ECE204_24

Basketball Player Performance Prediction

Kevin Fobare, Mukhtar Suleman, Ryan Pappania, Rodney Boone, *University of Rochester*

***Abstract*—The goal of this project was to create a machine learning model that could be used to predict the performance of any player solely based on their previous performances in games throughout their season and the opponent they were currently facing. The idea was inspired by the many wins and losses taken by one of our team members in sports betting. The National Basketball Association (NBA) was chosen as the sport to use for our model because of its ability to capture precise data of the performance of the players, familiarity and popularity of the sport, along with the endless source of information available. Our model uses every detail metric; from a player's shooting percentage, assist percentage, total rebounds, how many minutes the player played to each NBA team's defensive performance; defensive percentages, steals per game, blocks per game, etc allowing us the chance to get the most accurate results possible. After trying different machine learning models, we decided Random Forest was the best to use for our model.**

I. INTRODUCTION

Despite the controversy and setbacks in the industry, sports betting has grown throughout the 21st century, making people very wealthy and broke at the same time. Even people who aren't even sports fans to begin with, see it as a means of making some extra money. This is mainly due to the artificial intelligence suggestions that novices use to place their bets on sports. Without any knowledge of the player or history of the sports team, hundreds of thousands of people put their hard-earned money in the hands of these predictive models; our project aimed to make a model with the or higher reliability rate in performance by providing our own machine learning model. Other projects have been preformed such as 'Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle' by Ge Cheng et al. [1] and 'Predicting Outcomes of NBA Basketball Games' by Eric Scot Jones [2]. These two predict games using the maximum entropy principle and logistical models. These two provided inspirations but a different route was taken as described below.

II. METHOD

The first major step we took was gathering credible performance data for our chosen NBA players LeBron James, Kevin Durant, Giannis Antetokounmpo, Nikola Jokic, and Jayson Tatum. We chose these specific players for their consistent high performance for all or most of their career, and the availability of their performance data about their games. The data of their field goals, total assists, total rebound, and all the other player performance data was sourced from the website basketball-reference.com [3]. The defensive performance data was sourced for all the 30 NBA teams, such as defensive efficiency, points allowed for game, steals per game, blocks per game, and other metrics.

The next stage of our project after gathering our sample data was to begin programming. After importing our data for the five players, we decided to use Linear Regression as a baseline for predicting performance. This model produced a generalized result but due to its simplicity it produced accuracy that was undesirable. So to predict player performance in points, rebounds, and assists and ultimately predict game results, a random forest model was used. A random forest model generates many decision trees, how many trees can be adjusted. The branching within the trees chooses randomly from a subset of the features [4]. This randomness provides differences within the trees to prevent over and under fitting. When these trees are summed, it leaves just one decision tree which is then used to make our predictions of players performance.

III. PRELIMINARY RESULTS

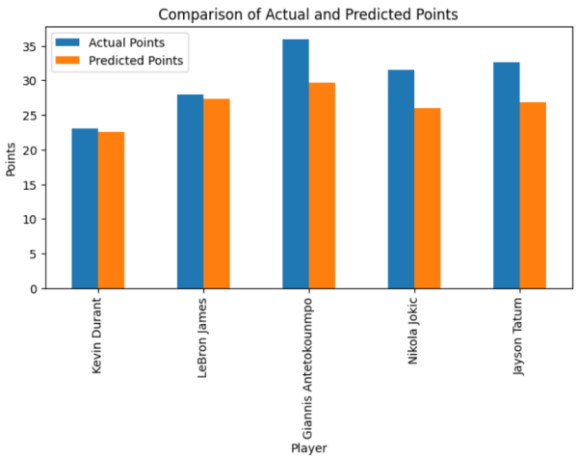

Figure 1 – Bar graph of players actual vs predicted points performance generated with a linear regression model

Results were generated from a linear regression model and our random forest regression model. Figure 1 + 2 show the predicted and actual results from points scored. The actual points scored was the average of the players performance over the regular season against the Washington Wizards. The predicted points is what our model predicted the player to score against the same team. The first figure shows the actual vs predicted for linear regression and then the second figure shows the same thing with the random forest regression. The predicted points more closely represents their actual performance against this team with the random forest model because it takes into account the opponent's defensive statistics. In this example, the opponent has a below average

defense so when we factor into account their defensive ranking it increases the players predicted points to more closely represent how they would perform against that team. We also predicted players' assists and rebounds against the same team and compared them to the actual numbers.

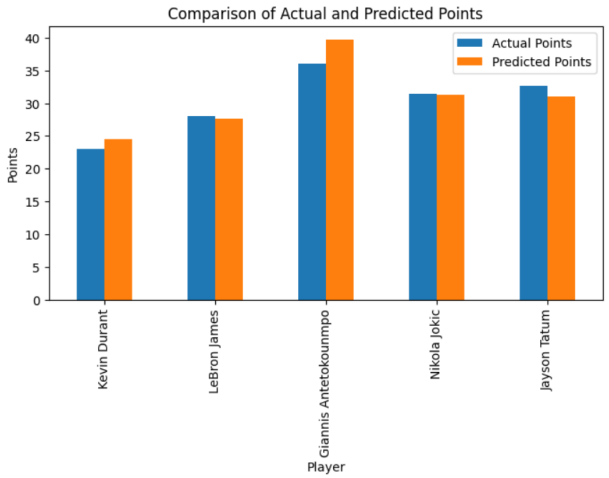

Figure 2 – Bar graph of players actual vs predicted points performance generated with a random forest regressor

Figures 3 and 4 show predicted assist and rebounds. As it can be seen the predicted vs actual values are quite close building confidence that this model is accurate enough to predict real life performance.

