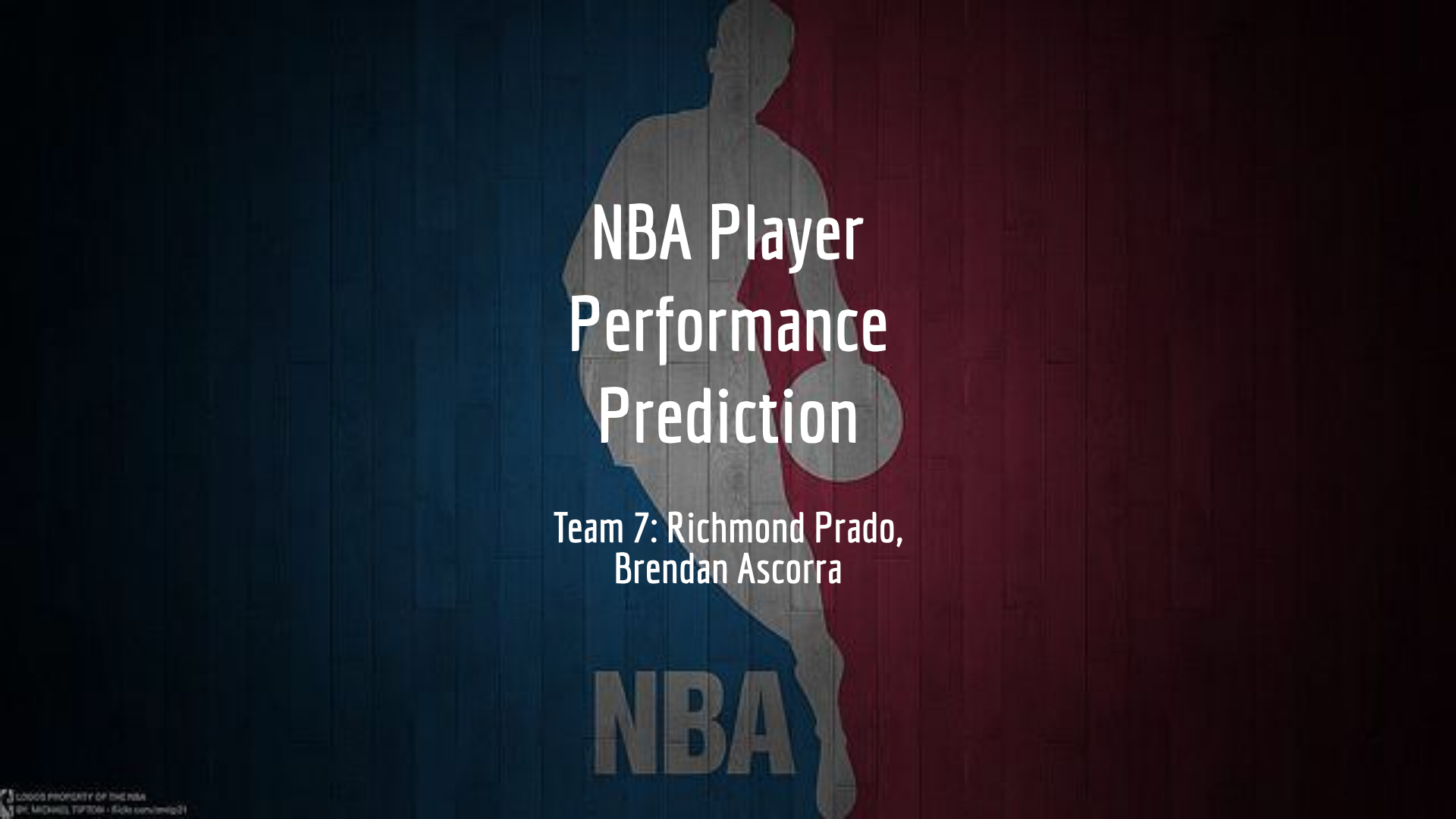
NBA Player
Performance
Prediction
Team 7: Richmond Prado,
Brendan Ascorra
NBA

