NBA Game Match-up Prediction Tool

Team 52

Yotan Demi-Ejegi, Diego Escalera, Ethan Garbow, Dan Goldin, Sebastián de la Hoz and Michael Lamontagne

INTRODUCTION & MOTIVATION

According to ESPN, NBA Teams and the sports betting industry spend well over $3.1 billion in research surrounding player stats and predictions to gain a competitive advantage in the respective fields.

Our objective is to provide a tool for fellow NBA fans, coaches and teams to do a deep dive into the game and be able to look at specific player matchup comparisons and uncovered intelligence using social sentiment.

THE APPROACH

Our modeling approach focuses on the individual player and understanding their performance against teams and whether they are at home or away.

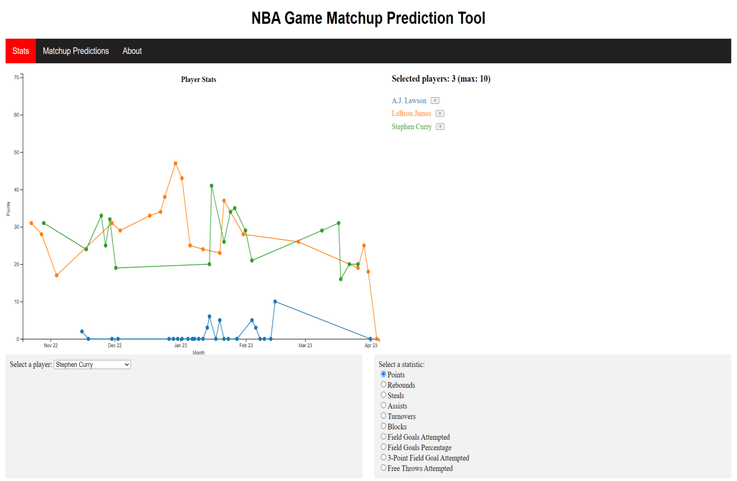
For each player, we completed the following tasks:

- Ran a multiple linear regression with the players' points per game as dependent variable and all teams and all players as categorical variables using historical and recent data of the player against each team they played again
- We analyzed past 7-days of social sentiment via Twitter to create a sentiment score using a -50 to 50 score range for positive, neutral and negative
- Using this model, we estimated points per game for each player
- For an upcoming game, we aggregated the players points for each team to determine the winner

Regression Model:

$$\text{Player}_A\text{Team}_A \text{ Points Per Game} = \beta 0 + WhereGamePlayed\,\beta 1 + TeamPlayed\,\beta 2 + \epsilon$$

Current prediction tools only show which team will play against which other team and show numbers to predict the most likely. We also have gone extra mile of creating an interactive display using D3 to allow users to explore and compare different players, teams, and attributes.



THE DATA

Data used to train the model and for prediction came from:

- NBA datasets from KAGGLE and NBA API. All games from 2004 season to end of 2023 season (April 2023) from API was pulled using Python code against exposed end points
- Over 26,000 games were analyzed with over 700,000 player results
- Social discussion data was captured from Twitter firehouse via the Tweepy Python Library

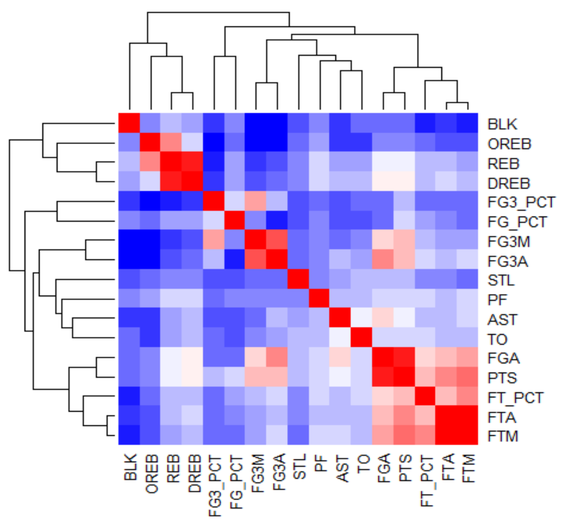
Data after collection was stored in GitHub and contains below characteristics

- Games Data containing total number of 21 variables of both continuous and categorical including game date, game id, home and away team ids, aggregate game stats for each team.
- Game Details data with 29 variables including Game ID, player ID, Team Abbreviation, Player Name, Game Statistics (Minutes played, field goal attempted and made, points, rebounds, assists etc.)
- For each player we took a sample of fifty tweets that had the following breakdown of fields like Tweet Text, Username,user_followers,attribute_score etc.

EXPERIMENT/RESULTS

As part of the data exploratory, we performed a correlation matrix for each basketball stat to identify and visualize any pattern in the data. Figure below presents the correlation matrix plot. Most variables related to the act of shooting such as: FGA, FG3A, FT, among others, are highly related to the variable of interest, points per game. Similarly, the heat map presents clearly those findings as well as the different clusters based on the correlations results. We can see that there are about 4 clusters, confirming the results of the player cluster analysis on the progress report. The 4 clusters consist of: defensive stats such as blocks and rebounds, 3-pointers, possession and shooting stats.

