NBA Player Performance ML Project

Ameesh Daryani

May 2024

1 Abstract

This project utilizes publicly available datasets from Kaggle, spanning three NBA seasons (2021-2024), to explore player efficiency, team dynamics, and performance trends. The objective was to obtain significant insights to support player appraisal and team strategy optimization by utilizing cutting-edge machine learning techniques, namely neural networks.

This project focuses on developing a predictive machine learning model to estimate players' ranks based on their performance metrics. By using a variety of player statistics including field goals, assists, rebounds, and more, the model attempts to use a players contribution to their teams in context with the performance of everyone else in the league to derive a ranking. Along with examining player stats in machine learning models, the research also looks into percentage statistics (such as field goal and three-point percentages) and how these vary over time for individual players. In order to comprehend offensive tactics and team dynamics, team-level performance data are also examined, such as average points per game.

Throughout the project, Python programming language and libraries such as pandas, numpy, matplotlib, and TensorFlow are utilized for data manipulation, analysis, and visualization. Machine learning techniques, such as neural networks, are employed for predictive modeling and analysis of player performance metrics. In conclusion, this project provides valuable insights into NBA player performance, efficiency, and team dynamics across multiple seasons. By leveraging data analysis and visualization techniques, teams, coaches, and analysts can make informed decisions to optimize player performance and team strategies in the ever-evolving landscape of professional basketball.

1.1 Index Terms

- **Machine Learning**: is a subset of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computers to learn from and make predictions or decisions based on data without being explicitly programmed.
- **Neural Networks**: are a type of machine learning algorithm inspired by the structure and function of the human brain. They consist of interconnected layers of artificial neurons, which process input data and make predictions or classifications based on learned patterns.
- **Dropout Layer**: is a regularization technique used in neural networks to prevent overfitting. In a dropout layer, a random subset of neurons is temporarily ignored or "dropped out" during training, reducing the network's reliance on any single neuron and improving generalization.
- **Dense Layer**: also known as a fully connected layer, is a type of layer in a neural network where each neuron is connected to every neuron in the preceding layer. This means that the output of every neuron in the layer is dependent on the input from every neuron in the previous layer.
- **Fully Connected Neural Network**: is a type of neural network architecture where each neuron in one layer is connected to every neuron in the subsequent layer, forming a fully connected graph between layers.
- **Data Analysis**: is the process of inspecting, cleaning, transforming, and modeling data to uncover insights, patterns, and trends that can inform decision-making and problem-solving.
- **Predictive Modeling**: is a process in which a statistical or machine learning model is developed to make predictions about future outcomes based on historical data. These models are trained on past data to learn patterns and relationships, which are then used to predict future observations.
- **Player Efficiency**: is a measure of a player's overall contribution to their team's performance on the court. It often encompasses various statistics such as scoring, rebounding, assists, steals, blocks, and turnovers, aggregated into a single metric or rating.
- **Performance Metrics**: are quantitative measures used to assess the effectiveness or efficiency of a system, process, or individual. In the context of sports analytics, performance metrics can include statistics such as points scored, rebounds, assists, shooting percentage, turnovers, etc.
- **Data Visualization**: is the graphical representation of data and information using visual elements such as charts, graphs, and maps. It is used

to communicate insights and patterns in data more effectively and intuitively, making it easier for users to understand and interpret complex information.

2 Overview and Motivation

Data Analytics is a rapidly growing field that is changing at an increasing rate. The advent of Big Data has brought abtout a boom in the need to be able to make sense of and draw meaningful conclusions from vast quantities of data. The rise of Machine Learning and iterative learning models as a way to derive meaningful inferences from data and make complex predictions based on those iteratively generated inferences has transformed Data Analytics into a rich and explosively expanding field. Utilizing Data-driven insights is a must for all companies, used to drive business decisions, increase product quality, and develop marketing audiences. In this project, machine learning is used to provide data-driven insights in the context of the highly competitive NBA. NBA teams compete at the highest level, and as such, must utilize every avenue possible to inform their decisions. The use of machine learning to provide a framework for the prediction of the relative ranking of every player in the league would be beneficial for several reasons. Chief among them is that it allows teams to assess player performance beyond traditional statistics. Analyzing patterns across multiple seasons and combining various performance metrics allows teams to identify undervalued players and potential stars before they become widely recognized. This can lead to better scouting, more strategic signings, and effective game strategy adaptations. Additionally, the model could help in injury management by predicting potential declines in performance due to physical stress, preventing the risk of long-term damage to players. In essence, the integration of machine learning in NBA analytics represents a transforming shift towards a more data-driven approach in sports management. It empowers teams to make smarter decisions, optimize player performance, and enhance fan engagement. Ultimately, this project demonstrates the application of advanced analytics in sports and sets the stage for future innovations in the field of sports data science.

