Analyzing the Future of NBA Players Through a Machine Learning Clustering Approach

Alex Bechakas, Albion Shoshi

College of Engineering, The Pennsylvania State University

University Park, PA 16802, USA

ajb8460@psu.edu

axs7352@psu.edu

***Abstract*— Evaluating the professional potential of college basketball players and young NBA players is critical for NBA teams, scouts, analysts, and the owner of the team. The study is using a two-phase machine learning driven framework to predict NBA career trajectories for NCAA collegiate basketball players and young NBA players using final career college basketball and early NBA statistics. The first phase consists of utilizing unsupervised learning through K-means clustering to categorize NBA players from 2000-2019 into five different performance tiers ranging from superstars/all-stars to minimal impact players/busts, based on their career statistics including career averages in points, assists, rebounds, and other impact metrics. The data driven clusters determine a player's "success" from measurable statistics and provide validated outcome labels from past completed careers in the NCAA and the NBA. The second phase uses supervised learning to train classification models we will use in Logistic Regression, Random Forest, and XGBoost. The models will be used on the NCAA collegiate statistics and the statistics from the first five years of the players' NBA careers, learning which performance indicators best predict success in the professional environment of the NBA. The trained models are applied to predict the future success of current NBA players under 22 years old using a dataset of NBA players who were rookies from the 2020-2025 seasons. By figuring out which collegiate metrics are most correlated with NBA success from different players in the NCAA, we can determine the potential a young player could have in the league. This research provides NBA teams and analysts with a different type of framework for evaluating prospects to aid the traditional scouting methods.**

Keywords: Players, Machine Learning algorithms, NBA, Classification, *Feature Selection, Supervised Machine Learning, Unsupervised Machine* ***learning.***

I. Introduction

The evaluation of basketball talent at the collegiate level represents one of the most challenging analytical problems in professional sports. Each year, NBA teams invest significant resources in young players based largely on subjective scouting assessment yet predicting which college players will succeed professionally remains highly uncertain. Traditional scouting methods suffer from inherent biases when evaluating talent and limited ability to quantify the complex features that determine NBA success. The transition from college to professional basketball involves dramatic changes in competition level, making historical performance an imperfect predictor of future outcomes. A collegiate player having high statistical metrics in the NCAA sometimes doesn't correlate directly into the NBA.

This project attempts to address a fundamental research question. Can we predict NBA career trajectory of young players using early NBA and NCAA collegiate basketball statistics? Machine learning is a powerful tool we can use to manage the large datasets of NBA and NCAA datasets and to predict player performance patterns that can be hidden. Machine Learning is very useful when it comes to predicting professional success in the NBA using collegiate metrics as well as supporting data driven decision making and creating rosters in the NBA.

The potential impact of having accurate predictions on young NBA players could be very important to stakeholders and the business side of the NBA. NBA front offices could use this framework to have more informed decisions about extending contracts and trade evaluations for players playing in their first few seasons by having measurable metrics for their potential career ceiling. It is also very important when it comes to player development programs in the NBA and allows coaches to have specific development programs into which players might increase their investment into the team. This same methodology can be transferred into different sports where predicting future performance from early career data still remains a challenging concept.

By using two decades worth of player data in the NBA and NCAA ranging from years 2000-2019, we developed a framework that uses unsupervised learning to categorize player outcomes into five different tiers based off their career statistics and supervised learning to map a combination of both collegiate statistics and early NBA metrics to the potential outcomes. The trained model is then applied onto our test set containing NBA players that played their rookie seasons from 2020-2025, who have both early NBA stats and NCAA collegiate statistics. In this paper, prediction of success in the NBA for young players is put into a two-phase machine learning approach using K-means clustering with classification models including Logistic Regression, Random Forest, and XGBoost.

II. Literature Review

Kannan, et al. [1] proposed a Machine Learning approach in attempting to predict the success of Division 1 NCAA collegiate basketball players at the professional level. Researchers used biometric combine data and NCAA season average statistics from 2009-2014 as features for success classification. This study defines a successful player as a player that moves on to play 174 professional basketball games in the NBA, which is the average number of games played within the dataset. Before modeling, the researchers performed a correlation test between biometric features to identify any highly correlated features for removal. After removing the highly correlated variable, a final subset of 12 biometric features and 18 game statistics were remaining to be used in the classification machine learning models.

