NBA Western Conference 2023-24 Season Playoffs Forecast

Ruiqi Zhang, Zichen Gong, Ziang Li

Problem Description

Forecasting the NBA Western Conference 2023-24 Season Playoffs is a multifaceted challenge with extensive implications. The unpredictability inherent in basketball, especially in a competitive cluster like the Western Conference, makes playoff predictions both significant and highly complex.

Why This Problem is Significant:

- **Financial Stakes**: Accurate playoff forecasts are crucial for teams' economic strategies, impacting player trades, marketing efforts, merchandise sales, and sponsorships. The financial health and strategic planning of teams hinge on these predictions.
- **Strategic Planning**: Teams plan their approach based on potential future adversaries and positioning in the playoffs. Reliable forecasts allow for better strategic preparation, risk assessment, and resource allocation throughout the season.

Implications of the Problem:

- **Advancement in Predictive Analytics**: This forecasting problem underscores the need for enhanced predictive models using machine learning, necessitating advancements in real-time data analysis, and adaptive decision-making tools.
- **Broader Real-World Application**: Solving this issue highlights the broader applicability of sports analytics in understanding complex, unpredictable scenarios. It

Introduction/Goals:

This project aims to forecast the outcomes of the NBA Western Conference playoffs for the 2023-24 season. This is a significant challenge in sports analytics, given the unpredictable nature of basketball, especially in a highly competitive environment like the NBA Western Conference. The predictions hold value not only for sports analytics but also for the strategic, financial, and marketing decisions of the teams. The goal of this project is to create a predictive model capable of accurately forecasting the outcomes of the NBA Western Conference playoffs. This involves analyzing extensive datasets, including individual player statistics and overall team performance.

Database management:

Basic architecture

playoffs:

- primary key Team

Tables with suffix "_rank":

- composite primary key (Season, Team)
- foreign key (Team) references playoffs(Team)

Tables representing each team derived from the full table merged_:

- composite primary key (Team, New_Rk)
- foreign key (Team) references their corresponding "_rank" table

Key features

- Custom Scoring System: A bespoke scoring system evaluates teams based on individual player statistics and team-level historical performance, highlighting key factors in playoff success.
- Database Management and Data Cleaning: The project emphasizes accuracy and reliability in data through meticulous data cleaning and management, standardizing measurements and resolving data inconsistencies.
- Web-Based Interface: A user-friendly web interface is included, allowing interactive exploration of predictions and data. This enhances accessibility and user engagement.

Description of data

This dataset offers a detailed snapshot of player statistics and team performance in the NBA, making it an invaluable resource for predictive modeling, player performance analysis, and team strategy formulation in the realm of basketball analytics.

Team: Indicates the NBA team to which each player belongs.

Player: The name of each individual player.

Position: The playing position of each player within their respective team.

Weight: The weight of each player, an important physical attribute in player performance analysis.

Height: The height of each player, another crucial physical characteristic in basketball.

Rk (Rank): The initial rank of each player within their team, based on performance metrics.

Age: The age of each player, offering insights into experience and career stage.

MP (Average Minutes Per Game): Represents the average amount of game time each player secures, indicating their role and significance in the team.

eFG% (Effective Field Goal Percentage): A measure of shooting performance that accounts for the value of three-point field goals.

TRB (Total Rebounds): The total number of rebounds secured by each player, highlighting their defensive contributions.

AST (Assists): The number of assists each player has, reflecting their playmaking ability.

PTS (Points Per Game): The average number of points scored by each player per game, a key performance indicator in basketball.

Coach-Playoff: The number of times each team's current coach has led the team to the NBA playoffs, providing a measure of coaching experience and success.

New_Rk (New Rank): A revised ranking of players within each team. This reordering addresses the gaps created in the original ranks due to the exclusion of certain players during data cleaning.

For 15 tables with "_rank" suffix:

Season: Represents the NBA season year.

Team: Indicates the NBA team.

GP (Games Played): The total number of games played by the team in the season.

W (Wins): The total number of games won by the team, a direct indicator of the team's success in the season.

L (Losses): The total number of games lost by the team, complementing the wins to provide a full picture of the team's performance.

WIN% (Winning Percentage): The percentage of games won by the team, offering a normalized measure of success irrespective of the number of games played.

PTS (Points): The total points scored by the team during the season.

PTS Rank: The rank of the team in terms of points scored compared to other teams, indicating their relative offensive capability within the league.

Division Rank: The team's rank within their division, providing a measure of their performance relative to their immediate competitive group.

