CS 584 Project Proposal

Using Machine Learning to Compare Win-Loss Prediction Models for NBA Games

Zain Farhat A20468542
Luke McDaneld A20477151
Elliot Tan A20394064

Abstract
The world of professional sports is highly competitive in nature, and teams are always looking to gain an edge over the competition. Our goal is to compare different predictive strategies and models to determine what factors have the highest impact on the win-loss outcome of an NBA game. We aim to compare not only a number of machine learning methods and models to generate accurate predictions, but also to investigate what predictive factors most impact the accuracy of those models.

The problem to be solved
Our goal is to predict the outcome of future NBA games based on historical data. We intend to use both player and team data to create a variety of predictive models. The end goal of this project is to compare and contrast different models to select the most accurate method for predicting future outcomes.

Many classifiers and prediction models already exist for this problem. Media outlets like ESPN and fivethirtyeight use their own predictive methodologies (Real Plus Minus and RAPTOR respectively) with varying degrees of success. A brief survey of the topic shows that the best performing models achieve near or below 70% accuracy in predicting the correct outcome of a given game [7]. Our goal is to implement a variety of models to try and replicate or surpass these results.

Because we acknowledge that many of these models already exist, our aim is to take a deeper dive into the accuracy of the models and analyze how the selection of predictors and models can improve or degrade the prediction accuracy. The end result of our project is to compile a comparative study of model performance as well as how different statistical approaches influence the accuracy of the models.

Our approach to solve the problem

We will use historical NBA game data gathered from a variety of sources including but not limited to: Basketball-Reference historical game data, fivethirtyeight's RAPTOR individual player data repository, and ESPN's Real Plus Minus player evaluation metric. We will collect historial

data from the previous 5 NBA seasons as we would like to test our model in the later stages against upcoming games. The initial approach will be to implement common classification algorithms like linear and logistic regression, Naive Bayes Classification, and random forests.

Once we have established a suite of models to perform our predictions with, we will use different subsections of predictors to study how predictor selection affects the various models. Some of the aspects we will compare include weighting individual versus team data, counting statistics (points, blocks, rebounds) versus other existing advanced metrics (PER, WAR, VORP), and offensive versus defensive performance.

After deciding on a final set of models to evaluate, we will put our model to the test against the current NBA season. The NBA season begins on October 19, so we will have opportunities to see how the models perform in real time.

Preliminary Project Milestones:

- Collect and compare data from multiple historical sources
- Prepare data to be used to modeling procedures
- Implement different classification techniques to compare and contrast
- Test different feature sets to determine the ideal features for our models
- Perform Ablation Studies and Hyperparameter tuning
- Draft report analyzing results

References

[1] Basketball-Reference (2021) https://www.basketball-reference.com/ Accessed September 2021

[2] Hangtian Jia, Chunxu Ren, Yujing Hu, Yingfeng Chen, Tangjie Lv, Changjie Fan, Hongyao Tang, and Jianye Hao. 2020. Mastering Basketball With Deep Reinforcement Learning: An Integrated Curriculum Training Approach. In Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1872–1874.

[3] Jasmin A. Caliwag, Maria Christina R. Aragon, Reynaldo E. Castillo, and Ellizer Mikko S. Colantes. 2018. Predicting Basketball Results Using Cascading Algorithm. In Proceedings of the 2018 International Conference on Information Science and System (ICISS '18). Association for Computing Machinery, New York, NY, USA, 64–68. DOI:https://doi.org/10.1145/3209914.3209921

[4] Mahboubeh Ahmadalinezhad, Masoud Makrehchi, and Neil Seward. 2019. Basketball lineup performance prediction using network analysis. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19).

Association for Computing Machinery, New York, NY, USA, 519–524. DOI:https://doi.org/10.1145/3341161.3342932

[5] NBA-RAPTOR (2021) https://github.com/fivethirtyeight/data/tree/master/nba-raptor Accessed September 2021

[6] Samuel Li. 2020. Revisiting the Correlation of Basketball Stats and Match Outcome Prediction. In Proceedings of the 2020 12th International Conference on Machine Learning and Computing (ICMLC 2020). Association for Computing Machinery, New York, NY, USA, 63–67. DOI:https://doi.org/10.1145/3383972.3383980

[7] Weiner, Josh (2021) Predicting the outcome of NBA games with Machine Learning https://towardsdatascience.com/predicting-the-outcome-of-nba-games-with-machine-learning-a810bb768f20 Accessed September 2021