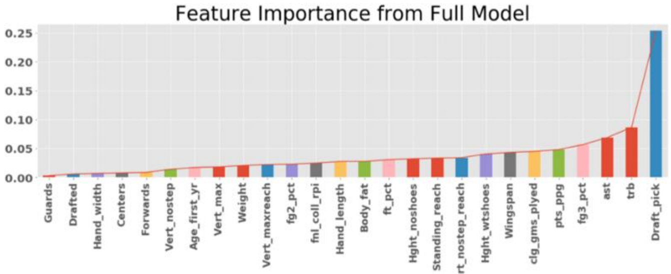

Fig. 1. Full Model Feature Importance Values

The researchers compared the results with the full data set and excluding the biometric data for logistic regression, random forest, and support vector machine, to test their hypothesis that the addition of biometric data would improve classification accuracy. The proposed hypothesis was supported through the random forest model—The most accurate model— as a 13% increase in accuracy and an increase in other supporting classification metrics was demonstrated while using biometric data. While observing the random forest model, they displayed the importance of features and found that the player's draft pick was significantly more important than any other feature used in the modeling. This strong importance of draft pick features could be too strong of a predictor with outside influence that impacts the meaning of the prediction. In a real situation, these predictions would occur before the draft process, excluding the draft pick feature, which would create a new setting for the prediction and revalue the remaining features. An observation created by the researchers was that many of the biometric features were not significant in relation to the model; the majority were overshadowed by the in-game statistics. Through the variable importance observation, the researchers determined that several features had insignificant importance levels as shown in Fig. 1, and if they were not addressed in the future, they could cause overfitting problems. The researchers successfully integrated biometric data into a machine learning approach in attempting to predict the success of Division 1 NCAA collegiate basketball players but have used a method that limits the models without a feature selection method. The future use of feature selection could improve the model's accuracy and create real world meaning to the important feature without using a Random Forest model.

Costa et al. [2] aimed to build off the limitations of Kannan's research by integrating machine learning models with a genetic algorithm to predict the success of NCAA basketball players to reach the NBA. Using decision tree and linear classifier algorithms, the goal of this research was to correctly predict the round the player was drafted based on NCAA basketball statistics as well as gathering the best subset of features for the classification model. Due to a large imbalance in classification data, a smaller pool of data was used with a reduced percentage in the skewed classification to prevent undertraining with a weak F1-score. The machine learning algorithms used were selected to diversify the approaches the models would take to gather a variety of results for further improvement.

The selected models were all tested with and without the genetic algorithm to create a comparison model, as well as testing multi-layer perceptron without the genetic algorithm to create a strong baseline for comparison. When the researchers tested the algorithms independent from the genetic algorithm, an average of 78% accuracy was obtained, but the remaining evaluation metrics were significantly lower and did not remain consistent throughout the model. The lack of consistency that was present could lead to misleading conclusions in cases where there is a large discrepancy between evaluation metric scores. The research demonstrates this discrepancy in the C4.5 model, as it displays a 92.81% precision with a significant drop off with a 47.29% recall.

The introduction of the genetic algorithm fixed the problem of inconsistent evaluation metrics and increased the accuracy and other evaluation metrics. The logistic regression model performed the best of the selected algorithms with the genetic algorithm, with an accuracy of 82.23% and supporting evaluation metrics with an average of 71%. The genetic algorithm granted the logistic regression a 4% increase in accuracy, a 7% increase in supporting evaluation metrics, and a subset of 6 features that display the traits with the best likelihood to predict the success of an NCAA basketball player. The hybrid model involving the genetic algorithm and logistic regression was outperformed by the multi-layer perceptron, but the slight decrease in accuracy is made up for in the form of the feature subset and runtime. The genetic algorithm does not require gradient information to converge to optimal features and hyperparameters, which drastically cut down on the runtime while maintaining high accuracies. While the differences in the best model with the genetic algorithm were not as high as the multilayer perceptron, the interpretability and comprehensive ability of the model increased. The feature subset allows for a better understanding of what to look for in NCAA basketball players as well as increasing the overall accuracy. The features selected in the subset are analyzed by the Shapley Additive Explanations explainability tool, which gives the magnitude of the features impacts on the model. The results from the SHAP model can present an interpretation of how valuable features are to the model, as shown in Fig. 2 which demonstrates the Logistic

Regression model's feature subset. Future improvements that the researchers aim to work towards are looking to increase the number of machine learning models tested and increase the features to involve advanced metrics as they are developed.