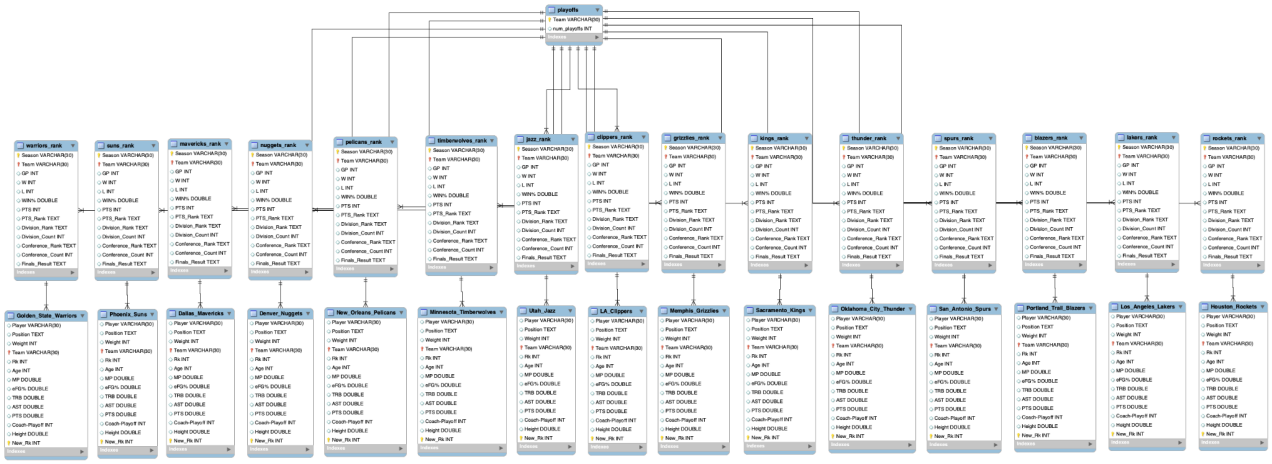
Division Count: The number of teams in the team's division, offering context to the Division Rank.

Conference Rank: The team's rank within their conference, indicating their standing in a larger competitive grouping.

Conference Count: The total number of teams in the team's conference, providing perspective to the Conference Rank.

Finals Result: The team's performance outcome in the NBA Playoffs, if applicable, showcasing their success in the ultimate stage of the competition.

ER Diagram



Implementation

Web Scraping & Table Creation and Merge:

We utilized the BeautifulSoup Python package to scrape information about players from NBA Western Conference teams, as well as data on how many times each team's current main coach has led their team into the NBA playoffs. This information was stored in the merged_2324_with_coach table. After conducting basic data cleaning, this extensive table was renamed to merged_ and exported as a CSV file. Once this comprehensive table was imported into MySQL, we manually created 15 separate tables based on team names for our database. Each table was named after the complete name of each team, such as Golden_State_Warriors.

Additionally, another dataframe was scraped from the official NBA website, specifically from the seasons information panel. This webpage provided access to each team's historical performance across different NBA seasons. The table compiled from this source includes data on aspects such as division rank, conference rank, and total games played. In a manner similar to the player information, we created 15 corresponding tables to incorporate this historical data into our database system.

As we only keep the teams who have a conference ranking that is at least 8, which means records that teams have entered the NBA playoffs in history, we can count the number of playoffs in total for each team in history simply by counting the number of rows for dataframes with "rank" suffix. After collecting this information, we manually create the _playoffs_ dataframe with team names and their number of playoffs in history.

Data cleaning:

Players who have a MP value smaller than 10 are dropped, since these players usually lack significant information and can leverage the prediction model. In the 'Season Summary' tables, we maintain a focus on playoff-relevant information. Consequently, entries pertaining to teams with a conference rank exceeding 8th are excluded, as they do not contribute to the playoff narrative. After that, all entries in the 'Season Summary' tables have been reorganized. This new sorting prioritizes chronological order, arranging the entries based on their respective season years to facilitate a clearer historical perspective.

Height Conversion:

In the dataframe merged_2324_with_coach, player heights were converted from a format in feet and inches (e.g., "6-6") to centimeters. This conversion was necessary for standardizing the height data across the dataset, and numeric values are easier for implementing prediction models.

Column Renaming and Modification:

Spaces in column names were replaced with underscores in various dataframes, such as those in the dataframe blazers_rank, to facilitate their importation into MySQL. Additionally, accents from player names in the merged_2324_with_coach dataframe were removed to ensure compatibility and prevent issues during MySQL import.

Removing Specific Data:

Specific players were dropped from the merged_ table to address the issue of duplicate entries. The original data source, obtained through web scraping, treated duplicate players as distinct entries, which affected both the ranking of players and the establishment of primary keys in subsequent MySQL operations.