Abstract

- Machine learning has been increasingly leveraged by organizations across the sports industry to predict athlete performance. These techniques have also been widely adopted in fantasy sports to forecast game outcomes and optimize strategic decision-making.
- In this study, we developed a predictive model to hypothesize the relationship between first-half performance metrics and second-half outcomes for NBA players. Using a linear regression framework, our model incorporates a range of features—including player season averages, shot selection, points, NBA team rankings, minutes played, and other performance indicators—to enhance prediction accuracy.
- Our findings suggest that first-half performance, particularly scoring metrics, holds meaningful—but not dominant—predictive value for second-half outcomes. While it is not the strongest predictor, it contributes to understanding performance trends and variability across games.

Introduction - Commonalities

- The works we researched all tackled sports related problems using machine learning algorithms and models. These works would cover:
  - Using different models and algorithms to predict the results of sports matches
    [1], [2], [3], [4], [5]
  - Going over player statistics as well as other factors for predicting player performance
    [6], [7], [8], [9], [10]
  - Using models to predict team win rate under certain factors or circumstances
    [11], [12], [13], [14]
  - Predicting player lineup performance given player statistics
    [15]

Introduction - Commonalities (cont.)

- The problems presented in these papers were handled pretty similarly.
  - They all have **rich statistical features** that they pull from [8]
  - They demonstrate the importance of **feature selection** in the sports domain [9]
  - And most papers had a **combination of different models** to improve their predictions [4], [5], [12].
- Though there were commonalities between these works, there were also differences such as:
  - A **difference in datasets** due to the different sports being sampled from or due to a difference in range of data taken
  - **Different prioritization** in feature selection

Introduction - Commonalities (cont.)

The papers and their different methods:

- Papers [2] and [8] use **decision trees** to predict team rankings and fantasy points
- Papers [7], [14], and [15] use **neural networks** for predicting player performance, sports team rankings, and basketball lineup performance
- Papers [4], [5], and [12] use a combination of **multiple algorithms** for predicting the results of basketball games
- Paper [11] uses **KNN** to predict the win rate of NBA teams
- Papers [1], [6], [9], and [10] are all papers that **compare the accuracy of different models** against each other in the realm of sports predictions

Introduction - Commonalities (cont.)

Debatable places or limitations

- Some papers claimed that they were limited by their data or features
  - NBA has years of data but other sports, like rugby, have way less
- Deciding which features were most important for their models
  - Paper [9] specifically goes more in depth about the importance of feature selection. The paper revolves around comparing different feature selection models
- Not everything is taken into account
  - According to paper [6], for cricket, there are other factors such as weather dynamics that need to be taken into account when making predictive models

Project Focus and Importance

Our work focuses on predicting individual player performance, which directly connects to broader outcomes such as team win rates, match result forecasting, and optimal team lineup decisions.

We aim to use first-half player data and other components to predict second-half performance — an area that, based on our review of existing literature, has not been extensively explored.

The significance of this means that:

- **Coaches** can make more informed decisions about player rotations and in-game strategies.
- **Players** can identify areas of weakness and target specific aspects of their game for improvement.
- **Fans and analysts** can leverage these predictions for more informed betting strategies and fantasy sports decisions.

Why A Linear Model?

- In the papers models related to sports prediction seemed to use linear regression due to how simple and easy to interpret it is. Because of this, linear regression is usually used to compare against the accuracy of other models. [1], [10], [12]
- It is also mentioned how the primary reason for using linear regression was due to how linear the relationship between features was to a target variable [10].
- So in our case, using first half metrics to predict second half metrics seemed pretty linear and we wanted to be able to easily interpret the relationship between our chosen features with our target variable.

Dataset

The primary dataset, titled *“23-24 NBA Dataset,”* was obtained from Kaggle. An additional dataset, *“23-24 NBA Dataset (New),”* comprising multiple CSV files, was also sourced from Kaggle. Furthermore, the *“23-24 NBA Rankings”* data was manually collected by scraping the NBA’s official website (NBA.com/stats) and organizing the information into a structured CSV file.

