Jacob Graham
Mechanistic Data Science for Engineering

Homework 2 - Real-life Problem Proposal

Predicting the assigned position of National Basketball Association athletes based on primary statistics

Introduction

The National Basketball Association is a professional sports league consisting of 29 teams based out of major cities in the US and 1 team in Toronto, Canada. In 2020, it was the third most watched sport among the US's five major sports leagues (football, baseball, basketball, soccer, and hockey). Each team is allowed to carry up to 15 players on their active roster and 5 players are on the court at a given time during the game. Players on the court are assigned to one of the following positions: small forward (sf), power forward (pf), center (c), shooting guard (sg), and point guard (pg). The assigned role gives a player an idea of what their positioning should be relative to the basket and coincides with unique expectations of that player on the offensive and defensive ends of the court. For example, a typical center is very tall and spends most of their time close to the basket. For this reason, their offensive stat lines are packed with two point field goals, attempted free throws, and rebounds, while they log a disproportionate number of fouls and blocks on the defensive end since they often are the last line of defense for opposing players driving to the basket. Although most teams follow conventions such as height, shooting percentage, and ball handling skills to assign positions, there are no restrictions regarding where or when a player can handle the ball based on their listed position. This makes basketball unique to the four other major sports which each have at least one special position that must follow particular rules. As a result, the true role of each position is subject to considerable interpretation based on each team's play style.

There are eight major statistics that each player can log during a basketball game which include steals, points, blocks, field goals made/field goals attempted (percentage), free throws made/free throws attempted (percentage), three pointers made, rebounds, and assists. Other statistics including height, points per attempt, three point percentage, offensive/defensive plus/minus, usage, turnovers, and others can also be defined. Some of these statistics are more obviously correlated with position than others.

For this project, I would like to apply a mechanistic data science approach to create an algorithm that can predict the assigned role of a player based on critical statistics, each of which can be preferred to as a mode in this context. Such an approach would offer insight into the breadth of interpretation that a player can make based on their assigned role on the court.

Project Outline

Module 1: Fortunately, basketball statistics are extremely well documented and are used by coaches and management to inform trade decisions and strategy. These statistics are also used for projecting player stats for fantasy sports leagues and sports betting. Basketball-reference.com is a vast and thorough compilation of decades worth of statistics, which are available for free. This will be an excellent source of data for the 8 primary statistics, but also player positions, height and other potentially useful statistics.

Module 2: The first step in analyzing data after it is collected is to run simple analyses to identify correlations between certain statistics and position. Since the ultimate prediction will be one of five discrete values, statistical distributions will be made for average and standard deviation of each statistic per minute on the court for each of the five positions. Once these plots have been made, average per minute for each statistic can be compared between the five positions to determine correlations of statistics to certain positions.

Module 3: Once the initial analysis is complete, the most critical statistics will be identified to reduce the complexity of the problem.

Module 4: Because there are no underlying scientific principles that have been identified to determine a player's position based on primary statistics, multiple mathematical models relying on subsets of data will be developed and tested for their efficacy in predicting position.

Modules 5 and 6: A more in depth regression analysis will be conducted based on critical statistics derived in modules 2 and 3 and useful principles resulting from module 4. Machine learning will be used to identify not only direct correlations between individual variables, but will also identify subsurface trends between groups of variables that will allow for more accurate differentiation of each position.