Rank Reassignment:

In the merged_ dataframe, the Rk (rank) values within each team were reassigned to be consecutive. This reassignment was necessary because players averaging less than 10 minutes per game were excluded from our analysis, and the drop of duplicate players led to non-consecutive rankings as well. In cases of tied rank values, the 'MP' (Minutes Played) value was used as a deciding factor: players with higher average minutes played per game were assigned a higher (numerically lower) new rank (New_Rk). This step was crucial for setting up Rk as part of the composite primary keys in the database.

Entity resolution:

As we connected tables with the suffix 'rank', which indicates the past playoff rankings of each team, with player information tables, we noticed that some teams have been known by different names historically. To address this, we used Python to replace former team names with the current name, as listed in the first row of each table. This standardization was crucial for ensuring consistency in team names across our datasets. By doing so, we were able to establish reliable relationships between these tables and others containing player information, using 'Team' names as foreign keys.

Furthermore, we encountered a specific inconsistency with team naming conventions. Unlike the 'Los Angeles Lakers', where all dataframes consistently used the full name 'Los Angeles' as part of the team name, the name of the 'Clippers' team appeared inconsistently as either 'LA Clippers' or 'Los Angeles Clippers'. To reconcile this discrepancy, we standardized all instances of 'Los Angeles Clippers' to 'LA Clippers', ensuring uniformity in team naming across our data.

Technical Challenges Encountered:

- Encountered problems with MySQL Workbench, including crashes and issues viewing foreign key constraints. Foreign key constraints not appearing in the GUI.
- BLOB/TEXT column used in key specification without a key length.
- Failed to add foreign key constraint due to missing index in the referenced table.
- Explored MySQL Workbench log files for diagnosing issues like missing resource files and OS detection warnings.
- In the construction of the website, a uniform wrapper was utilized across all web pages, with background details embedded directly within this wrapper. This approach, while ensuring design consistency, presents a significant challenge in terms of applying unique background images for each individual web page. The integration of the background information into the wrapper restricts the flexibility needed for distinct visual customization on a per-page basis.
- Another technical challenge arises in the context of CSS conflicts and the cascading nature of stylesheets when integrating a third-party table template into an existing webpage. The CSS files associated with the table template are observed to override certain styles defined in the original webpage's CSS. This leads to visual and functional discrepancies in the website's rendering. The issue is rooted in the cascading and specificity rules of CSS, where styles from the table template take precedence over the established styles of the webpage, negatively impacting the website's intended aesthetic and functional harmony.

Prediction

The scoring system implemented for ranking NBA teams is based on a combination of individual player statistics and team-level historical performance. Each player is scored across various attributes such as weight, height, age, average minutes per game, effective field goal percentage, total rebounds, assists, and total points per game. These scores are assigned by ranking the players in each attribute, with the highest value receiving the highest score, and then scaling these ranks into a score between 0 and 1. For the age attribute, the scoring is reversed, with younger players receiving higher scores. For each team, the scores of the top five players, determined by a new rank metric, are aggregated to form a player-based total score. This score is then augmented by two additional team-level metrics: the historical number of playoff appearances and the coach's playoff experience. These two metrics are directly added to the team's score, thus combining individual player performance with team and coaching experience. Teams are then ranked based on this composite score, with higher scores indicating stronger overall team performance and potential for success. This approach provides a multi-faceted evaluation of team strength, taking into account both current player statistics and historical team achievements.

In the revised table, as we are aware that players' age, height and weight might not be as determinant as their actual performance during games, while the 'Age', 'Weight' and 'Height' attributes retain a full mark of 1, indicating a standard scale, other player attributes such as 'Average Minutes Per Game (MP)', 'Effective Field Goal Percentage (eFG)', 'Total Rebounds (TRB)', 'Assists (AST)', and 'Total Points Per Game (PTS)' are now assigned a full mark of 2. This change effectively doubles the impact of these attributes on a player's total score, emphasizing their greater significance in determining player effectiveness and team strength.

Future Thoughts and Potential Improvements

The current scoring system, while effective in its basic form, acknowledges certain limitations that can be addressed with more sophisticated techniques in future iterations. For instance, leveraging advanced machine learning tools like Lasso regression could significantly refine the system's predictive accuracy. This enhancement would be particularly effective once the results of the current season are available, allowing for fine-tuning and calibration based on actual outcomes.

Additionally, incorporating information from the upcoming season could further enhance the model's precision. Implementing a propensity score matching system is one such strategy that could be explored to improve predictive accuracy. This approach would allow for a more nuanced analysis by matching similar entities and understanding the impact of various factors on team performance.

The decision to not rely heavily on time series analysis for this model stems from the rapidly changing nature of the NBA. Historical data from several years ago may not be as relevant or predictive for the current season due to various factors. Players can undergo significant changes, such as injuries, retirement, or shifts in performance levels. Additionally, the introduction of new talents and changes in coaching staff can markedly alter a team's dynamics. In the NBA, each season can vary dramatically, making newer information more valuable and indicative of future performance than older data. Therefore, the focus is on leveraging the most recent and relevant information to inform our predictions.

