 'bin' them using pandas' 'cut' function. Furthermore, I am looking at how many observations contain outliers an any of the variables, as variables with huge values may affect the significance of any visual trends seen.

By looking at all the plots, I see that the variables with extremely obvious effects on shot percentage are points type and shot distance. In the plot of shot percentage by shot distance, I see an obvious decrease in percentage as the distance increases. This decrease is large enough to be significant, even with the small sample sizes at larger distances. The bar plot of points type against shot result also shows an obvious difference.

Other variables have an effect on shot percentage, though they are less obvious. These variables are shot number, closest defender distance, shot clock, dribbles, and touch time. Plotting these variables against shot percentage, for the most part, show slight changes in shot percentage for the most part, and then some drastic changes the higher the variable becomes. These shot percentage changes are so drastic mainly because of the low frequency of observations with variables at that high magnitude and, therefore, is not enough evidence to prove statistical significance. However, the variable as a whole still has an impact on the shot percentage.

The only sense of correlation between independent variables occur with touch time and dribbles. There is definitely a strong relationship between touch time and shot result, which could present troubles in a final model. However, the correlation is far from perfect, and there are plenty of observations with a long touch time, but no dribbles, so it would be ok to use both in a final model.