Ensemble Predictions of NBA Career Longevity

Christian Loth
Department of Computer Science
University of Texas at Dallas
Richardson, TX, USA
CML230001

Divyam Patro
Department of Computer Science
University of Texas at Dallas
Richardson, TX, USA
DAP210001

Navaneeth Kumar Buddi
Department of Computer Science
University of Texas at Dallas
Richardson, TX, USA
NXB210086

Akshata Kinage
Department of Computer Science
University of Texas at Dallas
Richardson, TX, USA
AXK210231

***Abstract*—This project uses data derived from NBA players to construct a predictive model, specifically using the Random Forest, the XGBoost, the ADABoost, and the bagging algorithms, to forecast whether a player's career will span more than five years or not. We will compare and contrast the results of each of these. We will utilize datasets containing key features that correlate with prolonged career longevity in the NBA. This can provide necessary information to NBA teams who are looking to make their draft picks, as well as their contract negotiations. Higher longevity in a player can significantly raise the value that they contribute to the team.**

***Index Terms*—NBA, Random Forest, predictive modeling, career longevity, player statistics**

I. Introduction and background work

The length of a professional basketball player's career is subject to a myriad of influencing factors within the game, encompassing variables such as number of games played, points earned in each game, etc. The significance of developing a robust model capable of effectively harnessing these diverse statistics cannot be overstated, as it serves as a crucial determinant in attaining accurate predictive outcomes. The objective of this research is to forecast the likelihood of a player's career extending beyond the 5-year mark. This predictive insight holds substantial value, facilitating the distinction between prolonged and fleeting professional tenures. Moreover, this valuable information helps team optimize their decision-making processes when contemplating the duration of player contracts.

In the realm of sports analytics, machine learning emerges as a powerful tool, promising insightful outcomes when applied to the dynamic landscape of professional basketball. Within the domain of machine learning, ensemble learning emerges as a particularly potent subset, showcasing significant efficacy in predicting models for sports characterized by vast datasets and an extensive array of player-specific features and variables. Ensemble learning methods such as ADABoost, XGBoost, Bagging, and Random Forests have garnered acclaim for their ability to mitigate overfitting and enhance prediction accuracy by amalgamating the decisions of multiple learners.

ADABoost, as one of the prominent ensemble learning techniques, leverages a series of weak learners to create a robust model capable of predicting a player's career longevity. By iteratively adjusting weights assigned to misclassified instances, ADABoost refines its predictive capabilities, offering a nuanced understanding of the factors influencing career duration.

XGBoost, another noteworthy approach within ensemble learning, excels in handling diverse data types and has demonstrated prowess in predicting outcomes in the realm of sports analytics. Its ability to integrate diverse features and variables makes it a valuable asset in forecasting a player's career trajectory.

Bootstrap Aggregating, or Bagging, is an ensemble learning method that constructs numerous models by training on diverse subsets of the training data. The technique then combines the predictions from these diverse models. This method proves particularly effective in reducing the impact of outliers and enhancing the overall robustness of career prediction models.

Random Forests, characterized by their utilization of multiple decision trees, operate collectively to yield predictions with heightened accuracy. The synergy of these individual decision trees results in a comprehensive understanding of the intricate relationships between various player-specific parameters and career duration.

The exploration of ensemble learning methodologies within the context of sports analytics holds immense promise for predicting the career longevity of NBA players. This research endeavor seeks to harness the capabilities of ADABoost, XGBoost, Bagging, and Random Forests to provide teams with a comprehensive understanding of the factors influencing a player's career span. By delving into these advanced predictive models, teams can make informed and strategic decisions, optimizing their business strategies and fostering a more sustainable and successful future in the competitive landscape of professional basketball.

II. Theoretical and Conceptual Study of the Algorithm

Ensemble learning is based on the idea that a collection of weak learning models can surpass the performance of any single model in the group. That is, when multiple weak learning models are combined, their performance can provide us with very notable classification results. By combining the predictions of multiple models, ensemble learning capitalizes on the strengths and balances out the weaknesses of individual models [1]. This method is particularly effective in scenarios where no single model provides consistently accurate predictions across different data sets. This is why we would adopt to use of multiple learners on the same dataset instead. The diversity among the models in an ensemble reduces the likelihood of overfitting and often allows us to obtain more accurate and reliable predictions [1].

