Predicting NBA Draft Rankings with Machine Learning: Application of Pointwise Learning to Rank

Barnabás Szombat
John von Neumann Faculty of Informatics
Obuda University
Budapest, Hungary
barni6@stud.uni-obuda.hu

Natacha Moniz
Instituto Superior Tecnico
Universidad Lisboa
Lisboa, Portugal
natachamoniz@tecnico.ulisboa.pt

Gábor Kertész
John von Neumann Faculty of Informatics
Obuda University
Budapest, Hungary
kertesz.gabor@nik.uni-obuda.hu

***Abstract*—The NBA draft plays a pivotal role in shaping the future of professional basketball teams by selecting talented players from a pool of prospects. However, predicting draft rankings is a challenging task due to the complexity of evaluating player potential and the complexity of the decision-making process involving both objective performance metrics and subjective team preferences. Machine learning based ranking mthods have shown promise in addressing these challenges by systematically analyzing historical data to uncover patterns and trends. This study proposes a machine learning approach leveraging the Pointwise Learning to Rank (PLR) method to improve the accuracy of draft predictions. A baseline linear regression model was implemented for comparison, and the final PLR model demonstrated superior performance, achieving a Mean Absolute Error (MAE) of 7.3 and an R2 score of 0.647 on the validation set, with similar results on the test set. These metrics indicate that the model explains 64.7% of the variance in draft rankings, significantly outperforming existing methods such as Decision Trees and XGBoost, which reported MAE values nearly twice as large. The improved results are attributed to robust data preprocessing, feature selection, and hyperparameter tuning, alongside the PLR method's task-specific optimization. The results demonstrate that machine learning techniques can accurately predict certain aspects of the draft rankings. While subjective factors influencing the actual outcomes cannot be fully captured, the proposed approach provides a solid foundation for future advancements and integration of additional qualitative data to enhance predictive performance.**

***Index Terms*—NBA Draft, Learning to Rank, Predictive Analytics in Sports, Player Evaluation, Basketball Analytics**

I. Introduction

The National Basketball Association (NBA) draft is a cornerstone event in professional basketball, providing an opportunity for collegiate and international players to transition into their professional careers. Our aim in this research was to understand the draft process from a data analytical point of view [1].

To deeper understand the main aspects of priorizing a given player, we developed a machine learning-based application to analyze and predict the draft order of college players. The application assigns individual rankings to players based on statistical metrics, such as offensive efficiency and overall performance. These preliminary predictions are then used to generate a comprehensive ranking system, which is then compared against historical draft orders.

It must be pointed out that our model accounts for a wealth of quantitative data, the real life draft outcomes are significantly influenced by subjective team preferences and situational factors: the presented model treats these effects as noise.

II. Related Work

The tak of NBA draft prediction has gathered attention from researchers who have applied machine learning techniques to address various aspects of the process [2]. Previous studies have primarily focused on either predicting whether a player will be drafted or attempting to predict a rank to the players based on their statistical performance.

Ke et al. [3] proposed a unified machine learning framework for constructing optimal basketball team rosters in the NBA and WNBA, utilizing unsupervised and supervised learning methods to cluster players and model team composition based on performance trends, pressure-handling ability, and salary constraints.

In the paper by Czasonis et al. [4] a relevance-based prediction system to assess NBA draft prospects was introduced, emphasizing transparency and reliability in forecasting performance through mathematical measures of importance and prediction reliability.

Mamonov [5] applied XGBoost to predict NBA draftee success. Maymin [6] developed a leave-one-out random forest model using scouting reports, combine metrics, and other variables to predict NCAA-to-NBA performance, demonstrating that model-driven analytics can outperform traditional team decisions in value creation.

Ellison published an empirical analysis [7] of NBA drafts from 2006 to 2014, comparing models to evaluate draft success based on college statistics and combine data, emphasizing cross-sectional differences across draft years.

Aditya Kumar's work [8] published on Kaggle forms a foundational reference for this study. In his work, Decision

Trees and XGBoost were applied to predict player draft positions. His approach involved preprocessing the dataset to include only drafted players and leveraging ten key statistical metrics, such as minutes played, free throw rate, and shooting percentages.

