Machine Learning II Final Project

Comparison and Prediction of NBA Players Success Based On Their Combine Performance and Stats

Amaey Pandit, Derrick Gan, Ben Rice

Machine Learning 4420

Dr. Gerber

April 8th, 2025

#### Introduction And Literature Review

###### Problem and Research Topic:

Use NBA Combine performance metrics to predict whether a player will have a successful NBA rookie season, based on historical data of past prospects.

###### Background:

Our project's main source of data, which we will go into greater detail later in this report, is NBA Combine Data. For context, and to help clarify how we will be using this data, let us first explain what the NBA Combine is and the type of data that is collected.

According to Google, the NBA Draft Combine is “a multi-day event held before the annual NBA draft where college and international basketball players undergo various evaluations, including athletic testing, medical exams, and interviews, to showcase their skills to NBA scouts and team representatives.” In addition to player performance, the combine measures various physical attributes such as hand size, foot size, wingspan, body fat percentage, and standing reach. These are just a few examples of the types of data we have included. We intend on using the machine learning methods, MLP to predict player performance and Collaborative Filtering to support player comparisons.

###### Past Works:

Prior research that we found on forecasting NBA draft success typically fell into three categories: regression-based analysis, similarity-based modeling, and machine learning approaches. Each offering unique insights into how pre-draft data can translate to professional performance. Traditional regression models, such as those described in Breaking the Lines’ [Statistical Regression Models for Draft Success](https://breakingthelines.com/opinion/forecasting-greatness-how-statistical-models-predict-the-rise-of-future-nba-stars/#:~:text=algorithms%2C%20patterns%20emerge,term%20success) use linear or logistic regression to link combine and college statistics to career outcomes like win shares or All-Star appearances. These methods offer interpretability by quantifying the importance of each variable but may oversimplify complex relationships. In addition to this, a paper from the MIT Sloan Sports Analytics Conference [How To Predict The Performance Of NBA Draft Prospects](https://mitsloan.mit.edu/shared/ods/documents?PublicationDocumentID=10079#:~:text=The%20authors%20describe%20a%20new,expected%20reliability%20of%20a%20specific) introduces a similarity-based modeling approach that mirrors collaborative filtering, matching prospects to historical players with similar attributes to estimate future performance. This method aligns closely with our own work and adds transparency by explaining predictions through player comparables.

Machine learning approaches offer a third avenue. As explored in the article that talks about using [Machine Learning To Predict Long Term NBA Value](https://www.tothemean.com/2014/06/17/machine-learning-predict-long-term-value-in-draft.html), models like random forests and support vector machines trained on college stats and combine data can often outperform actual draft decisions, especially when incorporating draft position as a feature. These models capture nonlinear patterns and interactions that regression might miss. Supporting all of these techniques is a growing body of work on the [Predictive Power of NBA Combine](https://wilson-wang.medium.com/the-predictive-power-of-the-nba-draft-combine-a-statistical-analysis-b45d15931fe5). Another point touched upon in Wilson Wang’s article shows how agility, sprint speed, and vertical jump tests hold moderate predictive power, especially among top prospects, reinforcing the relevance of physical metrics in draft modeling.

These research articles are adjacent studies to our combine predictions as most have a pretty specific approach, hypothesis, and objective. Some use combine as their base data for prediction, some use more or different but all have a theme of identifying talent and longevity/success in the league but one variable that is seen to be important is the college stats of the players.

Lastly, an important article we felt should be included is about several researchers that challenge the overemphasis on raw athleticism as a proxy for potential. A piece about [How do we assess potential?](https://www.canishoopus.com/2014/2/26/5435374/potential-nba-draft-prospects) critiques the vague use of “potential” in scouting, arguing that metrics like assist-to-turnover ratio or steal rate better predict long-term development than vertical leap or wingspan alone. We will dive deeper into this later in the report.