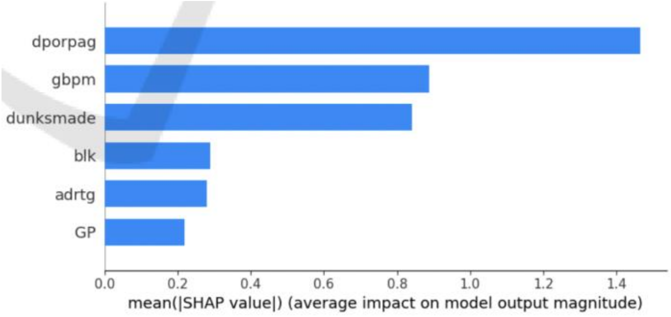

Fig. 2. Logistic Regression Results obtained from SHAP.

Basketball machine learning models struggle to add defensive features beyond rebounds and steals due to the lack of advanced tracking metrics that analyze the off-ball presence of a player. Li [3] proposed a hybrid model that would detect the frequency of defensive strategies, such as traps and switches, that will be able to support the groundwork for individual defensive grading in the future. Through a hybrid model featuring Long-Short-Term Memory and Convolutional Neural Networks, the research attempted to gather improved accuracy from already existing defensive strategy classification models.

The hybrid model used engineered features based on the NBA's SportVU Tracking System, a large-scale dataset including 32,000 offensive possessions recorded through a 25 Hz imaging system that maps all players and the ball position in a 400 x 360 grid. The features track the players' position, angle in relation to the ball, velocity, and the density of players around the ball. The sequential data for each possession, formatted as a time series vector, was input into the Long-Short-Term Memory to capture movement patterns and positioning across time. Simultaneously, the spatial grid representation is input into the Convolutional Neural Network to detect spatial patterns and identify traps effectively. The output vectors were combined in a fully connected layer to produce a powerful vector that is used for the final classification of a defensive strategy.

The results gathered from the research displayed an increase in accuracy from previous models classifying defensive strategies. The previous best performing model detected defensive traps at an 86% accuracy rate, and the proposed hybrid model identifies traps and switches at a 91.4% accuracy rate. The model was also compared with Random Forest and each model in the hybrid model to create a baseline for comparison. The hybrid model significantly outperforms the baseline comparison models in every evaluation metric as shown in Fig. 3. The model improved classification accuracy as well as presenting new beneficial information that previous models did not display. In game tendencies such as player density with respect to the court positioning, average location of defensive strategies, and the frequency of defensive strategies. These descriptive characteristics can be used to specialize in defensive or offensive strategies for game planning purposes and to get an advantage over your opponent. In future works, an advanced defensive metric can be generated from the additional player mapping data for individual player effort that can be used in player classification modeling. The current modeling successfully improves the classification of defensive strategies through a proposed hybrid machine learning model involving Long-Short-Term Memory and Convolutional Neural Networks.

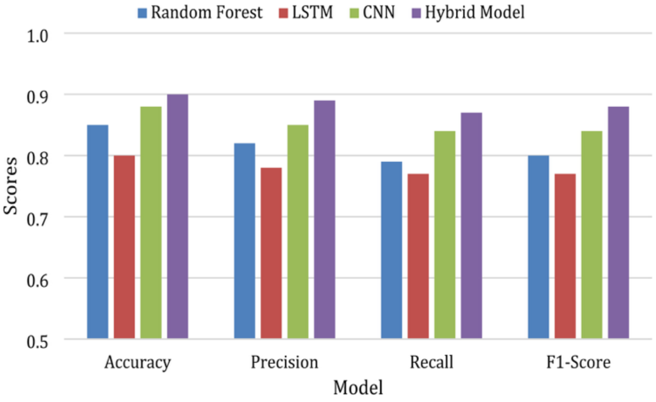

Fig. 3. Comparison of accuracy, precision, recall, and across models.

