Data Processes Project Report

Carlos Sanchez

Alvaro Arranz

Angel González

Daniel Saiz

12/16/2019

### Introduction and Related Work

On the project proposal we formulated 3 different questions. We decide to answer the second one about the gap between east and west conferences.

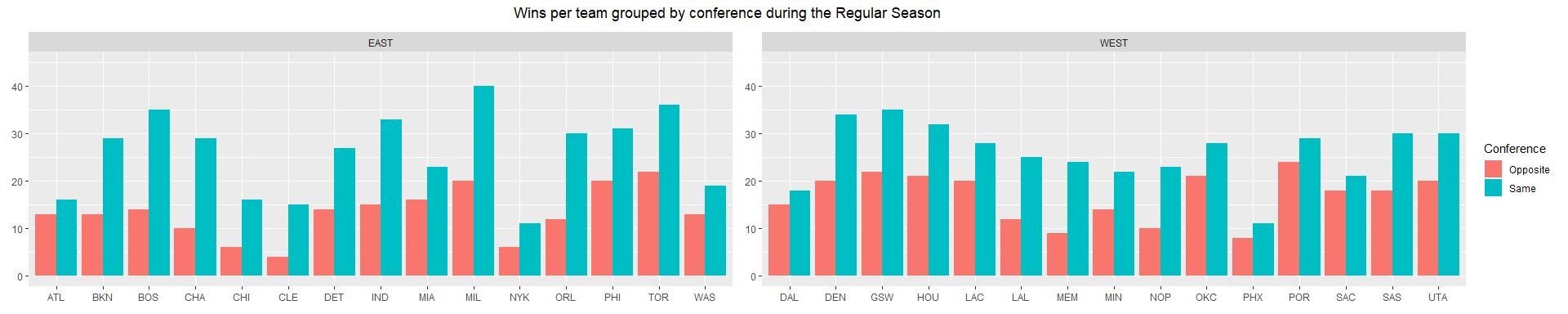
One of our member followed for years the NBA and during that period of time the Western conference teams ruled the league with Golden State Warriors on the head. Only the Cleveland Cavaliers and Miami Heat, thanks to LeBron James could fight against the best western team. During the last 8 seasons, LeBron James won 4 rings (3 with Miami Heats and 1 with Cleveland Cavaliers) and played 7 finals.

LeBron James is one of the best players in the NBA history, and he is able to lead a whole team to the victory almost by himself, but this is not the case with any other player of the league at this moment. LeBron moved to Los Angeles Lakers on the last season, and the project wasn’t complete, which is one of the reasons to select the season 2018/2019 to avoid an outlier on the eastern team where LeBron is playing.

To check if we were sure about our assumptions, we loaded the Match Statistics Dataset and performed several transformations to retrieve the data we wanted.

|  |  |
| --- | --- |
| Conference | Wins |
| EAST | 198 |
| WEST | 252 |

It seems like the gap between the West and the East is still there, but with this simple table we cannot be sure the data is not biased by low performing teams. To check this we plotted the wins against teams of the same conference and teams of the other, clustering the results by conferences. The script which generates the following chart can be found [here](https://github.com/AlmaProcesses/NBA_Project/blob/master/question_2.R)



We can see how West teams perform better than East ones on the cross matches. The eastern team with the most victories during the regular season was Toronto Raptors, which finally won the playoffs finals against Golden State Warriors. With this visualization we can assure the West conference was stronger than the East conference during the season 2018/2019.

Here you will find several relevant work about this question that we want to answer:

* [The Talent Difference Between the West and Eastern Conferences](https://thegruelingtruth.com/basketball/the-talent-difference-between-the-west-and-eastern-conferences/)
* [COMPETITIVE BALANCE IN THE NBA: Comparative analysis of Eastern and Western Conferences](http://pagines.uab.cat/appliedeconomics/sites/pagines.uab.cat.appliedeconomics/files/Brosed,%20M._paper.pdf)
* [NBA Predictions for the 2019-2020 Season](https://www.sbnation.com/nba/2019/10/22/20899122/nba-predictions-2019-2020-season-mvp-championship-nba-finals-lakers-clippers)

### Exploratory Data Analysis

#### Datasets

The dataset has been produced using the R package [nbastatR](http://asbcllc.com/nbastatR/), which parses information from the official [NBA API](https://stats.nba.com/) and different pages containing NBA historic data like [Hoopshype.com](https://hoopshype.com/). There are two different datasets:

##### Match Statistics Dataset

This dataset contains statistics about the NBA games during the 2018-19 season.

This dataset consists of one csv file. It’s named [regularseason1819\_games.csv](https://github.com/AlmaProcesses/NBA_Project/blob/master/data/regularseason1819_games.csv). It has 2460 observations and 47 features.