#### ML Methodology And Data Collection

##### Data Selection

To investigate the predictive potential of NBA Combine data, we compiled a dataset combining two major sources: the NBA Stats API for standard performance metrics for players rookie seasons, web-scraped data for strength and agility statistics, and another NBA API endpoint for anthropometric measurements. Our goal was to use these physical and athletic metrics to anticipate rookie performance without relying on in-season stats. To do this, we created a custom target variable called ROOKIE\_SCORE, designed to consolidate a player's rookie-year box score production into a single number. This metric was calculated using the formula:

ROOKIE\_SCORE = POINTS + 0.4 \* FIELD\_GOAL\_PCT \* POINTS + 0.7 \* REBOUNDS + 0.7 \* ASSISTS + STEALS + BLOCKS - TURNOVERS

The idea behind this formulation was to balance volume-based scoring with efficiency, rebounding and playmaking contributions, and defensive impact, while penalizing turnovers. This allowed us to benchmark rookies beyond points per game and more fairly assess overall output.

##### ML Method

Our primary model was a two-layer feedforward neural network (MLP) with a tanh activation in the first layer, a leaky ReLU in the second, and a linear output. The tanh activation offered smoother transitions and centered the data for early transformation, while the leaky ReLU in the second layer addressed the potential issue of dead neurons, ensuring gradients could still flow even for negative activations. The model took in normalized feature vectors and predicted a scalar rookie score.

The additional model we used was collaborative filtering to predict rookie performance for just the center position players using combine data comparative to actual player combine data - filtering for center like positions with at least 20 games and filling in missing values with average combine metrics. The prediction functions identify similar players using cosine similarity and/or Euclidean distance and estimate rookie scores by averaging the scores of the k nearest neighbors, incorporating error handling to provide robust predictions across various players. Our weights were created to rather than make an unbiased combined score was to emphasise attributes that recruits found most valuable when looking at center's features.