Random Forest is a popular ensemble learning technique. It can create multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees [1]. This is the ensemble technique that we are going to be coding by hand and will be the biggest constructed portion of our project. Random forests can simulate additional randomness within their own set of data. When a node is split during the construction of a tree, the best split is chosen from a random subset of features rather than from all features [1]. After this, the majority vote is then chosen from the output of the decision tree. This strategy increases diversity among the trees in the forest, allowing for low variance, but high bias results. The Random Forest algorithm is particularly noted for its ability to handle large datasets with high dimensionality, and and has been effectively used in sports statistics.

ADABoost is an ensemble technique that focuses on improving the accuracy of learning algorithms [2]. It starts by training a weak learner on the initial dataset and throughout each round adjusts the weights of wrongly classified points. It then uses these rounds to produce a single strong hypothesis with compelling results. Through this process, more attention is given to the examples that previous learners misclassified [2]. The process of adjusting weights and adding learners continues until a desired model can be built on the data classification, or until a specified number of learners have been added. ADABoost is particularly effective in reducing both bias and variance, giving it a "best of both worlds" advantage.

XGBoost is an abbreviation for Extreme Gradient Boosting, an advanced implementation of gradient boosting algorithms [4]. It has gained significant popularity due to its speed and performance. XGBoost offers an accurate solution for gradient boosting, where new models are built to correct errors made by existing models [4]. It comes with features for handling sparse data, regularizing to prevent overfitting, and efficiently processing large datasets. XGBoost has proven to be an award-winning algorithm in machine learning competitions because of its ability to handle different types of data and distributions.

Bagging, also known as Bootstrap Aggregating, is an ensemble technique that improves the stability and accuracy of machine learning algorithms [3]. It involves creating multiple versions of a predictor and using these to get an aggregated predictor. Each model in a bagging ensemble runs on a slightly different subset of the data, ensuring that the models vary from each other [3]. These subsets are created by randomly sampling with replacements from the original dataset, which allows the simulation of new samples without actually having to capture more data. The final prediction is typically made by averaging the predictions of all models (for regression problems) or by a majority vote (for classification problems). Bagging is particularly effective in reducing variance, preventing overfitting, and improving the accuracy of decision tree methods, though it can be applied to many types of predictive models.

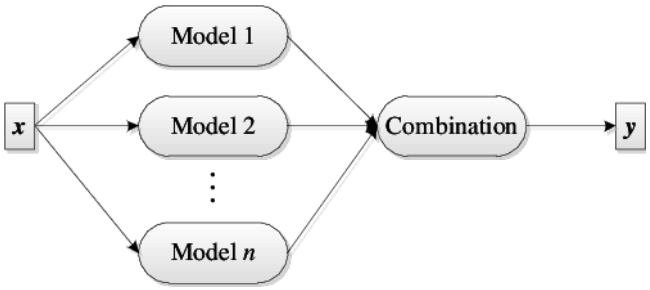

Fig. 1. General Structure of Ensemble Learning approach

III. Model Implementation

A. Random Forest Model Description

The Random Forest model is a robust and versatile ensemble learning algorithm tailored for predicting outcomes in the context of professional basketball player career longevity. The model is meticulously crafted from scratch to harness the complexities of the game, leveraging both bootstrap aggregating (bagging) and decision tree algorithms to achieve accurate and reliable predictions.

The backbone of the Random Forest model lies in its decision trees, which are constructed through a carefully designed recursive process. The decision tree's fundamental goal is to split the dataset into homogeneous subsets by selecting optimal features and split points. Two main methods are employed for this purpose one of which relies on a random selection of features, while the other adopts a more efficient approach by considering a subset of randomly chosen features. The splitting process continues until specific termination conditions, such as reaching the maximum depth or minimum samples for a split, are met.

