Mid-Quarter Progress Report:
Predicting the 2024 NBA Champion With Machine Learning

Zareb Islam | Aadhil Mubarak Syed | Jaynor Singson | Dylan Tran
ECS 171: Machine Learning | Professor Setareh Rafatirad
Department of Computer Science | University of California, Davis

---

Introduction - Jaynor

Our project aims to predict the 2024 NBA Champion using sophisticated machine learning techniques to analyze comprehensive team performance statistics. As the regular NBA season showcases increasing competitiveness and unpredictability, traditional predictive methods have proven inadequate, prompting our exploration of machine learning as a viable alternative. At this mid-quarter progress check, we have successfully completed the data extraction and pre-processing phases and are currently engaging in exploratory data analysis to identify the most influential factors for our model.

Data Extraction (Aadhil)

We sourced our data from Basketball Reference (https://www.basketball-reference.com/), a website containing "statistics, scores, and history for the NBA, ABA, WNBA, and top European competition." Initially, we considered employing web scraping techniques to collect the data; however, after encountering unexpected complexities with this approach, we opted to manually extract the data by copying CSV data directly from the website. We then imported these CSV files as strings into our Jupyter Notebook environment. Subsequently, we developed a Python script to transform these string representations of the data into pandas dataframes, facilitating the subsequent stages of data preprocessing in preparation for our data analysis.

Data Preprocessing (Aadhil)

Before initiating data pre-processing, we saved the original CSV files in a "raw_data" folder for potential future use. During pre-processing, we eliminated columns that were irrelevant or could cause collinearity in our predictive model, such as rank, minutes played, 3P%, FT%, and total rebounds. To differentiate team data from opponent data, we prefixed opponent metrics with an "O" (e.g., "PPG" became "OPPG"). We merged the datasets using the team name as the key and included a year column to distinguish team entries across seasons. We substituted the Win and Loss columns with a win percentage column to further reduce collinearity. After organizing the data by team name for consistency, we integrated a manually coded target values dataset indicating championship outcomes. Finally, we exported the comprehensive, pre-processed dataset into the "data" folder, generating individual CSV files for each season's data. This process can be found in the "Data_Extraction.ipynb" notebook.

Exploratory Data Analysis (EDA) - Dylan and Zareb

In our exploratory data analysis (EDA), we focused on comparing the statistical profiles of NBA championship-winning teams to league averages across various seasons through a series of 20 pre-determined questions. This comprehensive examination allowed us to identify key trends influencing a team's likelihood of winning a championship. Notably, we observed significant shifts in metrics such as an increase in pace and three-point attempt rates, alongside a decrease in free throw rates over the past two decades, reflecting the dynamic nature of successful strategies in the league. Our analysis also highlighted strong correlations between championship teams and league averages in metrics like net rating, points per game, and efficient field goal percentage, underscoring common attributes of successful teams. Conversely, we found weaker correlations in metrics such as turnover percentage and blocks per game, suggesting that while these factors can influence championship outcomes, they are not consistently critical for success. These findings not only enrich our understanding of the factors that contribute to NBA championships but also lay a foundational framework for our predictive modeling, emphasizing the importance of distinguishing between strong and subtle influences on championship outcomes.

Future Timeline - Aadhil

- [ ] May 19: Model Training, Evaluation, Deployment
- [ ] May 26: HTML Interface (Model Demonstration)
- [ ] June 2: Project Report Rough Draft
- [ ] June 6: Project Report Final Draft