3 Literature Review

3.1 Relevant Work

Using machine learning to create meaningful insights about the NBA and its players is not a new idea, and has been explored thoroughly in the literature. The following are three studies/models that give

different insights about the NBA using Machine Learning with large quantities of player data.

"Building My First Machine Learning Model — NBA Prediction Algorithm" by Fayad, A. introduces a beginner's approach to predicting NBA game outcomes using machine learning. Unlike more complex models, the author's methodology relies on basic statistical analysis and regression models, leveraging monthly NBA statistics as predictors. Achieving a commendable 72% accuracy in predicting game outcomes, the study distinguishes itself by prioritizing game predictions over player rankings and employing a relatively straightforward model due to the author's limited experience.

On the other hand, "Predictive Analysis of NBA Game Outcomes through Machine Learning" by Wang, J. delves into more sophisticated machine learning techniques to forecast NBA game results. Employing algorithms such as Logistic Regression, Support Vector Machines, Deep Neural Networks (DNN), and Random Forest models, this study rigorously evaluates model performance and emphasizes the significance of field goal percentage in predictions. Unlike the previous study, it focuses exclusively on game outcomes and provides valuable insights into the factors influencing these results.

Meanwhile, "Machine Learning Uncovers Nine Distinct Player Types in the NBA" by Macedo, A. B. shifts the focus to understanding player roles within the NBA through non-supervised machine learning techniques. By analyzing Per Game statistics for the current NBA season and utilizing clustering methods, the study identifies nine distinct player types based on their performance metrics. Unlike the previous studies, this research does not aim to predict game outcomes but instead provides valuable insights into the diversity of player roles and their impact on team dynamics.

In summary, these three studies are good examples of the growing body of literature on the application of machine learning in NBA analytics. While the first two focus on predicting game results using different levels of complexity in their methodologies, the final study discussed offers a unique perspective by examining player roles within the league. All together, these studies highlight the power of machine learning in analyzing the aspects of sports and their ability to offer valuable insights for coaches, analysts, and enthusiasts alike.

3.2 Contributions of The Project

This project amalgamates and builds upon several established methodologies to advance the predictive analytics of NBA player performance. In the study referenced above by Fayad, the use of basic statistical analysis and regression models for predicting NBA game outcomes shows foundational machine learning techniques being applied effectively in sports analytics. This project extends that initial approach, integrating more complex neural network architectures that handle nuanced relationships within player performance data, shifting the focus from game outcomes to individual player metrics. Further, the sophisticated machine learning techniques employed in study shown above by Wang, such as Deep Neural Networks (DNNs) and Logistic Regression, gave the backbone that this project adapted for predicting detailed player performance statistics. Unlike Wang's focus on game outcomes, however, this project uses machine learning to directly predict continuous performance metrics, which offers a tool for evaluating and forecasting player contributions in relation to every other player in the league.

Overall, this project focuses less on the classification or outcome based models shown in the literature review, instead aiming to give a direct quantitative analysis of player performance and their expected placement against each other.

4 Technology Details

The goal of this project is to utilize a Fully Connected Neural Network (FCNN) to predict the rank (Rk) of NBA players from their performance metrics across three seasons. The project leverages player statistics from the 2021-2024 NBA seasons, utilizing data-driven techniques to establish a predictive model.

Data Collection and Preprocessing

Data from three consecutive NBA seasons (2021-2022, 2022-2023, and 2023-2024) was compiled from CSV files, each representing a season's complete player statistics. The primary steps involved in data preprocessing included:

1. **Data Loading**: Each season's data is read into separate Pandas dataframes.

```
data_2122 = pd.read_csv('/content/2021-2022 NBA Player Stats - Regular.csv',
                        sep=";", encoding='latin-1')
```

2. **Data Merging**: Concatenation of the individual season dataframes into a single comprehensive dataset.

```
combined_data = pd.concat([data_2122, data_2223, data_2324], ignore_index=True)
```

3. **Normalization**: Selected statistical features (e.g., Field Goals, Minutes Played) are normalized using z-score normalization to facilitate effective model training by scaling all features to a similar range.

```
features = ['FG', 'FGA', '3P', '3PA', 'FT', 'FTA', 'MP', 'eFG%']
combined_data[features] = combined_data[features].apply(
    lambda x: (x - x.mean()) / x.std())
```