Dataset Size:

- *“23-24 NBA Dataset”* contains 232,541 rows and 29 columns, with a file size of approximately 57.28 MB.
- *“23-24 NBA Dataset (New)”* consists of 12 CSV files totaling approximately 721 KB.
- *“23-24 NBA Rankings”* includes 32 rows and 11 columns.

Dataset Preprocessing and Extraction

- **23-24 NBA Dataset**
- First_half_attempts
  - Number of shot attempts made in the 1st half
- First_half_made
  - Number of shots made in the 1st half
- First_half_points
  - Total points in the 1st half
- First_half_fg_pct
  - Field goal percentage in 1st half
- Avg_shot_distance
  - Average distance of shots taken in the 1st half

- **23-24 NBA Dataset (New)**
- MP_per_game
  - Average minutes played per game (season stat)
- PTS_per_game
  - Average points scored per game (season stat)
- AST_per_game
  - Average assists per game (season stat)
- TRB_per_game
  - Average total rebounds per game (season stat)
- eFG_pct
  - Effective field goal percentage (season stat)

- **23-24 NBA Rankings**
- Opponent_difficulty
  - Strength of opponent, calculated as `1/OVR_RANK`

Methodology

- **Data Collection**
  - Utilized the 3 datasets and extracted the important and most impactful features from each dataset
- **Feature Engineering**
  - First half metrics
    - Utilized attempts, makes, points, time periods, field goal percentage, shot distance, and 3-point percentage to create first half data metrics
  - Season-long Player Performance Averages
    - Utilized Minutes per game (MP), Points per game (PTS), Assists per game (AST), Rebounds per game (TRB), and Effective field goal percentage (eFG%)
- **Model Development**
  - Defined 13 features to predict second-half points
  - Split the data into an 80% training set and 20% test set
  - Applied StandardScaler to normalize the features to zero main and unit variance
    - So no feature dominates just because it is a larger scale
  - Trained Linear Regression model on the scaled training set

Methodology

Tools, Software, Platforms etc.

- **Pandas**
  - For loading data, filtering rows/columns, grouping data, and general data cleaning and manipulation.
- **NumPy**
  - For efficient mathematical operations, especially for array-based calculations (e.g., creating new columns with NumPy functions)
- **Scikit-learn (sklearn)**
  - To build and train the machine learning model (Linear Regression), split the dataset into training and testing sets, normalize the data (StandardScaler), and evaluate the model (MSE, MAE, R^2).
- **Matplotlib**
  - To create visualizations that compare actual vs. predicted player performance.
- **Joblib**
  - To save and load the trained model and scaler efficiently (better than pickle for large files).

The simulation was done through **Google Colab**, a Jupyter Notebook environment

Preliminary Results

By utilizing **Linear Regression**, we were able to predict the second-half point performance of NBA players based on their first-half performance.
Figure 1 displays the breakdown of the feature coefficients and their respective impacts within the predictive model.

- **PTS_per_game:**
  - The player's season average points per game is the most impactful feature.
- **First_half_attempts:**
  - Minor positive impact on second half scoring predictions
- **Avg_shot_distance:**
  - Minor negative impact on second half performance
- **eFG_pct:**
  - Minor negative impact
- **First_half_points:**
  - Has a very small positive effect on second-half performance, but is much less influential compared to shot attempts.
- **Opponent_difficulty:**
  - Shows no measurable impact on second-half scoring in this model.

```
Feature coefficients (sorted by absolute value):
            feature  coefficient
       PTS_per_game        1.485
first_half_attempts        0.650
  avg_shot_distance       -0.215
            eFG_pct       -0.179
  first_half_points        0.026
opponent_difficulty        0.000
```

Figure 1: Feature coefficients using Linear Regression predictive model

Preliminary Results (cont.)