The Random Forest model capitalizes on the power of ensemble learning by aggregating multiple decision trees. For each iteration, a bootstrap sample of the dataset is generated, and a decision tree is constructed. The resulting ensemble of trees collectively contributes to the predictive power of the Random Forest. The predictive strength of the Random Forest

Experiment Number	Parameters	Results
1	Estimators: 125, Depth: 5, Min Split: 18, Max Features: 3	Accuracy: 0.752, Precision: 0.780, Recall: 0.845, F1-Score: 0.811
2	Estimators: 80, Depth: 6, Min Split: 4, Max Features: 12	Accuracy: 0.737, Precision: 0.775, Recall: 0.821, F1-Score: 0.798
3	Estimators: 152, Depth: 13, Min Split: 17, Max Features: 18	Accuracy: 0.489, Precision: 0.603, Recall: 0.560, F1-Score: 0.580

TABLE I
RESULTS FOR RANDOM FORESTS MODEL FOR DIFFERENT MODEL PARAMETERS

Experiment Number	Parameters	Results
1	AdaBoost (n_estimators=50, learning_rate=0.5)	Accuracy: 0.7218, Precision: 0.7159, Recall: 0.7218, F1-Score: 0.71
2	AdaBoost (n_estimators=100, learning_rate=1.0)	Accuracy: 0.688, Precision: 0.68, Recall: 0.688, F1-Score: 0.6811
3	XGBoost (n_estimators=100, learning_rate=0.1)	Accuracy: 0.6692, Precision: 0.6606, Recall: 0.6692, F1-Score: 0.6624
4	XGBoost (n_estimators=200, learning_rate=0.05)	Accuracy: 0.6917, Precision: 0.6865, Recall: 0.6917, F1-Score: 0.6881
5	Bagging (n_estimators=5)	Accuracy: 0.5977, Precision: 0.5969, Recall: 0.5977, F1-Score: 0.5973
6	Bagging (n_estimators=20)	Accuracy: 0.6128, Precision: 0.6105, Recall: 0.6128, F1-Score: 0.6116

TABLE II
RESULTS FOR DIFFERENT MODELS AND COMBINATION OF PARAMETERS

model emerges from the consensus of its constituent decision trees. It employs a voting mechanism to make predictions based on the aggregated insights of individual trees. For each input instance, predictions from all decision trees are considered, and the final prediction is determined by the most frequently occurring outcome.

The model offers flexibility through the specification of key hyperparameters during its instantiation. These include the number of estimators, maximum depth of individual trees, minimum node size for splitting, and the maximum number of features to be considered during splitting. The model's performance is evaluated using the out-of-bag estimate, calculated as the average test error across all iterations. This provides a robust metric for assessing the model's predictive accuracy and aids in fine-tuning the hyperparameters for optimal performance.

B. Dataset Proprocessing

We utilized kaggle dataset describing a National Basketball Association(NBA) Basketball Players' Career Duration as the basis of our research. The data consists of performance statistics from each player's rookie year. The key attributes of this dataset include Number of games played, Number of minutes played per game, Average number of points scored per game, etc. The data set consists of 1,341 observations, each of which represents a distinct NBA player. The aim variable is a Boolean value that represents the likelihood that a specific player will play five years in the league.

We started by examining the dataset to have a comprehensive overview of the dataset, offering insights into the data types, general statistics, and the presence of any outliers. Then we systematically identified and quantified missing values within the dataset. This information is crucial for making informed decisions about potential imputation or removal of rows or columns with insufficient data. Additionally, we identified and subsequently eliminated duplicate entries, ensuring that the dataset remained free from redundancy and maintained data integrity. The generation of a visual representation of the

