Matthew Harris
Matthew Maroun
Ronit Patel
Byungkwon Moon
CS 532: Systems for Data Science
28 March 2023

Final Project: Proposal

The aim of our proposed project is to make various predictions on the season results and player contracts using the statistics and salaries of National Basketball Association (NBA) players' salaries. The main tools of our project would include:

- PySpark to manipulate the data
- TensorFlow and/or PyTorch to make machine learning (ML)-based predictions,
- Matplotlib for visualization of data
- Google CoLab for working on the document of the group most easily
- GitHub, on which we will push all code to produce all deliverables and visualizations

Data Sets

- NBA Players Statistics: https://www.kaggle.com/datasets/iabdulw/nba-player-performance-stats
- NBA Player Contracts: https://www.kaggle.com/datasets/jarosawjaworski/current-nba-players-contracts-history

The main questions we wish to analyze, at the moment, consist of:

- How does age, position, number of teams played for in a season affect production? (By "production," we specifically mean a quantifiable increase in the player's statistics that are constructive to a team's winning of games, such as Field Goal percentage (FG%), total points, total assists, etc.), while observing a quantifiable decrease in that player's stats that are destructive in the aim to win, such as total fouls, turnovers, etc.
- Does the 3-pointer still have the most effectiveness when it comes to winning games? (Qualitatively rephrasing, is the game still small-ball-dominant? Or is it true that basketball is, has been, and always will be the big man's game?)
- Evaluating total spend of a team vs. team's regular-season record - i.e. did the owners/GMs do their jobs effectively?

Questions will be answered primarily through visualization of data in Matplotlib. As previously mentioned, Spark will be used to cut and manipulate the data for visualizations. These questions will be used to aid in the testable predictions we will use to evaluate the performance of an ML algorithm to handle the data. The predictions we wish to use for such an evaluation include:

- Are players' contracts really worth it? That is, based on the "production" of the average NBA player (see above for details on what is meant production), what number of players are being paid more than their net season contributions suggest they should be paid, and what number of players are producing more than their pay would suggest? (Keep in mind that contracts are often evaluated on historical evaluations of the players' statistics.)
- Predict who will make the playoffs (We will know the playoff spots in a few weeks so we can see how good our predictions are based off the data we have)
- Predict the final standings for each team after the regular season. This can be based on each players' performance, number of games played, player efficiency, etc.

The goals we wish to have deliverable by this project's Milestone include:

- All data manipulation as handled by Pyspark, including proper respective filtering for ML predictions and for visualization
    - Use this data to later create tests(Final Project Goal) for when the NBA regular season ends to compare how our predictions compared to what actually happened in the NBA regular season
- Various histograms of relevant statistics in the "productive" and "unproductive" categories as binned by the age of the player, then again by the positions of the players, then their ages, etc. Possibly aggregate histograms of multiple statistics binned by age, positions, etc, too.
- Plots of number of team wins vs. total number of 3-point shots made and by 3-pt. FG% and effective FG%, then comparing wins to overall effective FG% and 2-pt. FG% for a basis of comparison between the effectiveness of the 3-pt. FG over the 2-point shot
- Generating a line chart of team wins vs. total team salary, then possibly binning this data into a histogram, to determine any deterministic relationship between team salary and team wins

Final Project Goals

- Gather data/information on the final team records and final team standings when the NBA regular season ends
    - Used for writing tests to compare predictions to actual results
- Write tests to compare our predictions to how the teams and players performed in the regular season (The NBA regular season ends 4/9/23)
    - Tests are built by ranking each team from 1-30 in the specific category we are testing. Then we compare the top teams and see if they made the

playoffs. If they do we pass the test and if they don't we fail. And then we can repeat the tests for some of the bottom teams in the ranking for each category

-
  - Compare team standings to our predicted standings
  - Compare playoff teams to our predicted playoff teams
  - Compare teams with the best 3-point efficiency to their actual team records
  - Identify which player statistics were significantly correlated with team records
  - Compare total cost of team to final team records and which teams got the most value for what they spent on the team
  - Etc
- Experimental Results
  - Write the analysis of the questions listed above based on the results of the data queries
    - Explain if what we predicted actually happened in the NBA
  - Include explanations for possible improvements to get better predictions
    - If predictor doesn’t display accurate results, explain possible oversights and suggest improvements
- Explain design description of the program
  - Descriptions of methods or classes
  - Identify parts that were difficult to accomplish
- Push all code, deliverables, and visualizations to Github

Possible augmentations and improvements on the project include, but are not limited to, the inclusion of data from other professional sports leagues such as the National Football League (NFL) and Major League Baseball (MLB).