CISC 455 NBA team Rank Prediction Using Evolutionary Algorithm

STEVEN WEN*, FEITING YANG*, and YUEHAN QI*, Queen's University, Canada

The popularity of sports games is immense and spans across the globe. Predicting the outcomes of sports games can provide valuable insights into the strengths and weaknesses of teams, which can help fans, club managers, and financial industries make informed decisions. In this paper, we propose a generic programming (GP) algorithm to efficiently predict the ranking results of NBA playoffs. Our proposed model demonstrates better mean square error (MSE) performance compared to traditional machine learning models. By identifying the most important features for NBA scoring, our algorithm can efficiently predict ranking results. The findings of this research demonstrate great usefulness for stakeholders in the NBA ecosystem, including fans, teams, and financial institutions.

CCS Concepts: • **Computing methodologies** → **Genetic Algorithms**; Machine learning algorithms.

Additional Key Words and Phrases: Genetic Programming, Machine Learning, Prediction algorithms

ACM Reference Format:
Steven Wen, Feiting Yang, and Yuehan Qi. 2023. CISC 455 NBA team Rank Prediction Using Evolutionary Algorithm. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Background

The National Basketball Association (NBA) is a professional basketball league in North America, where teams are ranked based on their win-loss record at the end of each season. However, the ranking system only takes into account the final outcome of the season and does not provide any insight into the performance of teams throughout the season. This creates a need to develop a more comprehensive ranking system that considers the team's performance throughout the season.

1.2 Objective

The objective of this project is to develop a genetic programming algorithm that can rank the NBA teams based on their performance in the previous year. The proposed GP algorithm aims to provide a comprehensive ranking system with higher accuracy compared to other methods. The proposed solution will allow teams to make strategic decisions, provide fans with a better understanding of the teams, and enable the betting industry to evaluate the odds for betting with better insights. It also aims to enable teams to make informed decisions, provide fans with a better understanding of the teams, and help the betting industry to evaluate the odds for betting.

*All authors contributed equally to this research.

Manuscript submitted to ACM

Table 1. Methodology in Sports Games Modeling and Prediction using Genetic Programming[1]

Data-set	Regular season performance statistics (16 statistics and 35 years (1984 to 2018) of sample) of each team in the NBA.
Model	A tree-based GP model to output the results of the playoffs for given input data of a season.
Parent Selection	Tournament selection
Fitness Evaluation	fitness = $\sum_s e_{\text{season}(s)}$,s is the season in the training set
Operator	Crossover
Mutation	Random mutation
Termination	Stagnation termination (stop if the best fitness remain unchanged for 300 consecutive generations)

1.3 Implication

The proposed solution has significant implications for various stakeholders. Firstly, the team managers can use the ranking system to make informed decisions regarding player trades, contract negotiations, and other strategic moves. Secondly, the fans can benefit from the algorithm's insights by gaining a better understanding of the teams' strengths and weaknesses. Lastly, the betting industry can use the algorithm to assess the odds and make more informed decisions, resulting in more accurate predictions and potentially increased profits.

2 LITERATURE REVIEW

Sports Games Modeling and Prediction using Genetic Programming[1]

2.1 Summary

The research proposes a new method of using genetic programming (GP) to predict sports results, specifically the final outcome of NBA playoffs. By utilizing regular season performance statistics of each team and historical data, the algorithm is able to achieve a good prediction accuracy and provide insights into what factors are more influential in winning the game. The GP algorithm can be used to predict various sports games as long as they have a similar games bracket as NBA playoffs. The research has implications for fans, financial assets, and club coaches and managers who can benefit from an analytical tool that suggests more efficient and suitable strategies to win.

2.2 Methodology

The study indicates that it employs per-possession and per-minute statistics, which are deemed more informative than per-game statistics for conducting analyses. Furthermore, certain features are selectively dropped due to their lack of suitability in basketball scoring preference. For the training method, it adopts a weight-based approach for assigning weights to different ranks, and utilizes a 7-fold cross-validation technique to assess the effectiveness of the model. The data-set is partitioned into 6 parts for training and 1 part for testing, and the algorithm is repeated 10 times. Table 1 shows the design of the model. After training, the study tests 72 combinations of parameters to tune the model for the best performance and reaches the size of population as 2000, mutate rate as 0.3, and tournament size as 24.