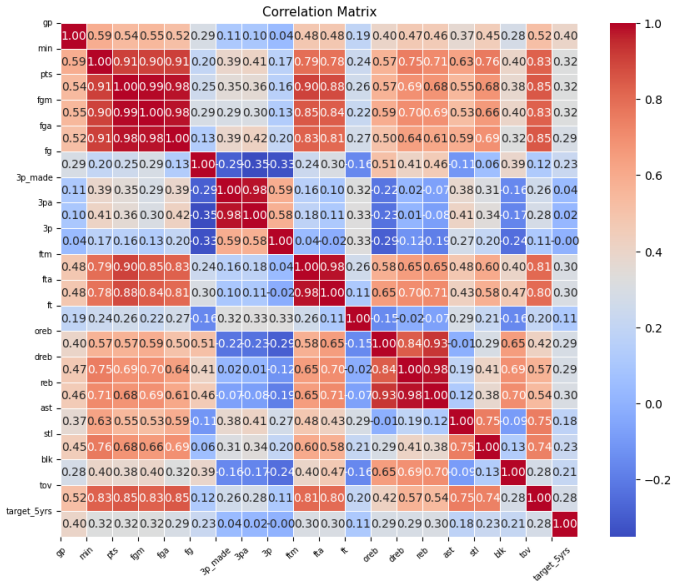

Fig. 2. Correlation matrix of all parameters of the dataset

correlation matrix, aided in the identification of highly correlated features. Finally, the dataset underwent normalization using the normalize method, ensuring that variables were on a consistent scale, which is particularly beneficial for machine learning algorithms that are sensitive to the magnitude of input features. Overall, these preprocessing steps laid a robust foundation for subsequent analysis and modeling, fostering a clean, organized, and optimized dataset for meaningful insights and predictions.

IV. RESULTS AND ANALYSIS

The custom Random Forest model, meticulously constructed from scratch, was subjected to a series of experiments to assess its performance under varying hyperparameter configurations. The model evaluation was based on key parameters such as the number of estimators, maximum depth of the trees,

minimum node size for splitting, and the maximum number of features considered during the splitting process. Each experiment aimed to explore the interplay of these parameters and their impact on the model's predictive accuracy.

The results of the experiments are presented in a comprehensive table I, providing an overview of the performance of random forest model across different configurations. The table includes experiment numbers, the specific hyperparameter settings for each model, and corresponding evaluation metrics, encompassing accuracy, precision, recall, and F1-score. These metrics offer a nuanced understanding of the model's ability to correctly classify instances, identify positive cases, and balance precision and recall. The ROC curve shown in Fig. 3 further visualizes the trade-off between true positive rate and false positive rate, providing a comprehensive depiction of the model's discriminatory power.

Next, we test popular ensemble learning techniques from the Scikit-Learn library. The Sklearn implementations, including ADABoost, XGBoost, and Bagging, were employed to assess their performance in comparison to the custom Random Forest model. Each algorithm was systematically evaluated across varying hyperparameter configurations, such as the number of estimators, learning rate, and maximum depth. The results, compiled in a detailed table, provide a holistic overview of each model's effectiveness in terms of accuracy, precision, recall, and F1-score.

The table II demonstrates the nuanced trade-offs between different models and their respective hyperparameter settings. Notably, ADABoost, XGBoost, and Bagging exhibit distinct strengths in capturing predictive patterns within the dataset, offering competitive performance across multiple metrics. The visual representation of Receiver Operating Characteristic (ROC) curve in Fig. 4 further enriches the analysis by illustrating the models' discrimination abilities.

While the custom Random Forest model showcases the flexibility of building models tailored to specific needs, the Sklearn implementations underscore the convenience and efficiency of leveraging well-established algorithms for predictive analytics. Researchers and practitioners can draw insights from this comparative analysis to inform their decision-making process when selecting the most suitable model for predicting professional basketball player career longevity.

V. Conclusion and Future Work

In conclusion, this study delved into the prediction of professional basketball player career longevity through a multifaceted approach, encompassing both custom-built Random Forest models and Sklearn implementations, including ADABoost, XGBoost, and Bagging. The comprehensive evaluation, considering a diverse array of hyperparameter configurations, highlighted the strengths and trade-offs of each model. While the custom Random Forest model showcased the adaptability of building models from scratch, Sklearn implementations demonstrated the efficiency and competitive performance of well-established algorithms.