Model Evaluation

- Mean Squared Error (MSE): 11.389
  - On average, the **squared difference** between your model's predicted second-half points and the actual second-half points is **11.389**
- Mean Absolute Error (MAE): **2.617**
  - On average, our predictions were off by about 2.78 points
- R-squared = **0.257**
  - The model explains around 25% of the variance in the second-half scoring. This is fairly low, indicating there is a lot of variability not captured by this feature alone.

Preliminary Results (cont.)

Evaluation Metrics

- Predicted Second-Half Points for LeBron James = 9.06
  - This means that, given the linear relationship the model learned from your dataset, LeBron James is predicted to average roughly 9 points in the second half (based on his aggregated first-half features).

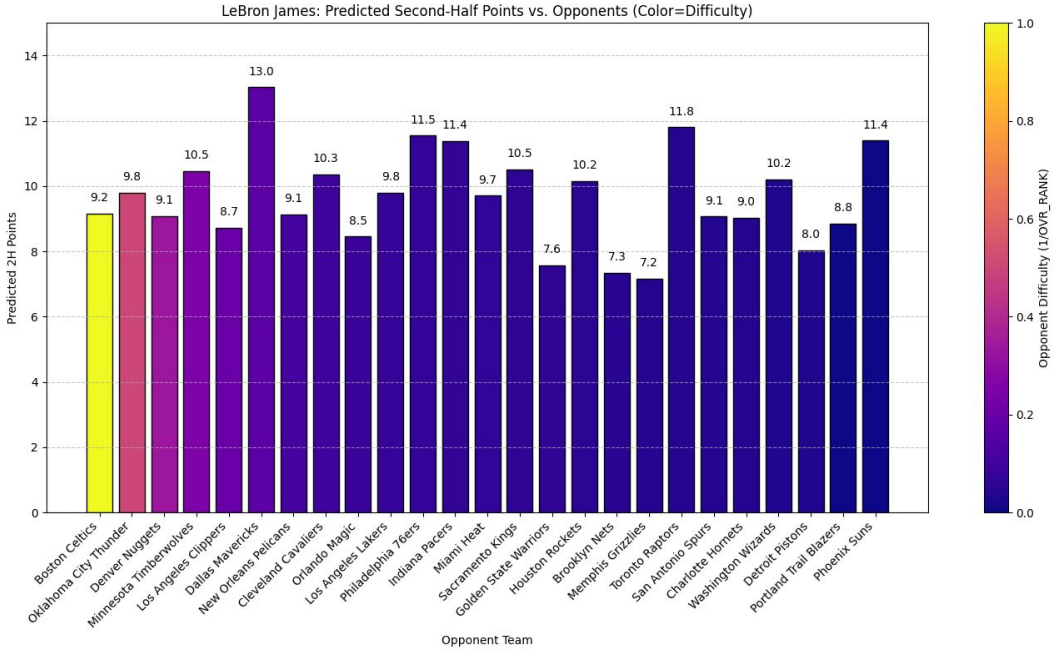

Figure 3: Bar chart that shows the predicted points of Lebron james based on team opponent

Preliminary Results (cont.)

Test Accuracy

- This chart visualizes **LeBron James' predicted vs. actual second-half points** across a series of games, with each data point representing a matchup against a specific opponent
  - Blue= Predicted
  - Orange= Actual
- There is a noticeable gap and fluctuation between the predicted and actual values in many games, which highlights the inherent volatility in sports performance—especially in the second half of an NBA game