2.3 Result

The overall performance of the genetic programming algorithm improves with the error and standard deviation decreases over time. The algorithm's success rate reveals that 52 out of the 70 best evolved models possess predictive accuracy surpassing 40%, with 21 of these models boasting accuracy exceeding 60%. The best performing model has a fitness score of 24064 and a final championship prediction accuracy of 0.8. Furthermore, with the most frequent occurrence in the following three features: DRtg(defensive rating), eFG(efficient field goal percentage), FT/FGA(free throws per field goal attempt), it indicates that Effective defense and efficient shooting are critical factors for winning the championship. The model accurately predicts the ranking of most teams in the 2008 season, including the final champion, the Boston Celtics, the runner-up, the Lakers, and the conference finalist, Detroit.

3 EA DESIGN

3.1 Overview

Our goal is to use the GP algorithm to forecast the order of all clubs in the playoffs using regular season data from the 2021–2022 season. We use a terminal set of 26 features

- Teams
- Games
- MinutesPlayed
- FieldGoals
- FieldGoalAttempts
- FieldGoalPercentage
- ThreePointers
- ThreePointAttempts
- ThreePointPercentage
- TwoPointers
- TwoPointAttempts
- TwoPointPercentage
- FreeThrows
- FreeThrowAttempts
- FreeThrowPercentage
- OffensiveRebounds
- DefensiveRebounds
- TotalRebounds
- Assists
- Steals
- Blocks
- Turnovers
- PersonalFouls
- Points
- year
- Rank

and a tree-based GP technique for the evolutionary process. Basic mathematical operations including addition, subtraction, multiplication, and division are included in the function set.

- `Representation of Chromosomes`: We aimed to predict the rank based on 26 features, and for this purpose, our Chromosome is expressed as a tree-based formula.
- `Initialization`: To initialize the Chromosome formula, we randomly selected elements from the function and terminal sets, with a depth of 3.
- `Selection`: For both parent selection and survivor selection, we currently employ the random selection method.
- `Fitness function`: To calculate fitness, we computed the total difference between the predicted and ground truth values. Since smaller errors indicate better fitness, the fitness score was calculated as 1 divided by the total difference.
- `Genetic Operators`: We used crossover as our genetic operator, which swaps subtrees at a chosen point with a rate of 0.7. Additionally, we used mutation to change the selected subtree with a new generated tree, with a mutation rate of 0.1.
- `Evaluation`:To compare the performance of our machine learning algorithm with others, we used mean square error as the evaluation method for the best formula obtained after evolution.

3.2 EA results

For the performance of GP we analyze the mean squared error (MSE) of the GP algorithm. By finding the maximum of the fitness value, the figure1 shows the trending fitness value of the operation.

From the best evolved predictive models, we choose the best prediction model with the lowest MSE value. The best individual found by the algorithm is "DefensiveRebounds ∗ FieldGoalPercentage" with an MSE value of 33.2 and a fitness of 0.0048This model is able to predict the ranking of NBA teams with reasonable accuracy. The model predicts the final rankings of the 30 NBA teams for the given season. The predicted rankings are based on the rating given to each team by the model. The rating of each team is calculated by feeding their regular season statistics into the best individual formula. The predicted rankings are compared with the actual rankings for the MSE. Which is shown on the right of figure 2 After comparing our data with the actual rankings, which are displayed in the graph below, we observed that many of the ranks were miscalculated within a 3-rank margin of error. In the graph, the red bars indicate the actual rank, while the green bars represent the rank with less than a 3-rank error. Upon further investigation, we discovered that the rankings were based on combining teams from both the Eastern and Western conferences, which may have resulted in some inaccuracies. For example, the Golden State Warriors won the playoffs in 2021-2022, but are not ranked first on this map. Therefore, we believe that using a 3-rank margin of error is a more accurate reflection of the rankings than the given ranks. According to the best individual formula, the most important features for predicting team ranking are DefensiveRebounds and FieldGoalPercentage. These features have a direct impact on a team's performance in terms of points scored and points conceded.

The algorithm was able to find the best individual formula in a running time of 1.13 seconds, which is fast and efficient for such a complex task. (without the graphing with graphing the time dropped to 2.97seconds depends on when the graph was closed)