The framework for this research was created by Kannan is the analysis of player success at the NBA level through a machine learning approach. This framework allows for expansion through other works in additional machine learning models or classification methods. Costa introduced expansion to Kannan's research through a new classification value and introduced feature selection to improve its accuracy and its interpretability. The new classification value introduced is the round the player was selected in the NBA draft, but it can be further evaluated as the level of player that the athlete is expected to be based on clustering or other methods. The increase in classification levels requires the features to exhibit strong correlation in prediction, which will require advanced metrics on both sides of the ball. The advanced defense rating that Li's research can extrapolate can be greatly influential on future classification models through accurately representing values that are not explicitly given through the game of basketball. As sports-based research continues to improve, the innovations in Kannan's framework will continue to improve and allow for further evaluation of players and a stronger classification value to increase the interpretability of models to the real world.

III. Contributions

The literature reviews provided a strong baseline for the foundation of our machine learning based research, but there were limitations that were present in the literature reviews that we aim to eliminate. The literature reviews all proposed thorough ideology surrounding the future predictions of the NBA players, but had visual data bias or a lack of interpretability involved in the modeling. These limitations are eliminated in our innovative approach along with providing additional information. Through combining the ideas from the literature reviews with our proposed methodology we aimed to improve the interpretability of the classification and analyze future NBA prospects potential.

A. Parent Paper Implementation

The original code and dataset used in our selected parent paper were not publicly available, which prevented direct innovation of the parent paper code. In an attempt to replicate the parent paper with up-to-date data, we had to come up with an initial plan to gather the data through scraping publicly available data from various sources. After gathering all the data and replicating the machine learning code present in the parent paper, our plan was to further improve interpretability of the classification model by changing the initial approach through a cluster approach. The parent paper approach created limitations that prevent the model from having impactful real-world implications and viability, that we attempt to solve in our innovation.

B. Parent Paper Limitations

Our plans for implementation came directly from looking to solve the various limitations presented in the parent paper code and dataset. The first limitation present in the parent paper is the lack of data involved in both the training and testing dataset. The data used in the parent paper was collegiate basketball players that participated in the NBA biometric combine, which consisted of 194 players in a six-year span of data. The lack of current players that participate in the biometric combine in full presented a lack of data that could be used for classification leading to undertraining the models. The features present in the dataset not only cause the limitation of a lack of data but also create interpretability and bias limitations. The biometric data, which was shown as less significant than the in-game statistics, creates the previously mentioned data limitation, and the main predictor of “Draft Pick” creates a strong real-world bias towards the model. The “Draft Pick” predictor is a feature that is based off the player's skill level, while we are attempting to do a prediction of a player's future skill level. This feature is essentially using an already established prediction to make a similar prediction. This also impacts interpretability of this model as a model of this nature will be used in a pre-draft process, which would eliminate this feature completely. The interpretability is also impacted by the final classifier of defining success as a binary value opposed to the level of which the player succeeded. When looking at future NBA prospects, a team will want to have an estimate of how the player will be compared to others and not just what is essentially a one-word answer about the player's potential. This limitation of success being ambiguously defined was the main motivator for our implementation and modification of the parent paper code.

C. Final Implementation

Our new goal for attempting to clear the classification ambiguity was to classify future NBA stars using a 5-level clustering approach. After averaging NBA players' seasonal data from 2000-2019, we used k-means clustering to classify the players into five clusters. These clusters represent the level at which the player is statistically performing in the NBA relative to players around the league. The clustered NBA player data is used alongside NCAA player data to predict the future potential of a young NBA player through machine learning models. The models we selected were Logistic Regression, Random Forest, and XGBoost. The selection of Random Forest and Logistic Regression is based off those models having the highest resulting accuracy scores in the parent paper, and XGBoost was selected due to the potential in high complexity of the modeling. The large number of features limiting the potential of Logistic Regression and Random Forest was identified as a potential problem, so XGBoost was chosen to potentially solve that problem.