Since teams combine player clusters to build balanced lineups, players' performance could vary throughout season matchups. These findings helped to support using non-traditional metrics to predict output such as the team matchup.

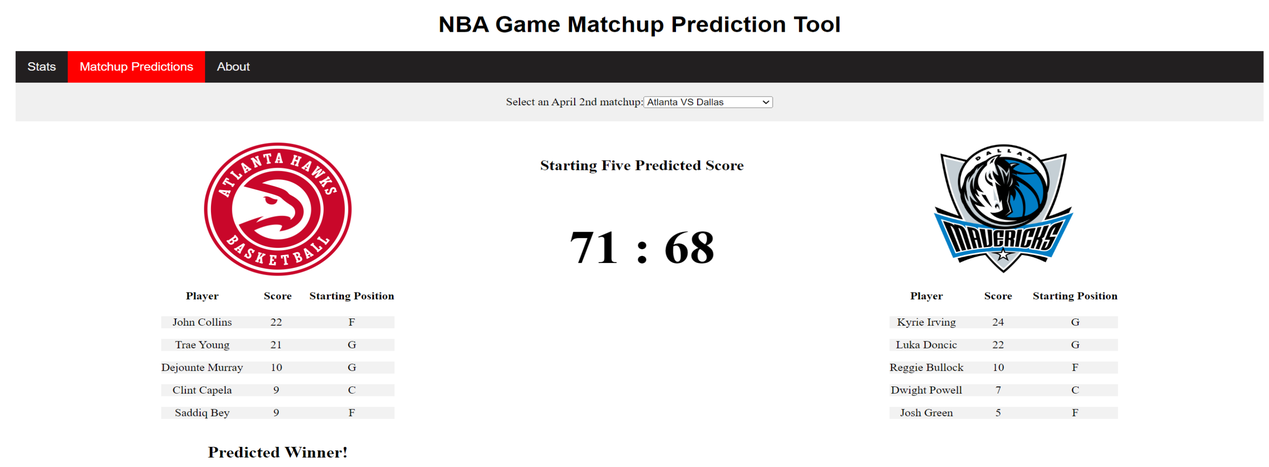


Example of model output with coefficients for each player on a team

PlayerID	1627749	203991	1629027	1630180	1628381
(Intercept)	13.2	15.4	26.1	10.4	21.8
WHERE_PLAYEDHOME_GAME	-0.7	0.9	1.8	0.4	0.3
BKN	-1.6	-3.8	-2.0	0.9	-6.0
BOS	1.6	-4.8	-5.5	5.5	-5.7
CHA	0.8	-3.7	-1.7	5.5	-5.7
CHI	0.8	-1.5	0.1	2.8	-8.2
CLE	3.5	-4.7	-7.0	2.5	0.4
DAL	-2.0	-6.8	-1.2	-1.6	-5.7
DEN	-0.1	-1.5	-1.6	7.2	-5.9
DET	-3.1	-3.3	-6.7	0.9	-0.7
GSW	-0.2	-6.3	2.5	4.7	-6.7
HOU	1.0	-5.3	0.4	8.6	-5.7
IND	-1.8	-1.7	-0.7	4.5	-6.1
LAC	-2.1	-5.3	-1.6	-1.5	-2.9
LAL	1.6	-4.7	-5.1	2.2	-7.9
MEM	-2.2	-4.2	-6.8	4.2	-7.0
MIA	-4.9	-6.2	-1.6	5.3	-8.3
MIL	3.6	-5.3	5.0	11.4	-5.3
MIN	-1.1	-0.6	-5.5	5.3	-8.4

Our modeling process was used to predict the winner of games on specific date (April 2nd). This does represent a small sample but however gives a promising outcome. We focused on the starting five players of each team and aggregated the predicted points. The teams with the highest points were predicted to be the winner. The model prediction accuracy was **83%**

GAME_ID	GAME_DATE	MATCHUP	ACTUAL WINNER	PREDICTED WINNER
22201167	2023-04-02	DAL @ ATL	ATL	ATL
22201165	2023-04-02	MEM @ CHI	CHI	MEM
22201166	2023-04-02	POR @ MIN	POR	POR
22201163	2023-04-02	CHA @ TOR	TOR	TOR
22201164	2023-04-02	UTA @ BKN	BKN	BKN
22201168	2023-04-02	WAS @ NYK	NYK	NYK



In addition, we tested 10 players to predict their points per game using the model for the games occurring after January 1, 2023. Our per game points prediction for each of the 10 players tested yields a Mean Squared Error $(\text{Prediction} - \text{Actual})^2$ in the range of 33-300 (average - 97.62).

Many NBA game prediction models have been created but focus on team results and statistics. Team-based outcome models largely ignore player performance and contributions. Player-based models largely ignore the impact of a specific matchups.

As a comparison to other models as outlined in our literature review, other models saw a prediction rate in the 70% range. Ours outperformed this rate in a limited sample.

Our model successful by focuses on player performance accounting for factors that are critical in today's NBA.