In a publication available on *Medium*, John Daniel describes his method using ensemble classifiers [9], including Logistic Regression, K-Nearest Neighbors, and Adaboost, to predict whether a player would be drafted. His analysis highlighted correlations between physical attributes (e.g., height and weight), statistical performance, and draft outcomes. While his models achieved an 82% accuracy in identifying drafted players, they were not designed to predict precise draft positions. Similarly to the publication of Kumar, articles from Marc Doucette [10] and Saadan Mir [11] explored draft prediction using advanced statistical analyses and modern machine learning models.

III. Methodology

There are multiple approaches for machine learning ranking (MLR) [12]. The Pointwise Learning to Rank (PLR) is a very simple method, similar to classical supervised learning. The loss can be formulated as

$$\mathcal{L}_{\text{PLR}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i), \tag{1}$$

where $\hat{y}$ and y refers to the predicted and actual ranks, respectively; the loss is given similarly as in case of the mean squared error.

Other methods, such as pairwise or listwise methods also exist [13], however these are significantly different approaches to the pointwise method. There are different popular applications to pairwise methods, such as facial recognition [14], [15], novel solutions are based on listwise approaches [16].

PLR is employed in this research, which prioritizes the relative ordering of players, enabling more detailed insights into the draft process. This approach allows for better handling of subjective factors.

This model serves as an ideal starting point for addressing ranking problems due to its simplicity and ease of implementation, as recommended in the literature [13]. These problems are treated as regression tasks, as the loss directly measures the distance between the expected values and the predictions.

This study outlines the detailed methodology used for developing a PLR based machine learning model to predict NBA draft rankings. The approach includes systematic data preprocessing, model selection, and ranking methodologies to achieve efficient predictions.

A. Data Processing

The dataset is historical college basketball player statistics between 2009 and 2021, available on Kaggle by Aditya Kumar [8], originally collected and published on the website of Bart Torvik [17]. It is worth mentioning, that college player statistics and especially the future path of young players are followed by multiple websites and is collected in different databases; the original dataset was extended by such data originating from the official website of the NBA [18].

1) Data cleaning, feature selection: Data cleaning involved removing irrelevant entries, such as players who did not declare for the draft, and handling missing values. Only players with sufficient statistical data and those who participated in the NBA Draft were retained, ensuring high-quality input for training. After an initial correlation analysis, it was discovered that among technical variables describing the picking round there are few variable showing weak correlation with the pick number: especially defensive and offensive ratings.

Key features, such as scoring efficiency (eFG%, TS%), offensive and defensive ratings, and advanced metrics like BPM (Box Plus-Minus), were identified. Pearson correlation analysis was used to measure the strength and direction of relationships between features and the draft position. Features with high correlation and statistical significance were prioritized.

B. Naive approach

To have a baseline analysis, a baseline neural network was prepared to perform linear regression. The network architecture consisted of 4 hidden layers with ReLU activations, while the single output neuron had linear activation. Dropout was applied to all hidden layers to deal with overfitting.

Training was applied to the baseline neural network using cross-validation to ensure model stability. To optimize performance, two key callbacks were utilized: *ReduceLROnPlateau*, which reduces the learning rate if the validation loss does not improve for several epochs, and *EarlyStopping*, which halts training when no improvement in the validation metric is observed within a specified timeframe. The Adam optimizer was employed, with the Mean Absolute Error (MAE) set as both the loss function and the evaluation metric. Training was conducted for a maximum of 1000 epochs.

The measured MAE was 14.31 for the validation set and 12.13 for the test set, meaning that the average difference between the actual and predicted draft order is above 10 spots.