IV. Data set

The data set being used in our implementation is derived from the parent paper data set with additional inclusions of data. The paper used five years of NCAA basketball player statistics along with the NBA’s biometric combine data, which limited the players available to be used in classification substantially. To avoid this problem of limited players, we used NBA player seasonal statistics and NCAA seasonal player statistics from 2000-2019 for the training dataset and from 2020-2025 for the testing set. The elimination of biometric data was essential to conducting proper research for predictions; with the dataset being significantly undersized when included. The final training and testing sets that were used come from joining the clustered NBA seasonal player averages and the collegiate seasonal averages to be used in classification models. The training set was limited to the players' first 5 NBA season averages to better align with the predictions of the test data with at maximum only 5 seasons of development. These datasets are not limiting the other entries as the NCAA data was derived from the list of players in the NBA dataset to not redundantly increase runtime and storage. The final datasets consist of 1,137 players in the training set and 247 players in the testing set, both having 17 features for classification.

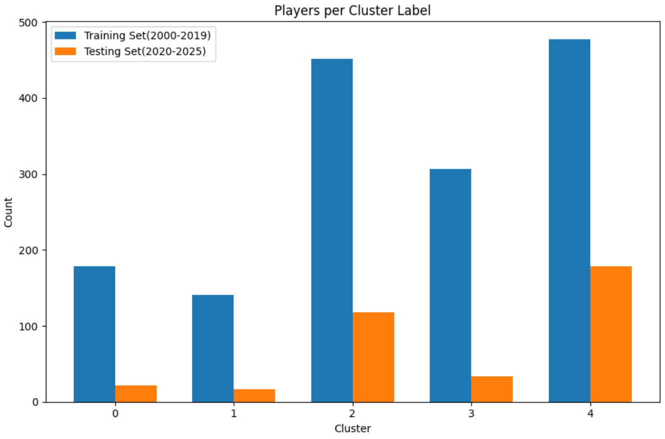

Fig. 4. Displaying the number of players in each cluster in both data sets

V. Implementation

The implementation of this research focuses on the process away from improving the parent paper code. The implementation process looks to expand on our contributions and describe the coding process in depth. The goal was to predict the future of young NBA players based on their college and early career NBA statistics through machine learning. The accuracy of our models is recorded in the same fashion as the parent paper, recording accuracy, precision, recall in F1-Score. Keeping these evaluation metrics the same as the parent paper is essential for comparison to the parent paper, although there is a trade-off in accuracy for model interpretability.

A. K-Means Clsutering

As we diverged from the parent paper as well as the other reference papers, we wanted to develop an interpretable classification value that has more meaning than the players' draft position. The classification value will establish the rank of the players' gameplay value with respect to the other players in the NBA throughout the given data set. A silhouette score test was developed to determine the optimal k-value for clustering where 4 was determined to be the optimal number of clusters based on the elbow method as displayed in Fig. 5. Although this a 4-cluster algorithm performed optimally inside the training set, the variability in the testing set prevented the 4-cluster approach from clustering accurately, forcing a pivot to our initial plan of 5 clusters. The clusters were labeled from worst to best in relation to their values as, Bust, Bench, Role Player, All Star/Starter, and Superstar, but beginning as arbitrary labels that will be reassigned once clustered. With the optimal clusters determined the NBA player data is standardized and a K-Means clustering method is performed creating initial clusters of players that need to be reassigned a label. The established clusters are analyzed to compare the average points and minutes, as those statistics are highly impactful on modeling and classification. Principle component analysis is done on the clustered data to observe the relation between clusters as well as limiting noise and overtraining the data.

