MONEYBALL

NBA STATISTICAL ANALYSIS

By: Reynaldo Cortez

Introduction

The most fitting application of data to investigate was how to apply data science to my favorite sport, basketball. My source of inspiration was the movie *Moneyball*, where a general manager decides to build his team, not by intuition or the "eye test", but instead by selecting players that have statistical qualities that will produce wins.

Business problem

This process could be implemented by any sports team who wants to find players that impact winning and ideally are not correctly under a big contract, especially the NBA in this case. Teams can try to attain these talented individuals at a "bargain" price and allow flexibility to attain even more talent. Budget management is key to the success of an NBA basketball team.

NBA free agency for the 2021-22 season began on August 2nd, 2021, where teams officially negotiated deals with players after the NBA's announcement of the Salary Cap being set at $112.414 million; the salary cap is the total limit the clubs can spend on all player salaries and benefits. The goal for every team is to fill the roster's total compensation under the salary cap, balancing your team's overall talent level from the superstars down to the rookies.

Data

To begin my investigation, I attained two main datasets from Kaggle and GitHub that gave me important information on every NBA game played for the last 20 years.

This data contained the following attributes, which were evaluated as my independent variables when modelling, to predict my dependent variable: wins.

- **FG_PCT:** Field Goal Percentage, mathematically expressed as: $\frac{\text{Total Shots Made}}{\text{Total Shots Taken}}$
- **FG3_PCT:** Thee-Point Field Goal Percentage, mathematically expressed as: $\frac{\text{Total Three Point Shots Made}}{\text{Total Three Point Shots Taken}}$
- **FT_PCT:** Free-Throw Percentage, mathematically expressed as: $\frac{\text{Total Free Shots Made}}{\text{Total Free Shots Taken}}$
- **REB:** Rebounds. How many times a team grabs possession of the ball after a missed shot
- **AST:** Assists. How many passes a team makes that lead to a shot made
- **STL:** Steals. How many times a team takes the ball from the opposite team
- **BLK:** Blocks. How many times a team blocks a shot from the opposite team
- **TO:** Turnover. How many times a team losses possession of the ball by their own mistake

The Process and Models

Using the data collected, I prepared different models that could accurately predict the result of a game depending on these statistical figures. The models created were:

- **Logistic Regression:** Accuracy on test set was 91.29%
- **KNN:** Accuracy on test set was 88.56%

- STL: Accuracy on test set was 78.69%

From the logistic regression model, I was able to ascertain which statistical figures where the most important by examining their coefficient weight. The model demonstrated that the top five most important statical figures, listed from highest to lowest, were:

1. Field Goal Percentage
2. Turnover
3. Rebounds
4. Three-Point Field Goal Percentage
5. Free-throw Field Goal Percentage

Armed with this information, I began my search for the players that possessed these attributes with machine learning to help me cluster the players and find trends. Since I am interested in results that are applicable in present day NBA, only using recent statistics to group players was necessary. I considered clustering players based on their average stats in the last available season in my dataset. However, I opted to use the average of the last three seasons to eliminate recency bias from my results that ultimately would bring down the average of players that had only one good season. I did not want to highlight players that haven't proved that they can be consistent yet.

After sorting the players table to contain the average of the last three season, I derived the optimal number of clusters that the players belonged in.

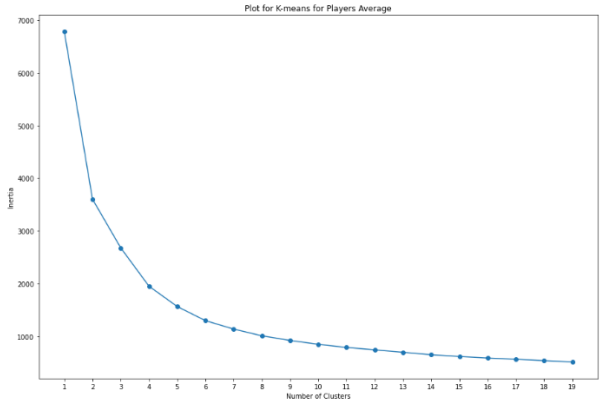

Figure 1: Cluster Inertia Score

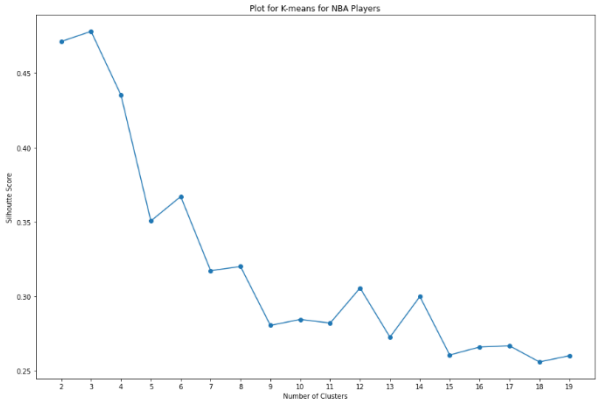

Figure 2: Cluster Silhouette Score

The most appropriate cluster number was 6, as it was at the elbow of the inertia graph (Figure 1) and also a high peak in my silhouette scores (Figure 2). Here's a 2D graphical representation of the clusters:

Figure 3: 2D Graphical Cluster Interpretation

The cluster groups helped find the average performance of each group (Table 1).

Table 1: Cluster Player Performance

CLUSTER PLAYER PERFORMANCE								
Cluster Type	FG_PCT	FG3_PCT	FT_PCT	REB	TO	AST	STL	BLK
0	0.39	0.29	0.40	2.39	1.28	2.85	0.69	0.21
1	0.55	0.17	0.62	10.56	2.12	2.79	0.82	1.17
2	0.44	0.31	0.63	4.78	2.58	5.70	1.20	0.38
3	0.39	0.19	0.3	3.17	0.70	0.97	0.47	0.35
4	0.29	0.15	0.17	1.12	0.44	0.59	0.28	0.12
5	0.47	0.24	0.47	5.65	1.25	1.71	0.70	0.71

Interpretation

The prominent groups in my opinion were cluster 1, and cluster 2. Cluster 1 had the highest field goal percentage and rebound percentage, and cluster 2 had the highest free throw percentage and three-point field goal percentage (highlighted in blue). I omitted turnovers (highlighted in orange) in this discussion as I believe a product of a good player is having possession of the basketball for longer period of times and thus statically being more prone to turn the ball over. These two clusters also contain the good values for each of the remaining categories.

When further examining the names of players cluster 1 and 2, I found even more fascinating insights. Cluster 1 is filled primarily of centers which are the tallest guys in the NBA. A high field goal percentage is usually on par with the fact that centers take most of their shots closer to the basket since they can shoot over their opponents more easily. Cluster 2 is filled primarily with “superstars”, the players that are currently the highest paid players in the league and are volume scorers.

My next evaluation introduced average salary for the last 3 seasons and ranked the players clusters 1 and 2 based on salary. I can now accurately give a recommendation to an NBA manager of players that are currently underpaid but are performing just as well as superstars.

Conclusion

Most people could easily spot a star player, but data science analysis allows for a team to create a balanced roster within the budget by transferring the focus from the superstars, to composing teams with good performance, allowing a lot flexibility to a team’s management. *Moneyball* was based on a true story, so this method is proven as by the real success of the Oakland Athletics.

I feel confident that with further analysis, I could take this a step further by analyzing specific talents or basketball positions. I could separate these categories by defensive and offensive stats or look solely at specific needs that a team currently lacks on.