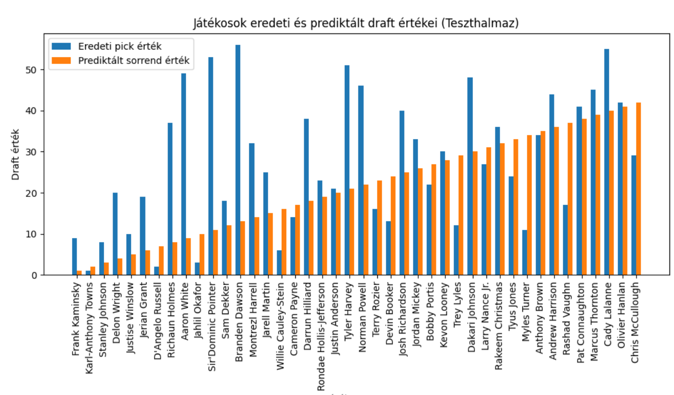

Fig. 1. Original and predicted draft pick order on the test set using the naive approach

Figure 1 visualizes the results on applying the linear regression model on the test set. It can be easily concluded that the model underperforms, especially when the expected pick number is small.

C. Pointwise Learning to Rank

The final model is based on PLR; implementation was done in Tensorflow and Keras.

1) Normalization: To address the varying magnitudes of features in the dataset, normalization was performed using RobustScaler, which is well-suited for handling outliers. This scaler removed the median and scaled features based on the interquartile range (IQR), ensuring that extreme values did not disproportionately influence the training process. The scaling parameters were calculated from the training set and then applied to the validation and test sets.

2) Data Preparation: The data was split into training, validation, and test sets to ensure robust evaluation. Sixty players were randomly assigned to the validation set, while the remainder formed the training set. The test set was composed of players from the 2015 NBA Draft, allowing evaluation on unseen data. To maintain consistency, a fixed random seed (random_state=42) was used, ensuring the same splits across multiple iterations.

3) Model Architecture: The model was implemented as a Pointwise Learning to Rank system, which predicts draft positions for each player based on their individual statistics. The Pointwise method independently evaluates each player, assigning a draft position as a continuous value without considering the relative rankings of other players. The model's architecture included an input layer, multiple hidden layers, and an output layer. The input layer accepted a feature vector representing each player's statistics.

The hidden layers were optimized for depth and neuron count, with the best configuration comprising three layers with 384, 448, and 160 neurons, respectively. Each hidden layer used ReLU activation to model complex relationships, and dropout layers with rates of 0.3, 0.2, and 0.4 were added to prevent overfitting. The output layer, employing a linear activation function, produced a single continuous value representing the predicted draft position.

4) Hyperparameter Tuning: Hyperparameter tuning was conducted using the Keras Tuner framework, which systematically explored parameter combinations to identify the optimal configuration. The tuning process resulted in the selection of the best-performing model parameters, which included the three-layer architecture, the aforementioned neuron counts and dropout rates, and a learning rate of 0.0064.

5) Training Process: The model was compiled with the Adam optimizer, configured to use the tuned learning rate. The Mean Squared Error (MSE) loss function was chosen to minimize large deviations in predictions, while Mean Absolute Error (MAE) was employed as an evaluation metric to quantify prediction accuracy. Two callbacks were integrated to enhance the training process: Early Stopping, which halted training after 20 epochs without improvement in validation loss, and ReduceLROnPlateau, which decreased the learning rate by a factor of 0.2 after 10 epochs of stagnating validation loss. The training process was conducted over a maximum of 100 epochs, though it typically converged earlier due to the Early Stopping mechanism. Validation data was used throughout training to monitor and fine-tune performance.

IV. Results and evaluation

The main aim of the PLR approach was to outperform the baseline approach, which employs a naive linear regression model for ranking predictions. The effectiveness of the PLR model is evaluated based on its ability to minimize ranking errors.

A. Numerical Results and Raw Predictions

The evaluation of the developed machine learning model is centered on its ability to predict NBA draft rankings accurately. To assess its performance, two well-established metrics were utilized: the Mean Absolute Error (MAE) and the R^2 score. These metrics are commonly used in regression tasks and provide clear insights into the model's prediction accuracy and its ability to explain the variability in the data.

The Mean Absolute Error quantifies the average absolute difference between the predicted and actual draft positions. This metric offers a straightforward interpretation, as it directly measures the typical deviation of the model's predictions from the true values [19]. Meanwhile, the R^2 score evaluates the proportion of variance in the target variable explained by the model. A higher R^2 value indicates that the model effectively captures the underlying relationships in the data, while lower or negative values suggest poor explanatory power.

