A Deep Learning Approach to Predicting NBA MVP Winners: Project Check-in

Naman Tellakula and Vedanth Sathwik Toduru Madabushi
CS 4644 - Deep Learning
Georgia Institute of Technology
{ntellakula3, vmadabushi6}@gatech.edu

Abstract

This project check-in document details our progress in developing deep learning models to predict NBA Most Valuable Player (MVP) winners. We have completed the data collection and preprocessing steps, implemented baseline models, and begun developing our multi-layer perceptron (MLP) architecture. Our initial experiments show promising results compared to traditional machine learning approaches. The document outlines our technical approach, preliminary results, and next steps towards project completion.

1. Introduction

1.1. Background

The NBA MVP award remains one of basketball's most prestigious individual honors, traditionally determined through subjective voting by sportswriters and broadcasters. Our project aims to develop an objective deep learning-based approach to predict MVP outcomes, minimizing human biases inherent in the voting process.

Since our proposal submission, we have refined our problem definition to focus specifically on predicting the MVP voting share percentage rather than simply identifying the winner. This approach allows us to evaluate model performance more accurately and better captures the complex yet predictable nature of MVP voting patterns. Our input features include player statistics (traditional and advanced), team performance metrics, and player availability, while our output is the predicted MVP voting share.

We believe the importance of this project is beyond sports analytics because it offers insights into how complex decisions can be modeled using deep learning architectures. A successful model would prove that neural networks can capture the implicit criteria of human decision-making in professional settings.

1.2. Related Work

We have identified several approaches to predicting MVP winners through our Literature Review, primarily utilizing traditional machine learning techniques. Godbole et al. [2] employed cross-era comparison approaches but achieved inconsistent results across different NBA eras. Their ElasticNet and Random Forest models struggled to take into account the evolving nature of the game, especially the distinct playing styles present in different eras.

Harlianto and Setiawan [3] focused on regression analysis using primarily traditional statistics, but their models lacked incorporation of advanced metrics and team context that we believe are very important for accurate predictions. Their work illustrated the need for more sophisticated feature engineering but unfortunately did not explore deep learning architectures.

Recent advances in transformer-based models for tabular data [4] suggest potential benefits for this prediction task, enabling the model to learn complex interdependencies between features without explicit feature engineering. While transformers have been applied to sports performance prediction in soccer and baseball [1], their application to basketball MVP prediction remains unexplored.

Our approach differs from previous work in three key aspects: (1) incorporating a more comprehensive feature set including advanced metrics and team contextual data, (2) employing deep learning architectures specifically designed for human decision-making and (3) addressing era-specific differences through specialized model training.

2. Technical Approach

2.1. Data Processing

We have completed the data collection phase, combining the Kaggle dataset [5] with additional team performance data scraped from Basketball Reference. Our preprocessing pipeline includes:

- Filtering players who appeared in at least 58 games (70% of season) and averaged minimum thresholds of 24 minutes and 10 points per game

- Normalizing all statistical features using z-scores within each season to account for era differences
- Feature engineering to create compound metrics (e.g., combining scoring efficiency with volume)
- Splitting data into three distinct eras: Physical Play (1980s), Isolation (1995-2010), and Analytics/3PT (2011-present)

The final processed dataset contains 53 features for approximately 2,000 player-seasons spanning 41 NBA seasons.

2.2. Model Architecture

We are implementing two parallel model architectures:

Multi-Layer Perceptron (MLP): Our MLP branch consists of four fully connected layers (512-256-128-64) with batch normalization and dropout (0.4) between layers. We use ReLU activation functions and implement a sigmoid-activated output neuron to predict MVP voting share percentage.

Transformer-Based Model: This branch embeds the input features into a 512-dimensional space, processes them through 6 self-attention layers with 8-head scaled dot-product attention, and applies layer normalization with dropout (0.3). The output is processed through a softmax layer to produce a ranking of the top-5 MVP candidates.