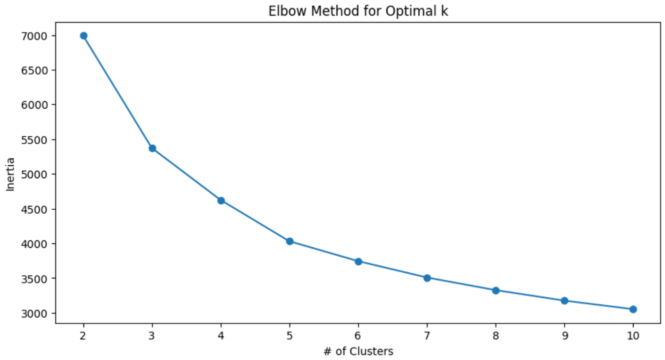

Fig. 5. Optimal Clusters for K-Means Clustering

Before the supervised learning can occur, the clustered data is joined with the NCAA seasonal average statistics to be used in predictions. As clustering before this join in data can be beneficial in the accuracy of classifying similar players across leagues, this research is looking to predict players at the NBA level, and not to find similarities at the college level. This can be put into real world perspective by finding similar players that have strong college careers but are unable to translate their game over to the NBA level. In order to not overvalue those players, clustering is exclusive to NBA statistics. The training set is also limited to only using the statistics from the first 5 years of the players' careers to ensure the training set better resembles an early career player and what their expectations as a career are. In our initial approach, we did not limit the training set to the 5-year split that we settled on, which resulted in players being under-classified in the predicted classification. This classification problem was due to the general trend of a player's career being able to be represented by a normal distribution bell curve. Early in a career, a player will have lower stats while still developing their skills until hitting their statistical peak in the middle of their career before slowing down due to old age. Using the first 5 years of the players career allows us to capture the "left tail" of the bell curve and display all the players in the development stage, similar to where the players in the test set currently are in their career.

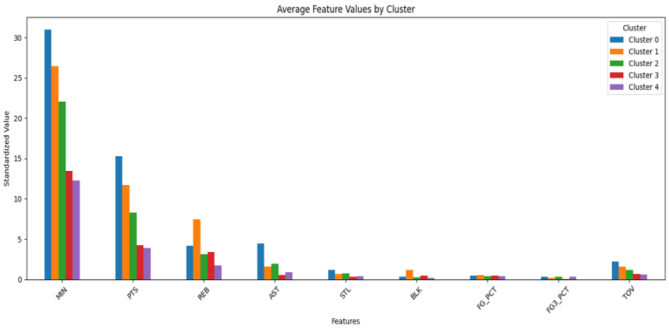

Fig. 6. Average Feature Values by Cluster.

B. Machine Learning Models

The machine learning models that we selected were based off the success of the parent paper modeling as well as some logical intuition on expansion. The machine learning models that we selected were Logistic Regression, Random Forest, and XGBoost. Random Forest and Logistic regression were the top two models in classification in parent paper, and as we were expanding off the parent paper code, the translation of accuracies should be present. XGBoost was an external model that we selected as it should perform well with strictly numerical data, as well as be able to identify non-linear patterns in the data which will lead to improved classifications. Along with the potential for increased classification accuracy, XGBoost also provides feature importance that can help the interpretability of the model and lead to assisting models to be created.

In the parent paper, logistic regression performed the second best of the three selected machine learning models, but in our implementation, it struggled performing the worst out of the three models selected. With an overall accuracy of only 45.75%, it drastically underperforms compared to the parent paper, which will be a common theme, and underperforms when comparing it to the other machine learning models of our implementation. The Random Forest model had the next highest accuracy resulting in an overall accuracy score of 51.42%. There is an improvement from the Logistic Regression as expected from the parent paper, but as previously mentioned still struggles to obtain accurate results. As predicted from our intuitive hypothesis when deciding on what models to use, XGBoost had the highest accuracy score at 51.82%. Along with the overall accuracy of the test set being measured, we observed the precision, recall, and F1-score of the individual classifications. A consistent theme across the different machine learning models was the decreased F1 score for the "Role Player" classification due to misclassifying the players as busts. Another area of inconsistency was the varying evaluation metrics for "Superstar" and "All-Star/Starter" due to the lack of players representing those clusters in the testing set. All of the model evaluation metrics were visually displayed in a confusion matrix, where all of them provided similar results due to the data distribution.