On the validation set, the model achieved an MAE of 7.3003, meaning that the predicted draft positions were, on average, approximately 7 places away from the actual positions. The R^2 score on the validation set was 0.6472, indicating that the model explained about 64.7% of the variance in the draft rankings. These results were consistent when applied to the test set, where the model obtained an MAE of 7.9065 and an R^2 score of 0.6599. The slight increase in the MAE on the test set reflects the challenges of generalizing to unseen data but still demonstrates robust model performance with no significant drop in accuracy.

When compared to the solution of Kumar, which utilized Decision Trees and XGBoost models, the results of this study show clear improvements. Kumar's models achieved MAE values of 15.35 and 14.51 on validation and test datasets, respectively, which are almost twice as large as the errors reported in this study. The superior performance of the proposed model can be attributed to the Pointwise Learning to Rank approach, which directly optimizes the ranking task and leverages hyperparameter tuning for enhanced model configuration.

The evaluation results highlight the effectiveness of the Pointwise Learning to Rank method for this task. By combining robust data preprocessing, feature selection, and hyperparameter optimization, the model significantly outperforms previous approaches. However, the remaining prediction errors, reflected in the MAE values, underscore the influence of subjective factors, such as team preferences, which are not captured in the model. These subjective factors introduce variance in the draft process, limiting the model's ability to achieve perfect accuracy.

Detailed comparison results are shown in Table I. This table provides a detailed comparison between the original draft positions and the model's predicted pick values. The table offers a clear view of the foundation upon which the generated rankings were constructed.

TABLE I
RAW PREDICTION VALUES ON THE TEST SET

Player Name	Original Pick	Predicted Pick
Karl-Anthony Towns	1.0	12.326
D'Angelo Russell	2.0	14.817
Jahlil Okafor	3.0	16.311
Willie Cauley-Stein	6.0	15.590
Stanley Johnson	8.0	16.898
Frank Kaminsky	9.0	18.793
Justise Winslow	10.0	16.299
Myles Turner	11.0	17.808
Trey Lyles	12.0	21.487
Devin Booker	13.0	25.872
Cameron Payne	14.0	30.907
Terry Rozier	16.0	22.573
Rashad Vaughn	17.0	30.933
Sam Dekker	18.0	20.968
Jerian Grant	19.0	21.824
Delon Wright	20.0	25.649
Justin Anderson	21.0	33.853
Bobby Portis	22.0	34.142
Rondae Hollis-Jefferson	23.0	32.921
Tyus Jones	24.0	35.466
Jarell Martin	25.0	22.896
Larry Nance Jr.	27.0	35.130
Chris McCullough	29.0	31.797
Kevon Looney	30.0	37.404
Montrezl Harrell	32.0	27.117
Jordan Mickey	33.0	27.648
Anthony Brown	34.0	35.645
Rakeem Christmas	36.0	29.403
Richaun Holmes	37.0	34.784
Darrun Hilliard	38.0	20.345
Josh Richardson	40.0	36.293
Pat Connaughton	41.0	35.888
Olivier Hanlan	42.0	36.310
Andrew Harrison	44.0	51.663
Marcus Thornton	45.0	54.279
Norman Powell	46.0	25.437
Dakari Johnson	48.0	55.530
Aaron White	49.0	50.890
Tyler Harvey	51.0	52.112
Sir'Dominic Pointer	53.0	46.272
Cady Lalanne	55.0	55.206
Branden Dawson	56.0	48.627

B. Predicted Rankings

The results on the test set are visualized in figure 2, showcasing the ranking generated by the model. The observations demonstrating that the model generalizes effectively and produces relatively minor errors at the beginning and end of the ranking. However, the most significant discrepancies occur in the middle third of the rankings.

The test set contained only 42 players with available data. This discrepancy arose from the lack of information on players drafted from international leagues, as these players' statistics are not present in the American databases used for this study.

This data deficiency created notable challenges in constructing the ranking. With only 42 players available, the model attempted to allocate these players across the 60 draft positions. This led to inherent errors, as 18 positions were necessarily incorrect due to missing players. Moreover, players ranked beyond the 42nd original draft position could not occupy their actual positions due to the missing data.

