CSE623 Machine Learning

Project 9: Athlete Profiling NCAA

Group 10

Anant Kedia	AU2220040
Maanit Shah	AU2220043
Samarth Bankar	AU2220001
Ayush Chaurasia	AU2220032
Viraj Bhatia	AU2220038

18th March, 2025

Problem Statement

The objective of this project is to analyze the NEC and MAAC conference game data (including all players and teams) for over the past four years and perform athlete profiling. This framework will be helpful for the recruiters to match player against their team specifications and select accordingly.

Deliverables

- Identification of individual team strengths and weaknesses (offense/defense).

Problem Statement

The objective of this project is to analyze the NEC and MAAC conference game data (including all players and teams) for over the past four years and perform athlete profiling. This framework will be helpful for the recruiters to match player against their team specifications and select accordingly.

Deliverables

- Identification of individual team strengths and weaknesses (offense/defense).
- Identification of athlete strengths and weaknesses.

Problem Statement

The objective of this project is to analyze the NEC and MAAC conference game data (including all players and teams) for over the past four years and perform athlete profiling. This framework will be helpful for the recruiters to match player against their team specifications and select accordingly.

Deliverables

- Identification of individual team strengths and weaknesses (offense/defense).
- Identification of athlete strengths and weaknesses.
- Based on opponent, identified strengths and weaknesses and game score, predict optimal athlete lineup for a team against each opponent.

Problem Statement

The objective of this project is to analyze the NEC and MAAC conference game data (including all players and teams) for over the past four years and perform athlete profiling. This framework will be helpful for the recruiters to match player against their team specifications and select accordingly.

Deliverables

- Identification of individual team strengths and weaknesses (offense/defense).
- Identification of athlete strengths and weaknesses.
- Based on opponent, identified strengths and weaknesses and game score, predict optimal athlete lineup for a team against each opponent.
- Athlete clusters - report general patterns.

Literature Review

Paper Name	Approach	Key Features	Limitations
Wang, J. (2023, October). Predictive Analysis of NBA Game Outcomes through Machine Learning. In Proceedings of the 6th International Conference on Machine Learning and Machine Intelligence (pp. 46-55).	Different ML approaches, such as Logistic Regression, Support Vector Machines, Deep Neural Networks, and Random Forests.	Comprehensive evaluation of various different ML models.	The approaches do not capture real world factors like player injuries, and team dynamics.
Islam, M. R., Ahmed, M. U., & Begum, S. (2024). iXGB: improving the interpretability of XGBoost using decision rules and counterfactuals. In 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) (Vol. 3, pp. 1345-1353).	iXGB, an approach to enhance XGBoost's interpretability by approximating decision rules from its internal structure and generating counterfactuals.	Enhances model interpretability.	The approximation of decision rules may not capture all the complexities of the model.
Ouyang, Y., Li, X., Zhou, W., Hong, W., Zheng, W., Qi, F., & Peng, L. (2024). Integration of machine learning XGBoost and SHAP models for NBA game outcome prediction and quantitative analysis methodology. Plos one, 19(7), e0307478.	XGBoost with SHAP.	Predictive ability of XGBoost integrated with SHAP to improve interpretability.	The performance indicators (e.g., field goal percentage, defensive rebounds) are not universally applicable.
Chen, W. J., Jhou, M. J., Lee, T. S., & Lu, C. J. (2021). Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association. Entropy, 23(4), 477.	Hybrid model with different approaches including KNN, XGBoost, and SGB.	Hybrid model to improve prediction accuracy.	Interpretability due to the complex nature of the proposed hybrid model.
Hu, H., Dimitrov, G., Menn, D., & Wu, S. NBA Player Performance Prediction Based on XGBoost and Synergies.	XGBoost	Player interactions quantified as synergies to enhance the prediction model.	Quantifying synergies for a larger set of players is difficult. Also, the method is extremely sensitive to outliers.

Dataset Discussion

The dataset consists of NEC and MAAC conference Basketball match data from 25/11/20 to 25/03/24. Each match is represented by the two teams playing it, and the teams' respective information such as the athletes participating in the match, various statistics of their performance in the match (points scored, assists, steals, blocks, etc.). Furthermore, the teams' total score in the match and whether they won or lost is present in the data. Finally, an aggregate score of each athlete's performance, called the Game Score is displayed. This quantity is calculated as follows:

Dataset Discussion

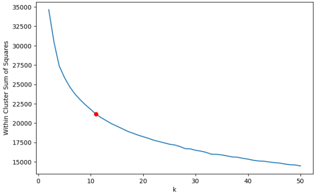
The dataset consists of NEC and MAAC conference Basketball match data from 25/11/20 to 25/03/24. Each match is represented by the two teams playing it, and the teams' respective information such as the athletes participating in the match, various statistics of their performance in the match (points scored, assists, steals, blocks, etc.). Furthermore, the teams' total score in the match and whether they won or lost is present in the data. Finally, an aggregate score of each athlete's performance, called the Game Score is displayed. This quantity is calculated as follows:

$$\begin{aligned} Game\ Score = {} & PTS + (0.4 \cdot FGM) - (0.7 \cdot FGA) – (0.4 \cdot (FTA - FTM)) \\ & + (0.7 \cdot OREB) + (0.3 \cdot DREB) + STL + (0.7 \cdot AST) \\ & + (0.7 \cdot BLK) – (0.4 \cdot PF) – TO. \end{aligned}$$

This is known as Hollinger's Game Score.