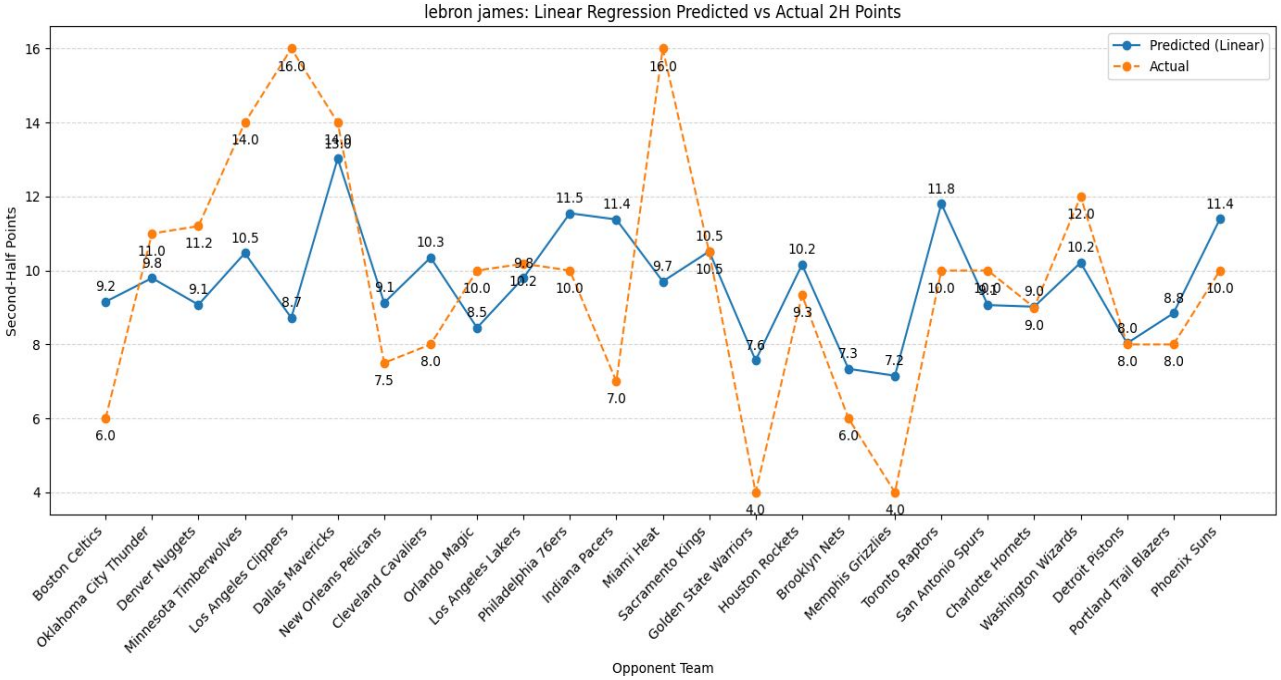

Figure 2: Chart that compares the predicted second half points of Lebron James Vs. actual points scored against every team

Compare linear to XGBoost

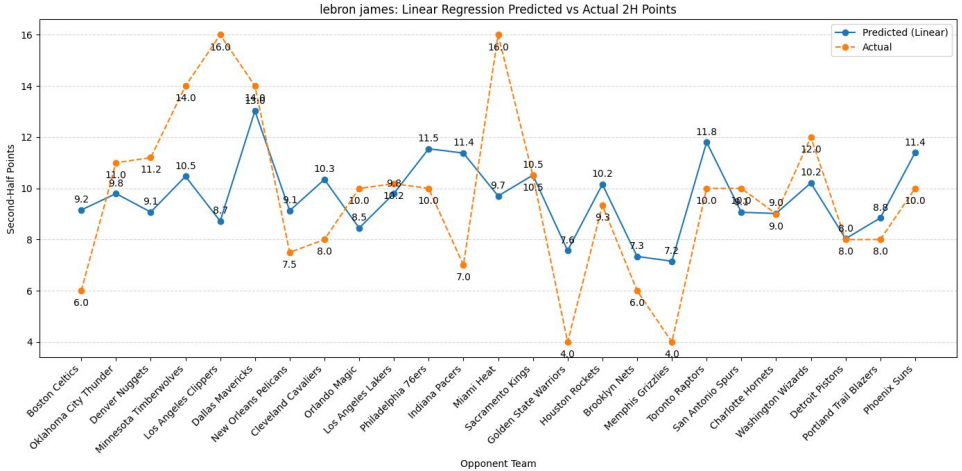

Figure 2: Chart that compares the predicted second half points of Lebron James Vs. actual points scored against every team using Linear Regression