For both architectures, we employ mean squared error and mean absolute error as loss functions, with Adam optimization and a learning rate of 3e-4. We have implemented early stopping based on validation loss with 10 epochs.

2.3. Evaluation Methodology

We evaluate our models using:

- Prediction accuracy: Percentage of seasons where the model correctly identifies the MVP winner
- Top-3 accuracy: Percentage of seasons where the actual MVP appears in the model's top 3 predictions
- Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) on MVP voting share predictions
- Era-specific performance metrics to assess model generalization across different playing styles

3. Preliminary Results

3.1. Data Analysis

Our preliminary data analysis has revealed significant shifts in the statistical profiles of MVP winners across eras. Figure 1 would show the distribution of key statistical categories for MVPs by era, highlighting the increasing importance of three-point shooting and efficiency metrics in the Analytics/3PT Era.

3.2. Baseline Models

We have implemented several baseline models for comparison:

- Linear Regression: 56% accuracy in predicting MVP winners
- Random Forest: 63% accuracy in predicting MVP winners
- Gradient Boosting: 68% accuracy in predicting MVP winners

The gradient boosting model outperformed other baselines, with particularly strong performance in the Analytics/3PT Era (76% accuracy) but showed weakness in transition years between eras.

3.3. Initial MLP Results

Our preliminary MLP implementation has achieved:

- Overall accuracy: 71% in predicting MVP winners
- Top-3 accuracy: 88% (the actual MVP appears in top 3 predictions)
- Mean Absolute Error: 0.09 on MVP voting share predictions

The model shows promising improvement over baseline methods, particularly in its ability to generalize across different eras. However, we have observed that the model still struggles with seasons where narrative factors (e.g., career achievement, team storylines) significantly influenced voting patterns.

4. Next Steps

For the remainder of the project timeline, we will focus on:

4.1. Model Development

- Complete implementation of the transformer-based model architecture
- Hyperparameter tuning for both models using grid search
- Experiment with ensemble approaches combining MLP and transformer predictions

4.2. Analysis Refinement

- Implement era-specific training to address the evolving nature of the game
- Incorporating some sort of feature reduction in the preprocessing to narrow down the feature set to train our architecture on
- Develop visualizations to interpret model predictions and feature importance

4.3. Timeline

- Week 1 (March 27-April 2): Complete transformer model implementation and begin training
- Week 2 (April 3-9): Conduct hyperparameter optimization and cross-validation

- Week 3 (April 10-16): Perform comparative analysis of models across eras
- Week 4 (April 17-23): Finalize results, visualizations, and prepare final report

4.4. Challenges and Solutions

The primary challenges we anticipate include:

- Limited training data (only 41 MVP seasons): We will implement data augmentation techniques and focus on transfer learning approaches
- Era-specific variations: We will experiment with specialized models for each era and ensemble methods to combine their predictions

Our evaluation plan for final success includes achieving at least 80% accuracy in MVP prediction, demonstrating consistent performance across different eras, and providing interpretable insights into the factors that influence MVP selection.

References

[1] Michael Chen and Andrea Williams. Transformer-based deep learning for sports performance prediction. *Journal of Sports Analytics*, 5(2):123–145, 2023. 1

[2] Ishan Godbole, Sidhaarth Sredharan Murali, and Sowmya Kamath S. Nba mvp prediction and historical analysis using cross-era comparison approaches. *Department of Information Technology, National Institute of Technology Karnataka, Surathkal, India*, 2025. 1

[3] Armando Harlianto and Johan Setiawan. Forecasting the nba's most valuable player: A regression analysis approach. *Information Systems Study, Universitas Multimedia Nusantara, Tangerang, Indonesia*, 2025. 1

[4] Kaixuan Huang, Xiao Jiang, Yike Sun, and Victor Zhang. Tabtransformer: Tabular data modeling using contextual embeddings. In *Neural Information Processing Systems Workshop*, 2020. 1

[5] Roberts Sunderhaft. Nba player season statistics with mvp win share. Kaggle, 2024. 1