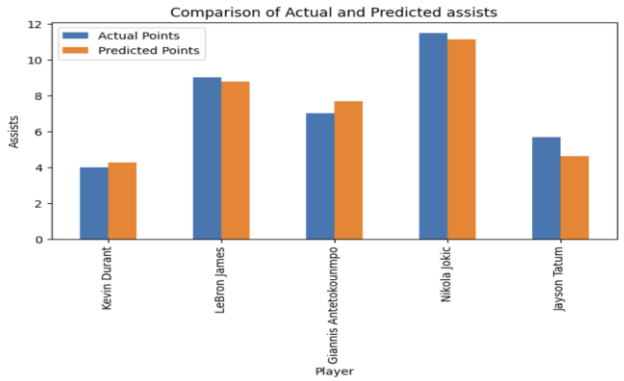

Figure 3 – Bar graph of players actual vs predicted assists performance generated with a random forest regressor

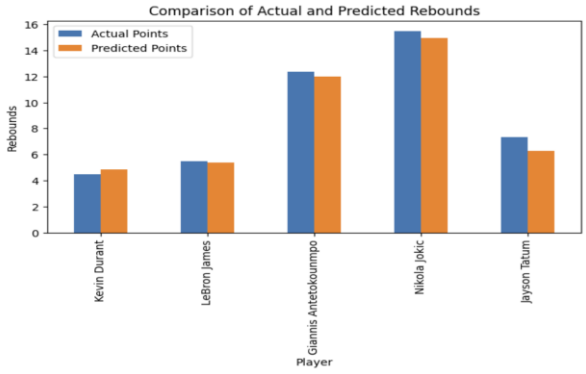

Figure 4 – Bar graph of players actual vs predicted rebounds performance generated with a random forest

IV. Experiments

With our model the number of estimators, or number of trees generated, is variable. As we want our model to be as accurate as possible, this value needs to be optimized. Using squared error a loss value was generated for each stat prediction. Visualizing this loss in figure 5 it can be seen that the loss decreases as the number of estimators is increased. Between 0-200, there is a sharp decrease in loss which eases up and becomes almost stagnant after 400 estimators. The conclusion is that 300 estimators is a good value as it keeps loss low while not unnecessarily increasing compassion time.

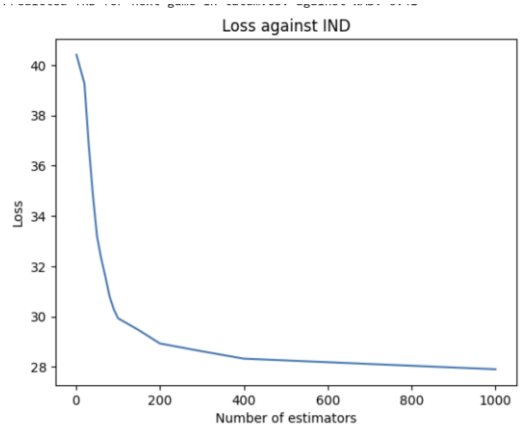

Figure 5 – Loss vs number of estimators, simulated against IND

V. Results

Testing the model with only each team's five starters produced results with least inaccurate predictions of the final, with teams sweeping each other and significant point differences between competitors. This testing also indicated that since the Celtics have the best starting five and the best on-paper stats, the flow of the playoffs into the finals would likely conclude with the Celtics nearly flawlessly sweeping to a championship

Figure 6 – Predicted NBA playoffs bracket

title.

Figure 8 – Predicted NBA playoffs bracket

Figure 7 – Points scored in simulated finals game

The final verdict of the random forest model is that the most accurate representation of playoff and finals games is displayed when the number of estimators is set to 300, and one bench player from each of the eight teams is incorporated. This method provided real game scenarios such as overtime, a one-point difference in game score, and accurate sweeps based on statistics from both competing teams.

IV. Conclusion

Our model produced accurate results in points, assists and rebounds against specified opponents. This was then used to predict a NBA Finals championship. Which was predicted to be the Dallas Mavericks. As the playoffs happen this year our prediction will be put to the test. This model and approach can be used for any team sport with significant statistics. With more statistics the more accurate the model will be.

IV. Who Did What

Mukhtar, Abstract, Intro on report, data acquisition, built linear regression model, worked on presentation slides (approach, issues, conclusions)

Kevin, worked on conclusion, experiment, and parts of method and formatted this paper. Worked on presentation slides (Model, Loss function). Wrote parts of random forest model. Coded game prediction function and loss visualization.

Rodney, Early development in code including data filtering, data acquisition, data gathering and reading in data to use in code, wrote results in paper and in presentation slides (results)

Ryan, worked on presentation slides (Purpose, dataset example, evaluation), data acquisition, wrote Preliminary Results in report, help code random forest model and visualizing data with bar graphs.

Everyone in our group did their fair share of work. We all worked extremely well together and were constantly meeting together to get his project done. We are all proud of what we have completed.

References

[1] G. Cheng, Z. Zhang, M. Kyebambe, and N. Kimbugwe, "Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle," Entropy, vol. 18, no. 12, p. 450, Dec. 2016, doi: https://doi.org/10.3390/e18120450.

[2] Basketball Reference, "Basketball Statistics and History | Basketball-Reference.com," Basketball-Reference.com. https://www.basketball-reference.com/

[3] Eric Scot Jones and Rhonda C. Magel, "Predicting Outcomes of NBA Basketball Games," Journal of Advance Research in Business Management and Accounting (ISSN: 2456-3544), vol. 2, no. 5, pp. 01-13, May 2016, doi: https://doi.org/10.53555/nnbma.v2i5.99.

[4] N. Donges, "Random Forest: a Complete Guide for Machine Learning," *Built in*, Jul. 22, 2021. https://builtin.com/data-science/random-forest-algorithm