Neural Network Model

Structure and Type

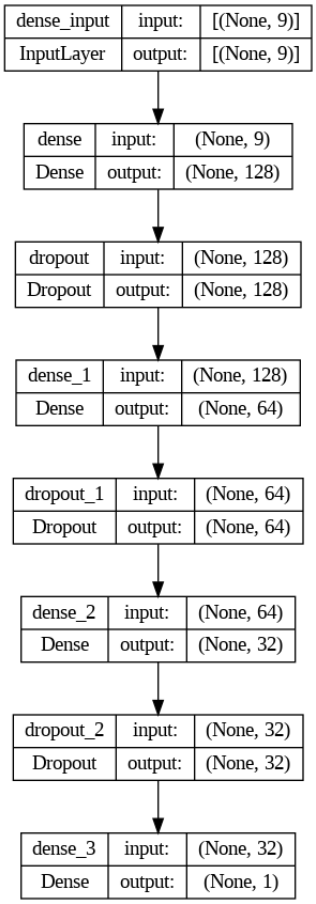
This FCNN, designed for regression, predicts the normalized ranks of NBA players based on their performance metrics. The structure includes:

1. **Input Layer**: Processes input features.
2. **Hidden Layers**: Includes multiple layers with regularization and dropout to prevent overfitting, featuring ReLU activation for non-linearity.
3. **Output Layer**: A single neuron outputs the predicted rank.

Configuration

- **Layer 1**: 128 neurons, ReLU activation, L2 regularization
- **Dropout**: 50% dropout rate to mitigate overfitting
- **Layer 2**: 64 neurons, similar configurations as Layer 1
- **Dropout**: 50% dropout rate to mitigate overfitting
- **Layer 3**: 32 neurons, continuing the pattern of regularization and dropout
- **Dropout**: 50% dropout rate to mitigate overfitting
- **Output Layer**

```
model = models.Sequential([
    layers.Dense(128, activation='relu', kernel_regularizer=regularizers.l2(0.001),
                 input_shape=(X_train.shape[1],)),
    layers.Dropout(0.5),
    layers.Dense(64, activation='relu', kernel_regularizer=regularizers.l2(0.001)),
    layers.Dropout(0.5),
    layers.Dense(32, activation='relu', kernel_regularizer=regularizers.l2(0.001)),
    layers.Dropout(0.5),
    layers.Dense(1)
])
```



Optimization and Loss Function

- **Optimizer**: Adam optimizer is utilized due to its effectiveness in handling sparse gradients and adapting learning rates.
- **Loss Function**: Mean Squared Error (MSE) is employed to quantify the model's prediction errors, ideal for regression problems.

```
model.compile(optimizer='adam', loss='mse',
              metrics=['mean_squared_error', 'mean_absolute_error'])
```

Training

The model undergoes training for 100 epochs with a batch size of 10. A 70-30 train-test split ensures sufficient data for learning while allowing validation of model performance against unseen data.

```
history = model.fit(X_train, y_train, epochs=100, batch_size=10,
                    validation_data=(X_test, y_test))
```