To address these gaps, the final ranking was supplemented with the original positions of the missing players, labeled as "Missing" in the figure. This adjustment allows the visualization to present the complete draft order, providing a clearer understanding of the model's performance. The graph highlights both the accurate predictions and the discrepancies introduced by the incomplete dataset, effectively illustrating the model's ranking capabilities and its limitations in handling missing data.

V. CONCLUSION

In this paper, we aim to find a solution to predict NBA draft rankings based on historical observations of player statistics.

In the study a naive and a recommended machine learning model was implemented and introduced, with the pointwise learning to rank (PLR) method selected as best performing. This final model, chosen based on its learning outcomes, exhibited the most optimal predictive capabilities among the tested approaches. Performance metrics such as Mean Absolute Error (MAE) and R^2 were employed to quantify the model's accuracy and explanatory power, and the results highlighted its effectiveness compared to the baseline models initially developed during the study.

The findings underscore that machine learning techniques can, in specific cases, be effective tools for predicting NBA draft rankings. However, it must be emphasized, that in real-world scenarios, the draft process is influenced by numerous environmental factors that cannot be formalized or incorporated into a purely statistical framework. Our experiments demonstrated that statistical features, including offensive and defensive metrics significantly impact the draft rankings.

In conclusion, while the developed model successfully highlights the predictive power of statistical data, the complex and subjective nature of the draft process remains a significant challenge. This study provides a foundation for further exploration and suggests that integrating qualitative data and team-specific insights could enhance future predictive models and bring them closer to replicating real-world draft outcomes.

REFERENCES

[1] C. R. Florke and M. D. Ecker, "Nba draft lottery probabilities," *American Journal of Undergraduate Research*, vol. 2, no. 3, pp. 19–29, 2003.
[2] A. Hussey, "Nba draft analysis," Ph.D. dissertation, Dublin, National College of Ireland, 2021.