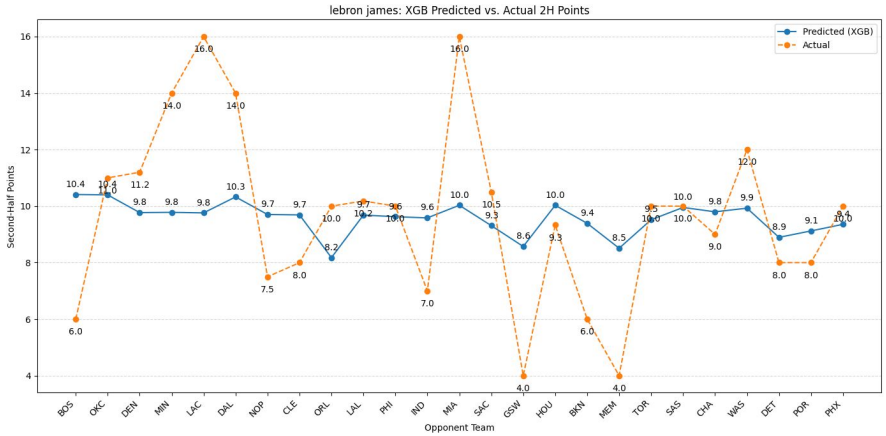

Figure 4: Chart that compares the predicted second half points of Lebron James Vs. actual points scored against every team using XGBoost

Linear Regression:

- Shows **more variance** in predicted values.
- Misses some of the **extremes**, especially underpredicting high actual scores (e.g., 16-point games).
- More sensitive to outliers and less flexible due to its **linear assumptions**.

XGBoost:

- Produces **more consistent** predictions.
- Closer tracking to actual points, especially on games with high or low values.
- **Better generalization** across teams, even when there are fluctuations in LeBron's real 2H scoring.
- Less volatile, thanks to XGBoost's ability to capture **non-linear relationships**.

Limitations

Limited Data Size

- Studies [2], [11], and [12] used small or single-season datasets, reducing model reliability.
- These limitations hinder the model's ability to generalize across different seasons or team compositions.

No Real-Time or Dynamic Updating

- Models in [1], [2], [4], [11], [12], [13], and [15] are static and not updated throughout the season.
- Unable to adapt to real-world factors such as injuries, trades, fatigue, or momentum.

Lack of Player-Level or In-Game Context

- Models in [1], [2], [4], [11], [12], and [13] rely heavily on aggregated team-level stats.
- This approach overlooks game rotations, specific roles, and situational context.

Future Modeling

- **Incorporate Additional Data Sources**
  - [1],[2],[11]Free throw stats and overtime performance should be added to reflect total scoring more accurately.
  - Include quarter-by-quarter splits, not just halves, for finer granularity
- **Include Contextual and Situational Features**
  - [1],[2],[4],[11]Add foul trouble, turnovers, and team score margin at halftime—these all affect second-half play.
  - Factor in rest days, travel distance, and back-to-back games to capture player fatigue.
- **Personalize the Model**
  - Instead of general models for all players, build individual models tailored to specific players like LeBron, since their playstyle and usage differ from role players

Conclusion

Linear regression offered a transparent way to understand how individual features influenced player's second-half performance. Its linear nature made it easy to trace how changes in each input directly affected the prediction.

- **PTS_per_game (+1.79)**
  This was the **strongest predictor**, meaning players who score more on average throughout the season tend to maintain or exceed that output in the second half. It reflects overall scoring ability and offensive role.

- **first_half_attempts (+0.68)**
  While not dominant feature, it gives indication that a high number of first-half shot attempts was positively correlated with second-half scoring

- **MP_per_game (−0.50)**
  Interestingly, more average minutes played per game was slightly **negatively correlated** with second-half scoring. This may suggest possible **fatigue**, **pacing**, or **rest patterns**, especially for veteran players like LeBron.

Reference List

[1] A. Patrot, H. H, S. B, G. P L and Sahana, "NBA Game Prediction Using Machine Learning Algorithm," *2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC)*, Mysore, India, 2023, pp. 1-6, doi: 10.1109/ICRTEC56977.2023.10111906.

