NBA Player Score Prediction

CS 439 : Data Science Course Project Proposal

Ryan Donofrio, Shreya Vyas

Problem Statement:

Can we accurately predict an NBA player's points per game in the next season based on their current and past performances?

In professional basketball, evaluating player performance is important for teams, analysts, and fans. Predicting a player's future performance can help teams make better decisions regarding trades, contracts, lineup adjustments and winning impact. This project falls under the supervised learning project category, as it aims to develop a predictive model that forecasts an NBA player's points per game (PPG) for the upcoming season based on past performance data. This project will focus on basic data analysis and visualization techniques covered in class. By using historical player statistics, we seek to determine the key factors influencing future performance and provide a data-driven method for making better informed predictions.

Strategic Aspects:

This project will leverage fundamental data science techniques, including data cleaning, visualization, and predictive modeling to analyze player performance. By understanding statistical patterns in player data, we can identify significant trends that impact future performance. This approach aligns with the data-driven decision making methods covered in class, demonstrating how data science can be applied to real world sports analytics. Furthermore, the project will focus on developing an accessible and interpretable model, ensuring that insights can be easily understood by basketball analysts and fans alike.

Novelty and Importance:

Accurate forecasting of player performance is valuable for team management, fantasy basketball players, sports analytics, and team general managers. While existing studies focus on predicting team success, individual player performance prediction is a more nuanced and challenging problem due to the influence of factors such as team dynamics, injuries, roster needs, and playing time adjustments.

Prior research in sports analytics has primarily used basic statistical models to evaluate player efficiency. However, many of these models fail to incorporate a broad range of player statistics and contextual factors. This project seeks to improve upon existing approaches by incorporating multiple variables, such as shooting accuracy, playing time, and historical trends, to enhance prediction accuracy. By creating a model that captures these complexities, we aim to provide valuable insights into what contributes to a player's future performance. This can not only help NBA scouts look for new up and emerging talent among the younger players, but also NBA GM's who are looking for the perfect fit to take a mediocre team to the next level by incorporating an able player perfect for the needed role.

Plan:

Data Collection and Preprocessing:

The dataset for this project will be sourced from publicly available NBA statistics databases, including Basketball Reference, the NBA API, and Kaggle NBA datasets. These sources provide comprehensive player statistics from past seasons, including points per game, assists, rebounds, field goal percentage, three-point percentage, free throw percentage, turnovers, steals, and blocks. Additional contextual factors, such as player age, years in the league, team performance, and injury history, will be included to provide a holistic analysis of player performance trends. Once the data is collected, it will be cleaned and preprocessed to ensure accuracy and consistency. This process will involve handling missing values, standardizing formats, and normalizing relevant statistics. Exploratory data analysis will be conducted to identify key trends and correlations within the dataset. Data visualization techniques, including scatter plots and bar charts, will be used to illustrate relationships between different factors and future player performance.

Model Selection and Training:

To predict a player's future performance, we will begin by using simple regression models, starting with linear regression to figure out a baseline prediction. Decision trees may be explored to capture nonlinear patterns inside the data. Feature engineering techniques may be used, such as for calculating the rolling average of player statistics over multiple seasons and categorizing players based on their position to identify position-specific performance trends. The model's performance will be evaluated using standard metrics such as mean absolute error and mean squared error. These metrics will help determine prediction accuracy and determine how well the model is generalized to unseen data. The model will be trained using past player data and validated by comparing predictions with actual player statistics from subsequent seasons.

Implementation and Evaluation:

Implementation of the project will follow a structured approach. The first step is data collection and cleaning, ensuring all relevant statistics are properly formatted. Next, exploratory data analysis will be performed to see relationships between different variables. Once the data is prepared, the model will be trained using regression techniques, followed by testing and validation on past season data. The final predictions can be compared with actual player performance to assess accuracy and reliability. We will measure success in this project by seeing how closely the model's prediction will align with actual next-season performances. For example, if we use the 2021-2022 season's player data for testing, we can see if the model accurately predicts player performance for the 2022-2023 season. A well performing model should have a low error rate, demonstrating its ability to make reasonable predictions. Additionally, insights gained from feature importance analysis will provide valuable takeaways about the factors that most significantly impact player performance. By applying fundamental data science techniques in a real world sports setting, this project can offer both practical insights and an engaging learning experience for us in the predictive analytics field. The results can be useful for teams, fantasy basketball players, basketball scouts, NBA general managers and front office staff, and analysts looking to better understand player development and performance trends.