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| yearSeason | num | The year when the statistics were taken |
| slugSeason | num | The season of the match |
| slugLeague | chr | League name abbreviation |
| typeSeason | chr | Type of season (always Regular Season) |
| dateGame | date | Date the match was played |
| idGame | num | Id of the match on the NBA API |
| numberGameTeamSeason | num | Match order for the team |
| nameTeam | num | Name of the team |
| idTeam | num | Id of the team on the NBA API |
| isB2B | logi | The match was a back to back |
| isB2BFirst | logi | The match was the first back to back |
| isB2BSecond | logi | The match was the second back to back |
| locationGame | chr | Location where the match was played |
| slugMatchup | chr | Abbreviation of the match |
| slugTeam | chr | Abbreviation of the team |
| countDaysRestTeam | num | Days the team had for resting before the match |
| countDaysNextGameTeam | num | Days before the next team’s match |
| slugOpponent | chr | Abbreviation of the opponent team |
| slugTeamWinner | chr | Abbreviation of the winner team |
| slugTeamLoser | chr | Abbreviation of the defeated team |
| outcomeGame | chr | The abbreviation of the result (W|L) |
| isWin | logi | True if the team won the match |
| fgmTeam | num | Field goals made by the team |
| fgaTeam | num | Field goals attempted by the team |
| pctFGTeam | num | % of field goals made vs attempted |
| fg3mTeam | num | 3-pts field goals made by the team |
| fg3aTeam | num | 3-pts field goals attempted by the team |
| pctFG3Team | num | % of 3-pts field goals made vs attempted |
| pctFTTeam | num | % of free throws scored |
| hasVideo | logi |  |
| fg2mTeam | num | 2-pts field goals made by the team |
| fg2aTeam | num | 2-pts field goals attempted by the team |
| pctFG2Team | num | % of 2-pts field goals attempted by the team |
| minutesTeam | num | Minutes played by the team |
| ftmTeam | num | Free throws made by the team |
| ftaTeam | num | Free throws attempted by the team |
| orebTeam | num | Offensive rebounds of the team |
| drebTeam | num | Defensive rebounds of the team |
| trebTeam | num | Total rebounds of the team |
| astTeam | num | Number of assists of the team |
| stlTeam | num | Number of steals of the team |
| blkTeam | num | Number of blocks of the team |
| tovTeam | num | Number of turnovers of the team |
| pfTeam | num | Number of personal fouls of the team |
| ptsTeam | num | Points scored by the team |
| plusminusTeam | num | Difference of points scored with the other team |
| urlTeamSeasonLogo | chr | Url with the logo of the team |

##### Teams Dataset

This dataset contains information about the Teams in the NBA.

This dataset consists of one csv file. The first one, [nba\_teams.csv](https://github.com/AlmaProcesses/NBA_Project/blob/master/data/nba_teams.csv), is a list of the teams of the NBA and their information. It has 30 observations and 9 features.

|  |  |  |
| --- | --- | --- |
| Column | Type | Description |
| nameTeam | chr | Name of the team |
| idTeam | num | Id of the team |
| slugTeam | chr | The acronym of the nameTeam |
| teamName | chr | The colloquial name of the nameTeam |
| cityTeam | chr | Name of the City where the team belongs |
| teamNameFull | chr | The abbreviation of the nameTeam |
| idConference | num | Id of the team conference |
| idDivision | num | Id of the team division |
| urlThumbnailTeam | chr | Url with the logo of the team |

* Creates 5 well designed and formatted graphics (**15 points**, 3 each)
  + The visual uses the appropriate visual encodings based on the data type (**1 point**)
  + Written interpretation of graphic is provided (**1 point**)
  + Clear axis labels, titles, and legends are included, where appropriate (**1 point**)

### Methods (**30 points**)

The appropriate methods are employed to answer the question of interest, including: - **Strength of relationships**: Uses the appropriate technique to assess the strength of relationships amongst your variables of interest. You should include: - A formula describing how you believe your features (independent variables) are related to your outcome of interest (dependent variable) (**5 points**) - A defense of the variables included in your formula (**5 points**) - Creating the appropriate model based on your dataset (**5 points**)

* **Prediction**: You must also make predictions for your outcome of interest. In doing so, you must demonstrate a clear use of:
  + Splitting your data into testing/training data (**2 points**)
  + Applying cross validation to your model (**3 points**)
  + Appropriately handling any missing values (**2 points**)
  + Appropriately using categorical variables (**3 points**)
  + Using a grid search to find the best parameters for you model of interest (**2 points**)
  + Employing the algorithm of interest (**3 points**)

### Results (**20 points**)

You must provide a clear interpretation of your statistical and machine learning results, including at least **one visual or table** for each. - **Strengths of relationships**: For the features you included in your model, you must describe the strength (significance) and magnitude of the relationships. This can be presented in a table or chart, and pertinent observations should be described in the text. (**10 points**) - **Predictions**: How well were you able to predict values in the dataset? You should both report appropriate metrics based on the type of outcome you’re predicting (e.g., root mean squared error v.s. accuracy), as well as a high quality visual showing the strength of your model (**10 points**)

### Discussion and Future Work (**10 points**)

Based on *specific observations* from the results section, the report clearly provides: - An analysis of the real world implications of the results, at least one full paragraph (**5 points**) - Clear suggestion for directions of future research, at least one full paragraph (**5 points**)