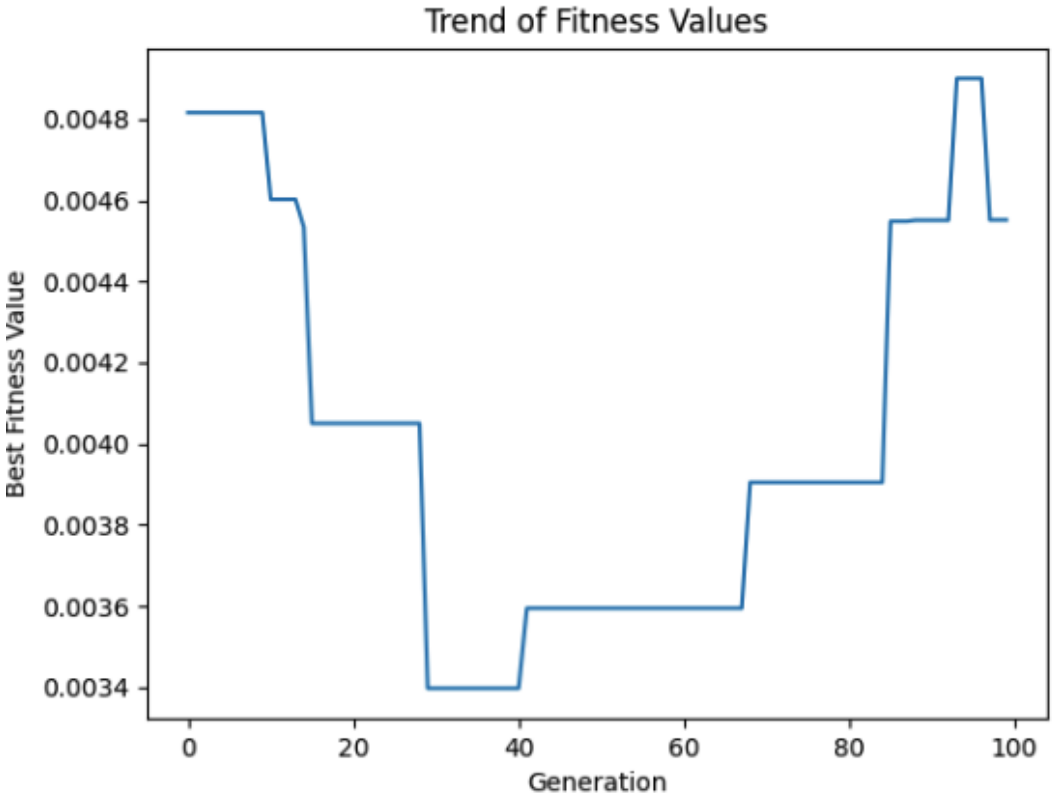

Fig. 1. Trend of the fitness value from the algorithm

4 COMPARISON

4.1 Comparison between selection methods

We have three methods for selecting parents and three methods for selecting survivors. The Table 2 below displays the performance measured by MSE and running time of various combinations of parent and survivor selection methods.

4.2 Comparison between Mechine Learning

In comparison to machine learning, the evolutionary algorithm has a Mean Squared Error of 33.2, which is slightly lower than the MSE of machine learning at 33.83. However, the evolutionary algorithm has an accuracy of 0%, which is a poor result for a supervised model. Additionally, the tree-based machine learning algorithm has a faster runtime than the evolutionary algorithm. Thus, a direct comparison between evolutionary algorithms and machine learning is challenging. Each approach has its own unique advantages and disadvantages, and the selection between them depends on the specific requirements and challenges of the problem at hand.

Fig. 2. Comparison of the Rank

Table 2. Comparison parent and survivor selection methods

Method	MPS	tournament selection	Random selection
$\mu + \lambda$	mse:78.73	mse:51.4	mse:97.13
selection	time:3.03s	time:9.15s	time:4.85s
Replacement	mse:65.53	mse:150.86	mse:33.2
selection	time:3.17s	time:3.00s	time:2.46s
Random	mse:55.0	mse:101.93	mse:104.8
selection	time:24.43s	time:9.40s	time:21.46s

5 DISCUSSION AND FUTURE WORK

After reviewing the results section, it is apparent that both machine learning and EA fail to generate accurate predicted values. One possible reason is the amount of data used was insufficient as we only used data from the 2021-2022 season. To produce more precise predictions using both machine learning and evolutionary algorithms, a more extensive

amount of training data is required. Another reason is that the evolutionary algorithm only employed two features(DefensiveRebounds and FieldGoalPercentage) for formula generation, which may have resulted in a significant amount of relevant information being disregarded. To improve the performance in the future EA algorithm, setting the constraint of the tree based individual formula may lead to comprehensive use of training data. Also, adding the coefficient value can make the formula more accurate. And a more effective way to improve the performance is to increase the amount of training data.

REFERENCES

[1] Shengkai Geng and Ting Hu. 2020. Sports Games Modeling and Prediction using Genetic Programming. In *2020 IEEE Congress on Evolutionary Computation (CEC)*. 1–6. https://doi.org/10.1109/CEC48606.2020.9185917

6 GITHUB

The Project code link: https://github.com/Stevenn2333/NBA-TEAM-Rank.git

Received 11 April 2023