Capstone Final Report

NBA Player Performance Prediction Project

Carl Riemann

September 20, 2024

Basketball has been a revolutionary sport for a long time. One of the perks about this sport being so big, is that there are several opportunities to be a part of the basketball phenomena, without having to play basketball. A fun way to enjoy basketball outside of the game is by participating in fantasy leagues. What are fantasy leagues? A fantasy league is a game where participants select any current player to create a team and compete based on the selected players' performance in real games.

The goal of this project is to develop a model that accurately predicts each player's points for 2025 season based on NBA player performance metrics. These predictions will then provide an advantage for fantasy league players by helping them draft the most promising players and optimize their lineups.

The datasets used for developing the predictive model include player information on all NBA regular season matches played from the 2019-2020 season through the 2023-2024 season. (excluding playoffs). Each dataset consisted of individual player statistics from each game during the regular season, including Player, Age, Games, GamesStarted, MatchesPlayed, FieldGoals, 3Points, 2Points, 2PointsAttempted, etc.

To create a predictive model for season 2025 there was important data wrangling to be done. This process involved reshaping the data to create new columns that capture player performance across the previous two seasons. Specifically, I created columns to track statistics for each player from the 2022-2021 and 2023-2022 seasons.

```
Index(['Player', 'Age', 'Age_last_year', 'Age_year_before_last', 'G',
       'G_last_year', 'G_year_before_last', 'GS', 'GS_last_year',
       'GS_year_before_last', 'MP', 'MP_last_year', 'MP_year_before_last',
       'FG', 'FG_last_year', 'FG_year_before_last', 'FGA', 'FGA_last_year',
       'FGA_year_before_last', 'FG%', 'FG%_last_year', 'FG%_year_before_last',
       '3P', '3P_last_year', '3P_year_before_last', '3PA', '3PA_last_year',
       '3PA_year_before_last', '3P%', '3P%_last_year', '3P%_year_before_last',
       '2P', '2P_last_year', '2P_year_before_last', '2PA', '2PA_last_year',
       '2PA_year_before_last', '2P%', '2P%_last_year', '2P%_year_before_last',
       'eFG%', 'eFG%_last_year', 'eFG%_year_before_last', 'FT', 'FT_last_year',
       'FT_year_before_last', 'FTA', 'FTA_last_year', 'FTA_year_before_last',
       'FT%', 'FT%_last_year', 'FT%_year_before_last', 'ORB', 'ORB_last_year',
       'ORB_year_before_last', 'DRB', 'DRB_last_year', 'DRB_year_before_last',
       'TRB', 'TRB_last_year', 'TRB_year_before_last', 'AST', 'AST_last_year',
       'AST_year_before_last', 'STL', 'STL_last_year', 'STL_year_before_last',
       'BLK', 'BLK_last_year', 'BLK_year_before_last', 'TOV', 'TOV_last_year',
       'TOV_year_before_last', 'PF_last_year', 'PF_year_before_last', 'PTS',
       'PTS_last_year', 'PTS_year_before_last', 'Points_scored_next_season',
       'Tm', 'Season', 'SF', 'C', 'SG', 'PG', 'PF'],
      dtype='object')
```

Figure 1. Columns transformation

The transformation was done using pandas lead and lag function and the primary goal was to generate a Points_scored_next_season column that contained the points per game (PTS) for the 2023-2024 season. This column will be the target variable for predicting, enabling the prediction of Points_scored_next_season (the player's points in 2024-2025) based on historical data from the 2022-2023 and 2021-2022 seasons.

During the exploratory data analysis (EDA) phase, I looked at the relationships between the variables and the target variable, Points_scored_next_season. I generated pair plots and scatter plots to identify potential outliers and relationships between features, histograms to assess the distribution and skewness of the data, and heatmaps to visualize the correlation between variables.

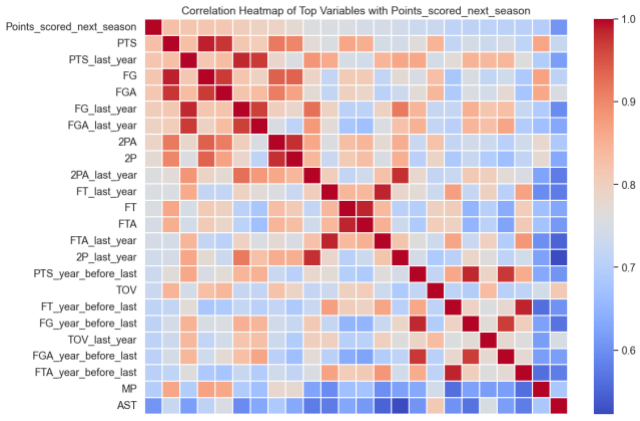

Figure 2. Heatmap of correlations to feature variable.

For the last part of my project (Pre-processing and Modelling) I built five machine learning models to predict Points_scored_next_season and compared them using Mean Squared Error (MSE) and R-squared.

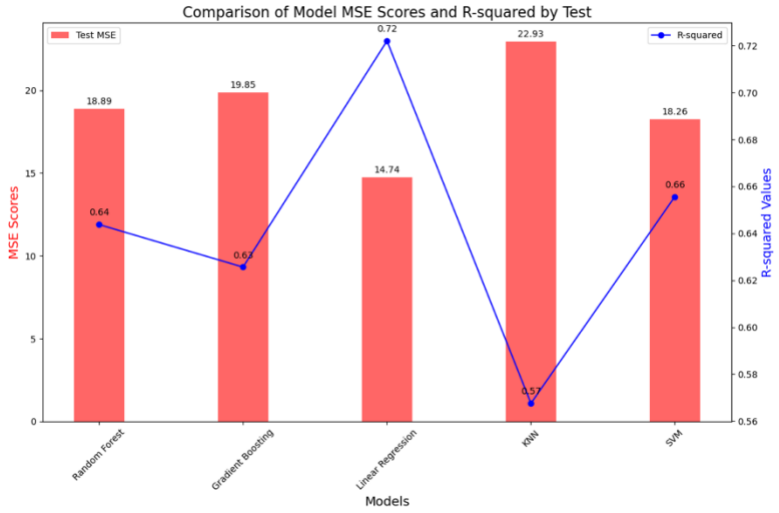

Figure 3. Model Metrics

Linear Regression had the lowest Test MSE (14.74) and the highest R-squared (0.722), indicating that it predicted the next season's points more accurately and explained the variance in the target variable better than the other models.

To conclude I used the model to predict Points_scored_next_season for every NBA player. These predictions were then stored in an Excel file, projecting the total season points for all players in the 2025 NBA season. This output is valuable for fantasy league players, and fantasy league owners planning their strategies for the upcoming season.