Height: 0.15

Weight: 0.1

Wingspan: 0.15

Standing Reach: 0.15

Body Fat Percent: - 0.1

Hand Length: 0.05

Hand Width: 0.05

Lane Adjility: -0.1

Three Quarter Sprint: -0.1

Standing Vertical Leap: 0.1

Max Vertical Leap: 0.1

Modified Lane Agility Time: -0.05

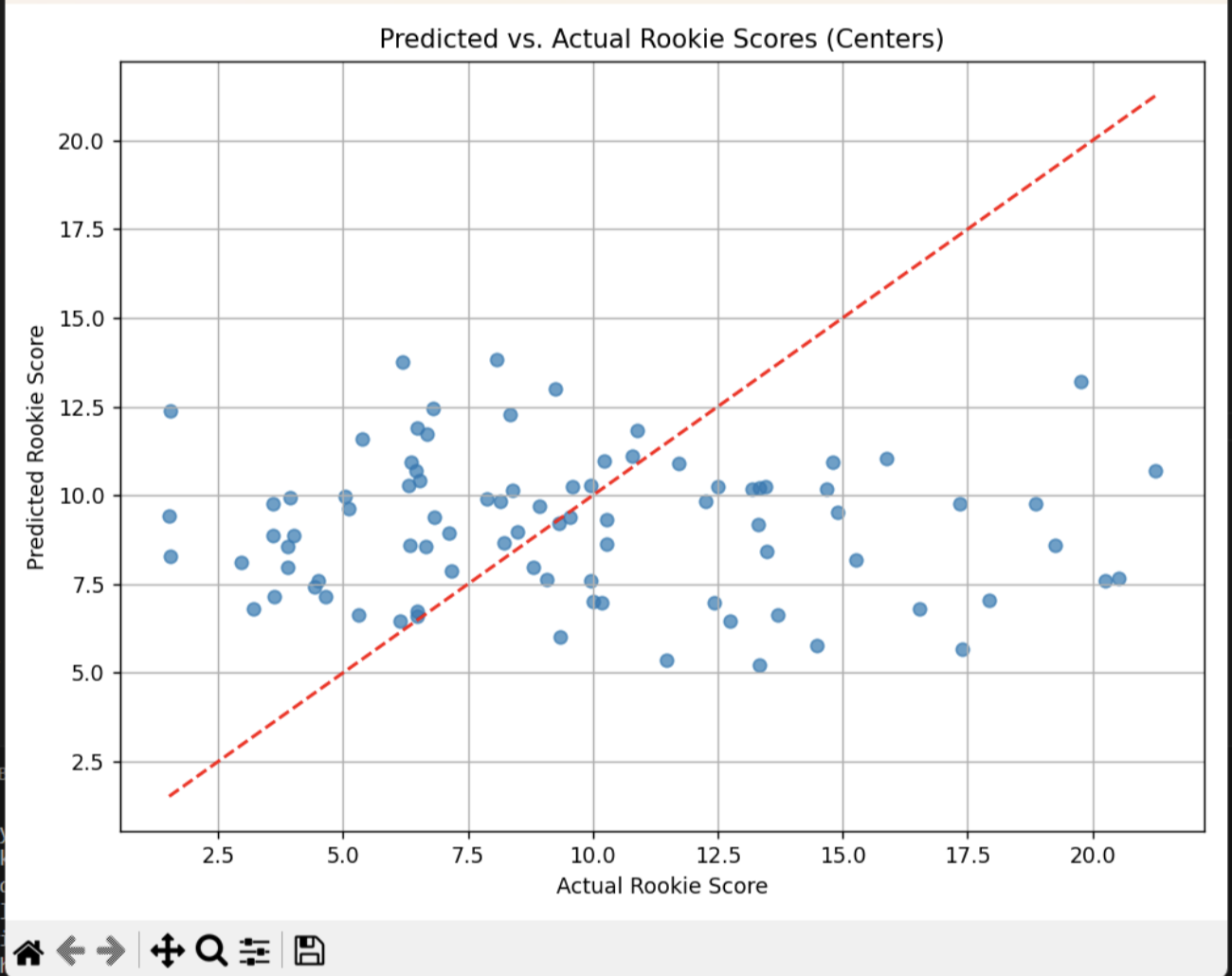
Training involved heavy tuning. Initial attempts with common learning rates were insufficient, leading to underfitting or unstable gradients. Additionally, we experimented with a variety of activation functions (tanh, ReLU, leaky ReLU, swish) and tested both one- and two-layer networks. Early models included a wide array of features from the combine dataset, but overfitting became a clear issue, especially as some features had near-zero correlation with the target. We resolved this by trimming our input to the most promising features (based on correlation with ROOKIE\_SCORE), eventually settling on a smaller set and adding GAMES\_PLAYED, which notably improved predictive accuracy. While its inclusion technically makes the model postdictive rather than predictive, it provides a helpful bridge to explore model potential in a limited-data setting.