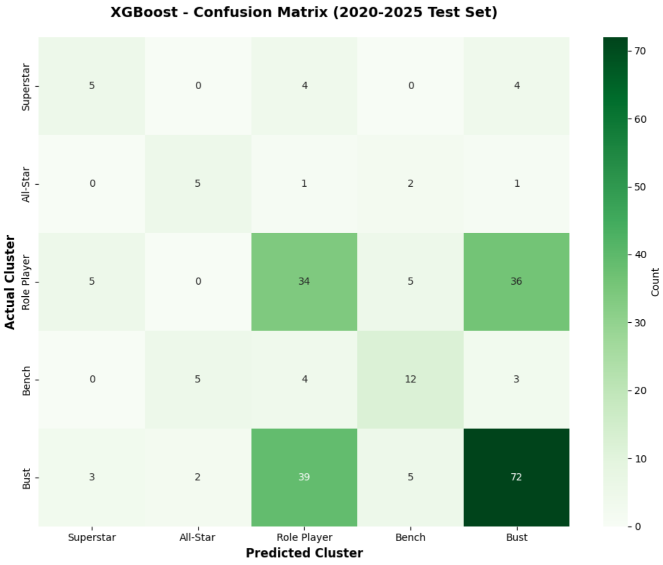

Fig. 7. Confusion matrix created by XGBoost's predictions.

When creating the machine learning models, we also obtained the importance of features for each machine learning model. These importance scores do not visually tell the entire story of why they vary in accuracy score, but some NBA connections can be interpreted to develop an idea. These important statistics can also follow the trend of the current NBA scene and can explain why they are the most important features. XGBoost, the most accurate model, displayed 3-point attempts as the most important feature; that was more than double as important as the next feature of blocks. 3-point attempts are a large part of the NBA and have recently evolved to be an even larger aspect of successful teams' offensive strategy. XGBoost picks up on the increased trend of 3-point attempts becoming more frequent among good players and heavily weighs the feature. Moving to the feature importance of Random Forest, field goal percentage, 3-point attempts, and blocks are the top 3 features without any significant gaps between them. Random Forest, similarly, to XGBoost, emphasizes 3-point attempts, but also efficient scoring. Efficient scoring tends to directly correlate with players that shoot a lot, as the most accurate players will be in the game plan to get more shots. Logistic Regression does not deviate off the offensive shooting trends with the most important features being field goals attempted, 3-point attempts, and assists. As scoring is the most offensive statistic in basketball, it makes sense that all these models would put large amounts of emphasis on the sporing methods.

All of the models observe similar features to be the most important features, with all of the top features involving scoring methods. When focusing on our most accurate model, XGBoost is unique to the other models due to the large gap that is created between its top feature and the next most important. It values 3-point attempts as the most important feature but uniquely values defensive statistics of blocks and steals as the next most important features. Not only did XGBoost uniquely heavily weigh the top feature but also focused greatly on defensive strategy rather than weighing more scoring methods as top predictors.

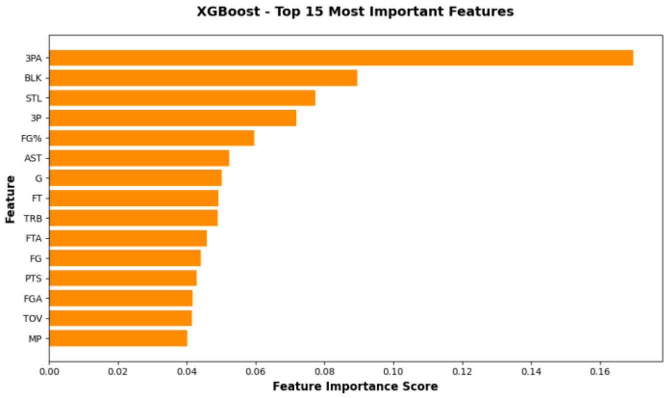

Fig. 8. XGBoost Top 15 Most Important Features.

Identifying the top features can increase the interpretability of the modeling by giving important features to the visual scouting team to look at the context of the statistics. NBA scouting consists of analytical and visual scouting, and the identification of these top features can help the visual scouting team to follow through with high predicted players. The visual scouting team would receive the results from XGBoost and would be able to analyze the context of the statistics in the superstar and all-star players. They may look at the 3-point attempts for context of the statistics and determine the status of the player based on statistics and visuals.