[2] S. Geng and T. Hu, "Sports Games Modeling and Prediction using Genetic Programming," *2020 IEEE Congress on Evolutionary Computation (CEC)*, Glasgow, UK, 2020, pp. 1-6, doi: 10.1109/CEC48606.2020.9185917.

[3] A. Šarčević, M. Vranić, D. Pintar and A. Krajna, "Predictive modeling of tennis matches: a review," *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, Opatija, Croatia, 2022, pp. 1099-1104, doi: 10.23919/MIPRO55190.2022.9803645.

[4] K. Trawiński, "A fuzzy classification system for prediction of the results of the basketball games," *International Conference on Fuzzy Systems*, Barcelona, Spain, 2010, pp. 1-7, doi: 10.1109/FUZZY.2010.5584399.

[5] Y. Wu, "NBA game outcomes prediction using several machine learning algorithms and Voting Classifier," *2024 7th International Conference on Data Science and Information Technology (DSIT)*, Nanjing, China, 2024, pp. 1-9, doi: 10.1109/DSIT61374.2024.10881130.

Reference List

[6] A. P. C A, Suhas, S. B J, S. M. Anvekar and U. R. C G, "Optimal Cricket Team Selection Using Machine Learning," *2024 Second International Conference on Advances in Information Technology (ICAIT)*, Chikkamagaluru, Karnataka, India, 2024, pp. 1-5, doi: 10.1109/ICAIT61638.2024.10690755.

[7] I. M. Devi and S. Juliet, "Game Statistics Forecast Based on Sports Using Machine Learning," *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, Kollam, India, 2023, pp. 645-650, doi: 10.1109/ICCPCT58313.2023.10245637.

[8] J. R. Landers and B. Duperrouzel, "Machine Learning Approaches to Competing in Fantasy Leagues for the NFL," in *IEEE Transactions on Games*, vol. 11, no. 2, pp. 159-172, June 2019, doi: 10.1109/TG.2018.2841057.

[9] J. Brito, J. Ferro, D. Costa, E. Costa, R. Lopes and J. Fechine, "A ranking between attributes selection models using data from NCAA Basketball players to determine their tendency to reach the NBA," *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, Aveiro, Portugal, 2023, pp. 1-6, doi: 10.23919/CISTI58278.2023.10211486.

[10] Y. Zhai and T. Xu, "Novel Metric to Predict NBA Regular Season MVP," *2024 10th IEEE International Conference on High Performance and Smart Computing (HPSC)*, NYC, NY, USA, 2024, pp. 36-42, doi: 10.1109/HPSC62738.2024.00014.

Reference List

[11] J. Han, "Using machine learning models to predict win rate of NBA teams," *2023 9th International Conference on Systems and Informatics (ICSAI)*, Changsha, China, 2023, pp. 1-4, doi: 10.1109/ICSAI61474.2023.10423336.

[12] D. Sikka and R. D, "Basketball Win Percentage Prediction using Ensemble-based Machine Learning," *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, Coimbatore, India, 2022, pp. 885-890, doi: 10.1109/ICECA55336.2022.10009313.

[13] V. Veluru, T. Xiao, S. Addagudi, S. Kumar and G. Mohanraj, "Machine Learning Optimization Model to Predict Fantasy Basketball Teams," *2024 International Conference on Computing and Data Science (ICCDS)*, Chennai, India, 2024, pp. 1-4, doi: 10.1109/ICCDS60734.2024.10560397.

[14] "Deep Similarity Learning for Sports Team Ranking," *2021 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, Potchefstroom, South Africa, 2021, pp. 1-6, doi: 10.1109/SAUPEC/RobMech/PRASA52254.2021.9377253.

[15] M. Ahmadalinezhad, M. Makrehchi and N. Seward, "Basketball Lineup Performance Prediction Using Network Analysis," *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Vancouver, BC, Canada, 2019, pp. 519-524, doi: 10.1145/3341161.3342932.