131-Final Project Data Memo

Ruoyu Li(9522913)

2022-10-02

An overview of your dataset

What does it include?

It includes a csv file, containing data about all drafted NBA players from 1989 to 2021. Data includes draft year, number of pick, team, games/seasons/time played, points/assists/rebound per game etc.

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

It's available on kaggle and the link is https://www.kaggle.com/datasets/mattop/nba-draft-basketball-player-data-19892021

About how many observations? How many predictors?

There are 1923 observations, and there are 19 possibly useful predictors.

What types of variables will you be working with?

mostly numeric variables

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

There are missing data about some players, I will say there's about 5% missing data. The reason might be a player is drafted but never played in NBA. If I can't make sure why the data is missing, it's probably okay to just delete those observations.

An overview of your research question(s)

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

I'm interested in predicting the variable win_shares_per_48_minutes, since it basically tells how much wins can a player contribute to the team. I want to answer questions like how would the number of overall pick of a player affect his performance in NBA.

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

win_shares_per_48_minutes, it is an estimate of the number of wins contributed by the player per 48 minutes (league average is approximately 0.100).

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

Regression, since the outcome would be numerical values.

Which predictors do you think will be especially useful?

Number of pick, or points/assists/rebound per game.

Is the goal of your model descriptive, predictive, inferential, or a combination? Explain.

I think it can be both predictive and inferential. I aim to predict the win shares by a player given his draft pick, but also I can investigate which of the player's statistics contributes the most to his win share.

Your proposed project timeline

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

I already have the csv file so I can upload it in week2 and start cleaning up the missing values.

Provide a general timeline for the rest of the quarter.

I can do general EDA in the week 3-4 I think. Then I will try to produce my predictive model from week 5 and on. One thing I'm not sure is whether I will learn the stuff I need to use for my project before I proceed to that part.

Any questions or concerns

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

I can't decide if this problem is too hard or too easy. Because based on what I have learned before, this problem might just be solved by a linear regression model. Or it could be too hard that even if I tried many ways, I still can't get a satisfactory predictive model.

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

Can you post more specific examples of R code for the future hws/projects since they're typically not given in lecture notes?