In the context of the website design, further optimization for various mobile devices and screen sizes could enhance user experience. This includes adjusting layouts, image sizes, and interactive elements for better usability on smaller screens. In addition, ensuring the website is accessible to all users, including those with disabilities, is important. This can involve adding alt text to images, ensuring proper contrast ratios, and making navigation keyboard-friendly.

Reference

Real data_source: (web scraping data source):

Basketball Statistics & History of every Team & NBA and WNBA players. Basketball. (n.d.). Retrieved from: https://www.basketball-reference.com/

The official site of the NBA for the latest NBA Scores, Stats & News. NBA.com. (n.d.). Retrieved from: https://www.nba.com/

Images_source:

Ganguli, T. (2023, June 13). *Denver Nuggets beat Miami Heat for first N.B.A. championship*. The New York Times. Retrieved from: https://www.nytimes.com/2023/06/12/sports/basketball/denver-nuggets-miami-heat-nba-finals.html

*NBA Teams & Rosters|NBA.com. (n.d.-a).*NBA.com. Retrieved from: https://www.nba.com/teams

Other sources used:

Keane, S. (2022, April 15). *Reviewing the NBA's playoff slogans*. Golden State Of Mind. Retrieved from: https://www.goldenstateofmind.com/2022/4/15/23027279/nba-playoff-slogans

Massively by HTML5 UP. HTML5 UP. (n.d.). Retrieved from: https://html5up.net/massively

Responsive table HTML and CSS only - codepen. (n.d.). Retrieved from: https://codepen.io/florantara/pen/dROvdb

24 Motivational Basketball Quotes To Build Confidence. Retrieved from: https://www.basketballmindsettraining.com/blog/24-motivational-basketball-quotes

Index of files submitted:

Folders:

- **cleaned_dataframes**:
  Contains all CSV files ready for direct import into MySQL. These include:
  - ☑ CSV files with the "rank" suffix after data cleaning.
  - ☑ The complete table merged_ including all player information.
  - ☑ The manually created table playoffs.

- **ConferenceRank**:
  Contains all rank tables that have been scraped and undergone basic data cleaning. Not yet ready for import into MySQL.

- **webpage_code**:
  Contains code for all web pages.
  - ☑ 'index.html' serves as the primary homepage.
  - ☑ 'reference.html' is designated for references.
  - ☑ 'prediction.html' hosts the outcomes from the prediction model.
  - ☑ Specific web pages dedicated to each team are labeled following the convention 'team_name.html', such as 'lakers.html'.
  - ☑ Folder 'images' stores all the images utilized on the site.
  - ☑ Folder 'assets' contains various auxiliary functions and resources.

PDF files:

- **CPSC_437_Project_Fall_2023**.pdf—Final project instructions.
- **ER_pdf**.pdf—Directly generated ER diagram output from MySQL Workbench.

ipynb files:

- **NBA_website_2 (1)**.ipynb: web_scraped data
- **NBAstat (1)**.ipynb: web_scraped data
- **CPSC537 final project**.ipynb:
  Performs dataframes web scraping for all player information tables. 15 tables are combined into combined_nba_data and exported. No additional data cleaning addressed.
- **merged_df**.ipynb:
  Imports merged_2324_with_coach and performs a series of data cleanings to prepare merged_.csv for MySQL and prediction model import. Also cleans all rank files, and creates the playoffs dataframe, exporting it as playoffs.csv.

CSV files:

- **Files named in formats such as pelicans_rank, warriors_rank in the cleaned_dataframes folder:**
  Rank information tables that are part of the database.
- **merged_.csv in cleaned_dataframes:**
  Complete player information. Later split into smaller tables representing individual teams for database storage.
- **playoffs**.csv:
  Manually created table storing the history of playoff entries and corresponding team names.
- **combined_nba_data**.csv:
  Player information table post-web scraping without data cleaning, lacking main coach information.
- **merged_2324_with_coach**.csv:
  Includes some data cleaning (e.g., removing players with less than 10 average minutes per game) and adds a column for coach information. Not yet ready for MySQL import and prediction modeling.
- **rank_equal**.csv:
  Teams' ranking information with all features weighted the same is stored in this file.
- **rank_adjusted**.csv:
  Team's ranking information with other features weighted more heavily than age, height and weight is stored in this file.

SQL file:

- **nba**.sql:
  Contains all SQL codes used for creating and separating tables, changing variable types, setting up primary keys and foreign keys, and creating the database.

RMD file:

- **prediction**.rmd:
  An R Markdown file that processes the creation of the prediction model.