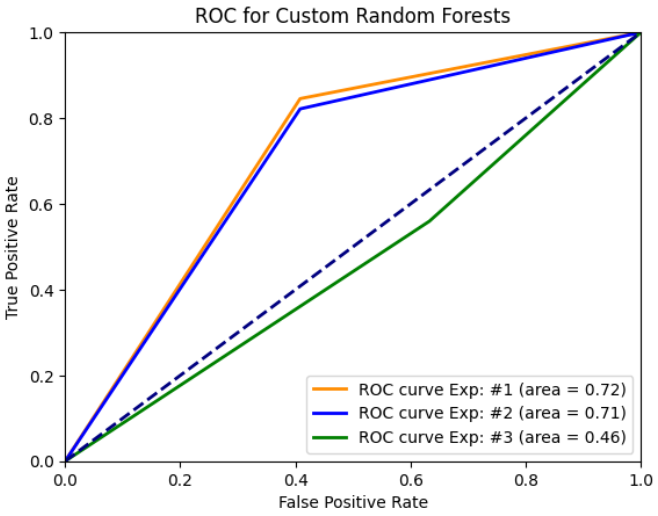

Fig. 3. ROC for Custom Random Forests Model

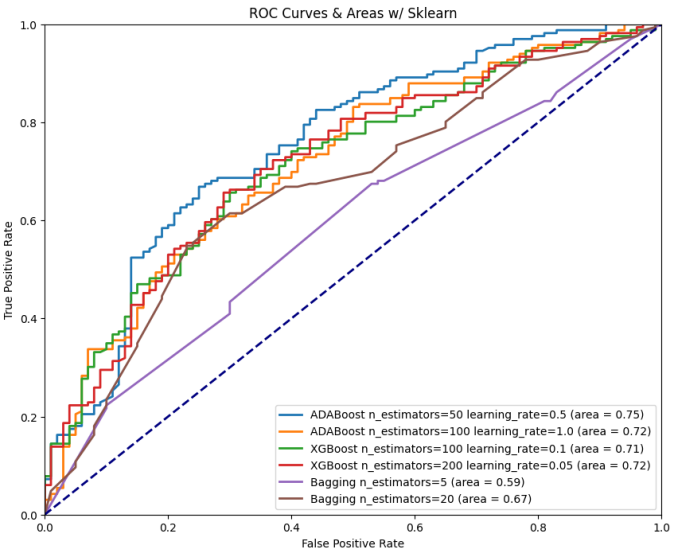

Fig. 4. ROC comparison and Areas with Sklearn for different models

Future work could extend this research by exploring additional ensemble learning techniques, experimenting with alternative feature engineering strategies, and incorporating more sophisticated methods for handling missing or ambiguous data. Moreover, the predictive models could be enhanced through the integration of player-specific data, such as injury history, training regimens and player statistics evolution over time. The inclusion of temporal aspects in the analysis could provide a more dynamic understanding of career trajectories. Additionally, an in-depth exploration of interpretability tools and techniques could shed light on the factors driving model predictions, fostering a deeper understanding of the features influencing player career longevity.

REFERENCES

[1] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," in IEEE Access, 2022. [Online]. Available: https://dx.doi.org/10.1109/ACCESS.2022.3207287

[2] F. I. E. Sari, F. W. Edlim, F. A. Ramadhan, Muhtadin, and D. A. Navastara, "Performance Analysis of Resampling and Ensemble Learning Methods on Diabetes Detection as Imbalanced Dataset," presented at the 2022 IEEE Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems (VECIMS), 2022. [Online]. Available: https://dx.doi.org/10.1109/ICVEE57061.2022.9930467

[3] H. Huang, L. Huang, R. Song, F. Jiao, and T. Ai, "Bus Single-Trip Time Prediction Based on Ensemble Learning," in Computational Intelligence and Neuroscience, vol. 2022, Article ID 6831167, 2022. [Online]. Available: https://dx.doi.org/10.1155/2022/6831167

[4] H. Jafarzadeh, M. Mahdianpari, E. Gill, F. Mohammadimanesh, and S. Homayouni, "Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation," in Remote Sensing, vol. 13, no. 21, 2021. [Online]. Available: https://dx.doi.org/10.3390/rs13214405

[5] https://www.kaggle.com/datasets/yakhyojon/national-basketball-association-nba/data