VI. Conclusion

The goal of our research was to create a machine learning model using clustering that would predict young NBA players' careers based on their NBA and collegiate basketball statistics. Using a 5-cluster approach on the players' NBA data, we then limited the NBA data to the first 5 years of the players' career and joined the players collegiate stats before moving on the modeling. Once the clustering dataset was finalized, we tested Logistic Regression, Random Forest, and XGBoost, where XGBoost performed the best with 51.82% accuracy on the testing set. The machine learning models also presented feature importance where XGBoost determined that 3-point attempts were the most important feature by a significant amount. While this modeling did not meet the accuracy results of the parent paper of 64%, it provides an increased interpretability using clustering opposed to a binary success classifier.

VII. Future Works

The project showcases a foundation for predicting NBA success based on college statistics, but additional extensions could be added to improve accuracy and generalization of our research question. One of the more impactful enhancements can be adding position-specific modeling. By training seperate models based on positions and their appropriate features, we could possibly achieve higher accuracy improvements. Emphasizing rebounds and block features for centers, assists and steals for guards, and balanced efficiency for wing players, players can be properly evaluated amongst each other. A center averaging 12 points per game and 10 rebounds per game could be more valuable than a guard player with identical stats, but our model ends up judging all players by a single standard. The position-based modeling approach could remove cross-positional comparisons and possibly give us better insights into predicting young players success.

Beyond position specific modeling, integrating advanced defensive metrics can capture essential aspects like perimeter defense, defensive positions, and help rotations that offer insight beyond traditional box score defensive statistics in steals and blocks. We aim to improve the current model's accuracy by integrating these defensive metrics alongside the already established steals and blocks. Players in the league like Draymond Green and Rudy Gobert built their historic NBA careers from elite defense. Their college statistics however understate the true defensive talent they posess because steals and blocks don't fully capture defensive excellence in the sport of basketball. The current models we have struggled in identifying defensive stars because their offensive statistics are subpar. Including advanced defensive metrics like defensive rating, defensive plus/minus, opponent field goal percentage when defended, and defensive win shares could provide complete judgment of a player's impact on the court. The addition of an advanced defensive metric will improve classification and reduce the importance value of blocks which is standing out as potentially wrongful important feature. However, the development of the metric will strongly impact the way the feature interacts.

Finally, we can also look into adding additional machine learning based models to help better understand the complex relationships between college and NBA success. Multi-Layer Perceptron's (MLPs) could find these feature interactions and nonlinear patterns that our current models might miss or not see. MLP could discover that a high scoring player with high efficiency ratings in ways our models can't. Neural Networks in general can capture nonlinear relationships and can find patterns in complex ways. Support Vector Machines (SVMs) can find optimal decision boundaries between career tiers in high dimensional feature space. SVMs are robust to outliers and perform extremely well with sparse and small datasets. SVMs are also prone to overfitting and can potentially handle both linear and nonlinear classifications. The increase in machine learning models will lead to an increase in accuracy in a corresponding model. Multilayer Perceptron is likely to be the best performing model when comparing results to the previously reviewed papers. Along with additional accuracy, we will receive a proper subset of important features that display real world interpretability.

References

[1] **A. Kannan, B. Kolovich, B. Lawrence**, and S. **Rafiqi, "Predicting National Basketball Association Success: A Machine Learning Approach," SMU Data Science Review: Vol. 1: No. 3, Article 7**, 2018.

[2] D. de Araujo Costa, J. Macedo Fechine, J. Rubens da Silva Brito, J. Victor Ribeiro Ferro, E. de Barros Costa, and R. Vilhena Vieira Lopes, "A Machine Learning Approach Using Interpretable Models for Predicting Success of NCAA Basketball Players to Reach NBA" Proceedings of the 16th International Conference on Agents and Artificial Intelligence Vol 3, 2024.

[3] J. Li, "Machine learning-based analysis of defensive strategies in basketball using player movement data" Scientific Reports Vol 15, 13887, April 22, 2025