Final Project Data Memo_Revised

Joseph Chang, Akul Bajaj, Tom Wei

2/3/2022

- An overview of your dataset

What does it include?

Our data set includes NBA players and their salaries, the team they play for, their ranking, and more statistics regarding the individual players.

Where and how will you be obtaining it? Include the link and source.

We will be obtaining data from kaggle and Basketball Reference.

https://www.kaggle.com/annettecatherinepaul/nba-salaries

https://www.basketball-reference.com/contracts/players.html

Added: https://www.kaggle.com/umutalpaydn/nba-20202021-season-player-stats?select=nba2021_advanced.csv

About how many observations? How many predictors? Observations are statistics of each NBA player, such as points per game. The predictor will be their salary in the next year.

Feedback:

NBA data should work for a final project.

You can use multiple datasets for the project as long as you can merge or join them together; https://hollyemblem.medium.com/joining-data-with-dplyr-in-r-874698eb8898 See there for more information on how to join datasets.

How many observations will you have? That's one of my main concerns. If there's only one row per player, how many rows is that?

Let's meet and talk about this this week, either after class or during my office hours. Then you can resubmit the memo based on the conversation and I'll regrade it.

Revised: There are 481 players in our dataset, which means there will be 481 observations. We will merge the two datasets -the salary dataset and the Nba 2020-2021 Season Player Stats dataset by players' names.

What types of variables will you be working with? Most of our variables will be numerical. Character and logical variables will also exist.

Is there any missing data? About how much? Do you have an idea for how to handle it?

In the basketball reference data, some salaries for the 2021-2022 season are missing. This could be because of injury or other external factors. We will handle this by disregarding the data.

- An overview of your research question(s)

What variable(s) are you interested in predicting? What question(s) are you interested in answering?

We want to know variables such as MPG, PPG, RPG, APG, SPG, BPG, PER, Age, (basketball statistics), etc. Given these variables, how much will a player earn?

Name your response/outcome variable(s) and briefly describe it/them.

The response variable is the salary we want to predict of nba players. How much salary will a player earn for a future season?

Will these questions be best answered with a classification or regression approach?

Regression approach will be the best because it deals with numbers and multiple predictors.

Which predictors do you think will be especially useful?

The most useful predictor is probably the rank of the player because it determines how valuable they are to the team.

- Your proposed project timeline and group work

How is your group dividing up the work?

We will all work together on the code and we will be meeting via zoom, or in person in the case that school continues in person to discuss tasks that need to be completed, and who is responsible for each task.

When do you plan on having your data set loaded, beginning your exploratory data analysis, etc?

We plan on having our data set loaded and beginning our exploratory data analysis 1/24/21.

Any questions or concerns

Can we have more than one dataset? And if we do, how will that affect our project?

Are there any problems or difficult aspects of the project you anticipate?

Some players are currently signed to contracts so their salaries for the next couple of years are already confirmed. Maybe we can use this to check the variability of our machine learning algorithm. Salary could also be affected by inflation or lockouts, so errors may be common.

Any specific questions you have for me/the instructional team?

What is the maximum number of datasets we can use, and how would that affect our project.