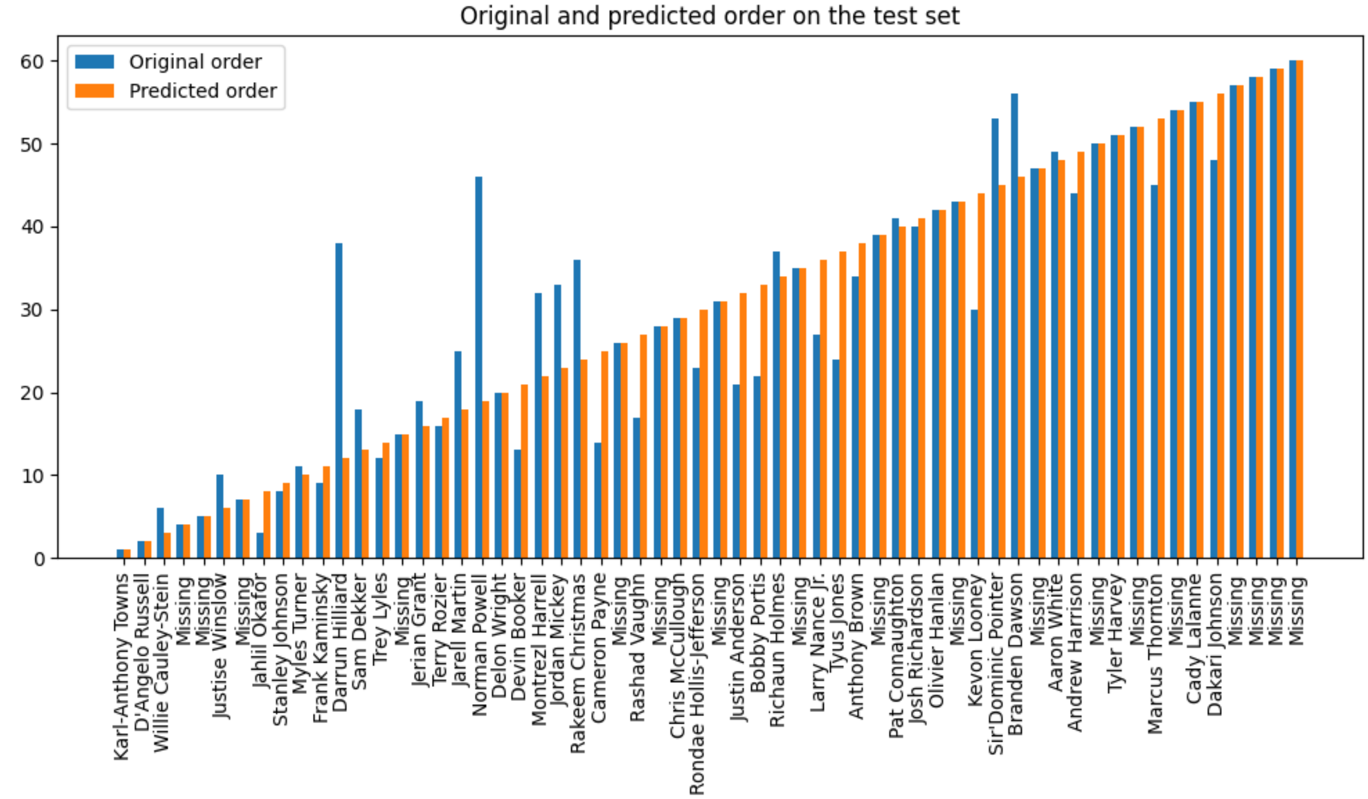

Fig. 2. Original and predicted order on the test set

[3] Y. Ke, R. Bian, and R. Chandra, "A unified machine learning framework for basketball team roster construction: Nba and wnba," *Applied Soft Computing*, vol. 153, p. 111298, 2024.

[4] M. Czasonis, M. Kritzman, C. Kulasekaran, and D. Turkington, "How to predict the performance of nba draft prospects," 2023.

[5] N. Mamonov, "Maximizing draft outcomes: An ml-based approach to nba draftee success," 2023.

[6] P. Maymin, "Using scouting reports text to predict ncaa→ nba performance," *Journal of Business Analytics*, vol. 4, no. 1, pp. 40–54, 2021.

[7] L. Ellison, "An empirical analysis of the nba draft from 2006-2014," *Empirical Economic Bulletin, An Undergraduate Journal*, vol. 13, no. 1, p. 5, 2020.

[8] A. Kumar, "Predicting the NBA Draft Using College Stats," https://www.kaggle.com/discussions/general/285535, [Online].

[9] J. Daniel, "Predict NBA Draft using Machine Learning!" https://medium.com/analytics-vidhya/predict-nba-draft-using-machine-learning-7023503e33e7, [Online].

[10] M. Doucette, "Machine Learning Applications in Predicting NBA Draft Picks, and Reflecting on Rushed Year," https://medium.com/@tmarcdoucette/machine-learning-applications-in-predicting-nba-draft-picks-and-reflecting-on-rushed-year-b61c2cf2e9f, [Online].

[11] S. Mir, "NBA Draft Analysis: Using Machine Learning to Project NBA Success," https://towardsdatascience.com/nba-draft-analysis-using-machine-learning-to-project-nba-success-a1c6bf576d19, [Online].

[12] T.-Y. Liu *et al.*, "Learning to rank for information retrieval," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[13] M. Ibrahim and M. Carman, "Comparing pointwise and listwise objective functions for random-forest-based learning-to-rank," *ACM Transactions on Information Systems (TOIS)*, vol. 34, no. 4, pp. 1–38, 2016.

[14] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 129–136.

[15] B.-N. Kang, Y. Kim, and D. Kim, "Pairwise relational networks for face recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 628–645.

[16] M. Buyl, P. Missault, and P.-A. Sondag, "Rankformer: Listwise learning-to-rank using listwide labels," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 3762–3773.

[17] B. Torvik, "T-Rank," https://barttorvik.com/, [Online].

[18] NBA Official website, "NBA," https://nba.com/, [Online].

[19] G. E. Box, J. S. Hunter, W. G. Hunter *et al.*, "Statistics for experimenters," in *Wiley series in probability and statistics*. Wiley Hoboken, NJ, 2005.