#### ML Results And Interpretation

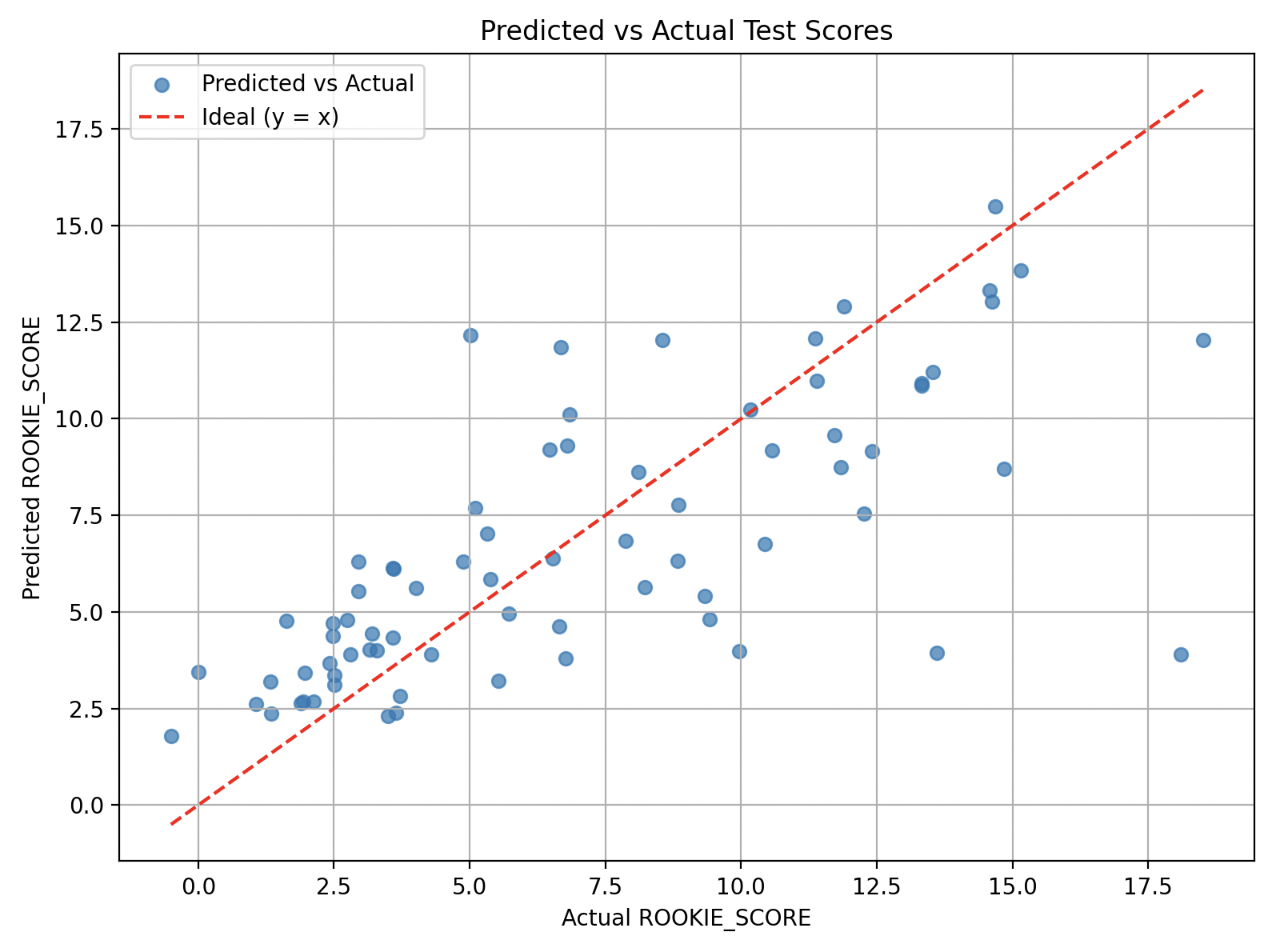
The best-performing MLP model achieved a test MSE of 10.7493, significantly outperforming the baseline MSE of 21.7912, where predictions were simply the mean ROOKIE\_SCORE across the training set. This model also achieved an R-squared score of 0.5066, meaning it explained just over half of the variance in rookie outcomes.

Our CF model wasn’t as promising, achieving a test MSE of 28.468 and R-squared score of -0.224, which indicates that our actual and predicted score differences are large as well as our data may be overfitting and a prediction of just using the mean would give us better results. The hypothesis, can you predict a rookies score based on the comparison of an existing player with similar stats is not possible. We assessed different tuning when trying to perform collaborative filtering but the oversimplicity of the model made it not possible to have a “good” prediction.

Our visualization to show our results relative to a perfect 1:1 line is below. This visual aligns with the negative R-squared score, indicating that the model’s predictions are performing worse than if we simply used the average rookie score for every player. Meanwhile, the relatively high MSE confirms that, on average, the squared errors are quite large which means the predicted values differ significantly from the true scores. Taken together, these metrics and the visualization suggest that the collaborative filtering approach in its current form is not effectively capturing or modeling the factors that drive a player’s rookie performance.



The plot below shows the predicted versus actual rookie scores on the test set. Although the points are scattered, there is a clear upward trend, and the predictions cluster reasonably close to the identity line. Some underprediction at the high end is visible, which is common in regression models trained on relatively small samples with skewed target distributions.