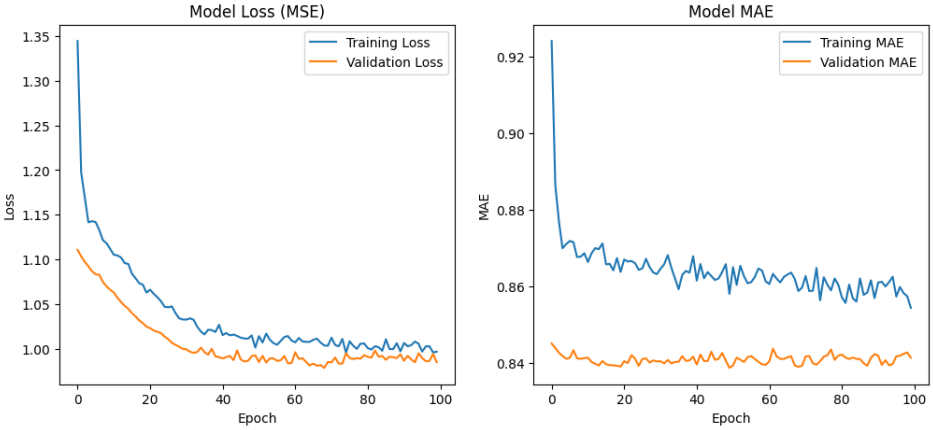
5 Performance Evaluations

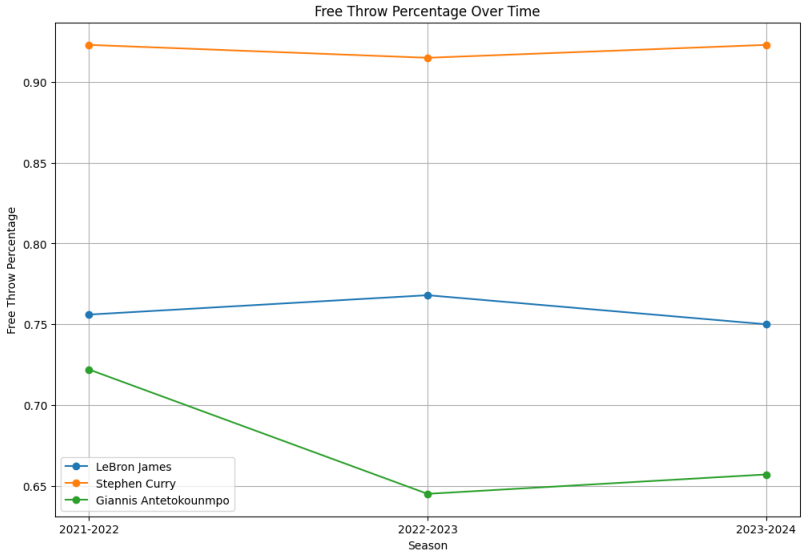
The model's performance was analyzed based on loss, mean absolute error (MAE), and mean squared error (MSE) metrics, alongside their validation counterparts.

1. **Loss and Validation Loss:** Both loss and validation loss decreased consistently over epochs, indicating the model's learning progress. They eventually stabilized, suggesting convergence.

2. **Mean Absolute Error (MAE) and Validation MAE:** MAE and validation MAE decreased steadily, indicating improved prediction accuracy. However, they reached a plateau in later epochs, indicating stabilization.

3. **Mean Squared Error (MSE) and Validation MSE:** MSE and validation MSE followed similar patterns, decreasing over epochs and stabilizing towards the end.


Model Loss (MSE)
Training Loss
Validation Loss
Loss
Epoch
Model MAE
Training MAE
Validation MAE
MAE
Epoch


Free Throw Percentage Over Time
Free Throw Percentage
LeBron James
Stephen Curry
Giannis Antetokounmpo
2021-2022
2022-2023
2023-2024
Season

6 Conclusions

In conclusion, this project aimed to develop a predictive model using Fully Connected Neural Networks (FCNN) to rank NBA players based on performance data from three seasons (2021-2024). The effort focused on meticulous data preparation and the implementation of advanced machine learning techniques. The neural network structure incorporated dropout layers to avoid overfitting and utilized ReLU activation to effectively handle non-linear data.

Throughout the project, the model demonstrated consistent improvement in its ability to predict player rankings, as evidenced by decreasing loss and error rates over 100 training epochs. This indicates that the chosen approach for setting up and training the neural network was effective.

However, several challenges were encountered. A primary issue was the inability to immediately test the model's predictions due to the time lag in obtaining new NBA season data. Additionally, difficulties were faced in transforming the model into a more accessible format, such as a web app or API, which would enhance its practical application.

Plans to address these challenges include the creation of a simulation environment that can use past data to predict and verify the model's accuracy until new data becomes available. Furthermore, collaboration with experts in software development is considered essential to facilitate the integration of the model into user-friendly platforms. This would make the model more accessible to basketball coaches and sports analysts, thereby broadening its applicability.

Overall, the project has laid a solid foundation for future exploration and enhancements in sports analytics, showcasing the potential of machine learning technologies to revolutionize the analysis and optimization of player performance and team strategies in professional sports.

7 References

C. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, 2011.

Vivovinco. (n.d.). NBA Player Stats [Data set]. https://www.kaggle.com/datasets/vivovinco/nba-player-stats/data

Vivovinco. (n.d.). 2022-2023 NBA Player Stats - Regular Season [Data set]. https://www.kaggle.com/datasets/vivovinco/20222023-nba-player-stats-regular?select=2022-2023+NBA+Player+Stats+-+Regular.csv

Vivovinco. (n.d.). 2023-2024 NBA Player Stats [Data set]. https://www.kaggle.com/datasets/vivovinco/2023-2024-nba-player-stats/data

Fayad, A. (2020, July 9). Building My First Machine Learning Model — NBA Prediction Algorithm. *Towards Data Science*.

Wang, J. (2023). Predictive Analysis of NBA Game Outcomes through Machine Learning. In *The 6th International Conference on Machine Learning and Machine Intelligence (MLMI 2023)*.

Macedo, A. B. (2023, March 21). Machine Learning Uncovers Nine Distinct Player Types in the NBA. *Samford University Center for Sports Analytics*.

Brownlee, J. (2019). A Gentle Introduction to Dropout for Regularizing Deep Neural Networks. *Machine Learning Mastery*. Available at: Link

(2020). The Four Most Important Deep Learning Layers. *Towards Data Science*.

(2024). Dense Neural Networks: Understanding Their Structure and Function. *DataScientest*.