Approaches

K-means Clustering
We applied k-means clustering to the dataset on individual athletes' statistics (averaged over all their appearances) and estimated the ideal k value to be 11 via the "Elbow Method", however other unsupervised clustering metrics showed the performance of the algorithm to be unsatisfactory. Similar results were obtained after dimensionality reduction via PCA to 2 components/features. We assume that the poor results are due to the Curse of Dimensionality, as k-means utilizes distances.

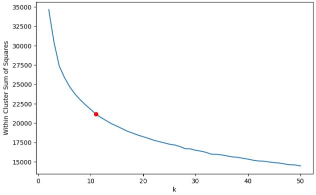


We also implemented DBSCAN for clustering, however we got similarly disappointing results.

Approaches

K-means Clustering

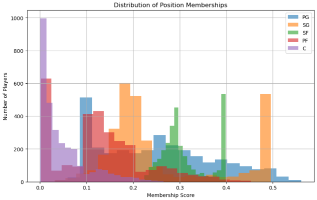
We applied k-means clustering to the dataset on individual athletes' statistics (averaged over all their appearances) and estimated the ideal k value to be 11 via the "Elbow Method", however other unsupervised clustering metrics showed the performance of the algorithm to be unsatisfactory. Similar results were obtained after dimensionality reduction via PCA to 2 components/features. We assume that the poor results are due to the Curse of Dimensionality, as k-means utilizes distances.



We also implemented DBSCAN for clustering, however we got similarly disappointing results.

Fuzzy C-means Clustering and ANFIS

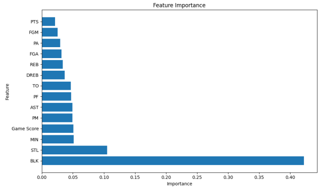
We used Fuzzy C-Means (FCM) to cluster athletes based on key performance metrics (FG%, PTS, 3P%, AST, REB, STL, BLK) and assigned soft membership scores to five basketball positions (PG, SG, SF, PF, C). Then, we trained an ANFIS model to predict a player's position membership based on these features, helping recruiters understand player strengths and fit within a team.



Approaches and Future Work

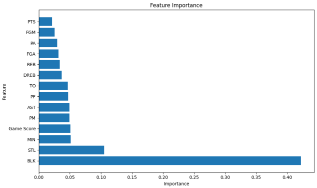
Random Forest Prediction

We applied random forest prediction to dataset on individual athletes' statistics (averaged over all their appearances) and trained the model over their number of wins. The model had a R^2 value of 0.45, which is undesirable. This is most definitely due to the fact that it match wins are a result of team efforts, and the statistics of individual athletes is not enough to predict them.



Approaches and Future Work

Random Forest Prediction
We applied random forest prediction to dataset on individual athletes' statistics (averaged over all their appearances) and trained the model over their number of wins. The model had a R^2 value of 0.45, which is undesirable. This is most definitely due to the fact that it match wins are a result of team efforts, and the statistics of individual athletes is not enough to predict them.

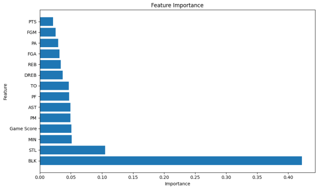


Future Work

- XGboost based winner prediction.

Approaches and Future Work

Random Forest Prediction
We applied random forest prediction to dataset on individual athletes' statistics (averaged over all their appearances) and trained the model over their number of wins. The model had a R^2 value of 0.45, which is undesirable. This is most definitely due to the fact that it match wins are a result of team efforts, and the statistics of individual athletes is not enough to predict them.



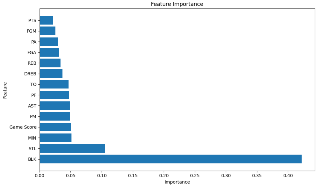
Future Work

- XGboost based winner prediction.
- Improving the explainability of the clusters created.

Approaches and Future Work

Random Forest Prediction

We applied random forest prediction to dataset on individual athletes' statistics (averaged over all their appearances) and trained the model over their number of wins. The model had a R^2 value of 0.45, which is undesirable. This is most definitely due to the fact that it match wins are a result of team efforts, and the statistics of individual athletes is not enough to predict them.



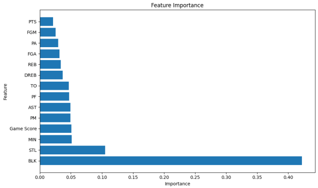
Future Work

- XGboost based winner prediction.
- Improving the explainability of the clusters created.
- Possible approaches to explore in order to improve explainability: SHAP, Neurosymbolic AI, LIME

Approaches and Future Work

Random Forest Prediction

We applied random forest prediction to dataset on individual athletes' statistics (averaged over all their appearances) and trained the model over their number of wins. The model had a R^2 value of 0.45, which is undesirable. This is most definitely due to the fact that it match wins are a result of team efforts, and the statistics of individual athletes is not enough to predict them.



Future Work

- XGboost based winner prediction.
- Improving the explainability of the clusters created.
- Possible approaches to explore in order to improve explainability: SHAP, Neurosymbolic AI, LIME
- Explore potential ensemble learning approaches to improve prediction accuracy

References

- "Glossary — Basketball-Reference.com," Basketball-Reference.com. `https://www.basketball-reference.com/about/glossary.html` (accessed Mar. 17, 2025).
- ANFIS GitHub Repository — Tim, "`https://github.com/twmeggs/anfis`" (accessed Mar. 17, 2025).