In terms of feature importance, we approximated input relevance by summing the absolute values of the first-layer weights (W1) for each feature:

MLP Input Feature Importances (abs sum of W1):

WEIGHT : 4.7823

HAND\_WIDTH : 6.3449

LANE\_AGILITY\_TIME : 2.6194

THREE\_QUARTER\_SPRINT : 6.9945

MAX\_VERTICAL\_LEAP : 3.3099

MODIFIED\_LANE\_AGILITY\_TIME : 4.2163

GAMES\_PLAYED : 9.1867

These values provide a rough sense of which inputs influenced the hidden layer most, although they shouldn't be interpreted as strict measures of importance. GAMES\_PLAYED, unsurprisingly, had the strongest contribution, as it's closely linked to opportunity and production. Sprint time and hand width followed, suggesting that both speed and size characteristics had some relation to rookie-year impact in this context.

The model did not show strong evidence of overfitting in its final form, and the trimmed feature set helped improve generalization. However, several predictions still hovered near the mean, particularly for players with lower ROOKIE\_SCORES, suggesting the model may be smoothing out too aggressively.

#### Conclusion And Future Work

This project demonstrates that even a relatively simple neural network can uncover modest patterns linking physical combine metrics to rookie-year output. Our MLP model achieved a meaningful improvement over a baseline and showed potential for use in player scouting or roster planning whereas our collaborative filtering model proved to be less impactful and worse at predicting player success.

However, the approaches had limitations. The inclusion of GAMES\_PLAYED improves performance, but undermines the claim that the model is truly predictive. In future iterations, we'd ideally model playing time separately or use it only for filtering the dataset (e.g., focusing on players who played at least 500 minutes). Additionally, the model’s ability to generalize to higher-performing rookies remains a challenge, possibly due to the limited number of high-end outcomes in the dataset.

Another important limitation is that not all prospects participate in the NBA Combine. In fact, some of the most anticipated draft picks opt out entirely, meaning that a purely combine-based model cannot be universally applied. This introduces a selection bias that needs to be accounted for.

To strengthen the model, future work could incorporate college statistics or international league performance metrics to help contextualize players’ combine numbers. Expanding the variety and resolution of combine data, including timing splits, shooting drill results, and other tests, could also improve model expressiveness. We would further explore unsupervised techniques to cluster rookies before prediction or to identify archetypes could also be useful. Ultimately, the goal is to develop a tool that can flag under-the-radar prospects based purely on pre-draft measurements—a challenging but worthwhile objective.

#### 

#### 

#### References

Davis, A. (2024, April 9). *Forecasting greatness: How statistical models predict the rise of future NBA stars*. Breaking the Lines. https://breakingthelines.com/opinion/forecasting-greatness-how-statistical-models-predict-the-rise-of-future-nba-stars/

[Berri, D. J., & Schmidt, M. B. (n.d.). *Using the NBA draft to evaluate the expected value of college basketball players*. MIT Sloan Sports Analytics Conference. https://mitsloan.mit.edu/shared/ods/documents?PublicationDocumentID=10079](https://breakingthelines.com/opinion/forecasting-greatness-how-statistical-models-predict-the-rise-of-future-nba-stars/)

[Wang, W. (2020, July 12). *The predictive power of the NBA draft combine: A statistical analysis.* Medium. https://wilson-wang.medium.com/the-predictive-power-of-the-nba-draft-combine-a-statistical-analysis-b45d15931fe5](https://breakingthelines.com/opinion/forecasting-greatness-how-statistical-models-predict-the-rise-of-future-nba-stars/)

Tothemean. (2014, June 17). *Can machine learning predict long-term value in the draft?* To the Mean. https://www.tothemean.com/2014/06/17/machine-learning-predict-long-term-value-in-draft.html

Eric\_in\_StPaul. (2014, February 26). *Potential NBA draft prospects*. Canis Hoopus. <https://www.canishoopus.com/2014/2/26/5435374/potential-nba-draft-prospects>

Cui, Y., Liu, F., Bao, D., Liu, H., Zhang, S., & Gómez, M.-Á. (2019). Key anthropometric and physical determinants for different playing positions during National Basketball Association draft combine test. \*Frontiers in Psychology, 10\*, 2359. https://doi.org/10.3389/fpsyg.2019.02359