Project Proposal

CS396-4 Causal Inference

February 10, 2022

Instructions

This assignment is due on Thursday, Feb 10 at 11:59pm CST. Assignments will be accepted up to 72 hours late, but with a 14.3% (1/7th) penalty per day late. If your assignment is less than 24 hours late, we'll grade it and you'll receive 85.7% of those points; if it's less than 48 hours late, you'll receive 71.4% of those points. If it's more than 72 hours late, you'll receive zero points. This assignment is worth 5% of your final grade. If you are working in a group and are concerned that some members of your group are not contributing equally, please email me or let me know in an anonymous course survey.

Please upload your (group's) proposal to Canvas as a single PDF. You can use this assignment TeX to fill in your answers. If you are working in a group, please only submit one PDF for the entire group. While this template is rather long because it includes both questions and partial example answers, you may delete these instructions from the pdf you submit. If you do, we expect your proposal to be *at the most* three pages long.

1 Group members

Please list your group members.

Michael Diersen and Ben Caterine

2 Problem Statement

Describe the problem you're considering, why it is important, and who else cares about it.

We want to understand how much individual college basketball player statistics impact their NBA draft stock. This is important to basketball coaches, scouts, NBA front offices, and basketball analysts breaking down draft prospects. It is important for them to identify the primary factors that contribute to a player's draft stock and determine whether these factors actually have a causal effect on the player's professional performance.

3 Causal Questions or Hypotheses

Describe at least one causal question or hypothesis you would like to investigate.

(a) What are the treatment(s) and outcome(s)?

Treatments: Player performance across various stats (points, rebounds, assists, blocks, year in college, etc.)

Outcome: Draft position (round and pick) and whether a player is drafted

(b) Frame your question as a contrast of counterfactual random variables.

We are interested in the causal risk ratio $E[Y^{a=1}]/E[Y^{a=0}]$, or expected draft position had they scored more points, assists, etc. divided by the expected rate of their draft position had they scored less points, had less assists, etc.

(c) If you haven't been able to decide which causal question(s) you want to ask, please list multiple possible treatments and outcomes and possible arguments for or against each.

We are interested in using a variety of stats as treatments to determine which ones have the largest causal effect on draft position. We have also considered focusing primarily on recruiting rankings out of high school as a causal factor impacting professional performance, which would allow us to determine whether these recruiting rankings can truly predict a player's career.

4 Dataset(s)

What dataset(s) do you plan to use?

For the following questions, answer (a) through (c) for *each dataset* you might use. You only need to answer (d) through (f) for *one dataset* – preferrably the largest dataset or the one you plan to work with first.

(a) Describe the dataset's background: how was it collected, what does it contain?

The data was collected through watching the basektball. One of the data sets contains a list of all college basketball players from the year 2009 through 2021 as well as several columns worth of different individual player statistics. The other data set (on the same link) has the draft position of all college basketball players drafted from 2009 through 2021.

(b) What are the limitations of this dataset?

Certain obscure players stats such as rim shots and recruit rank have null values.

The draft dataset also contains international players who were drafted, which is not useful to our investigation because they did not play college basketball.

(c) Describe the format of the data. Can it be represented as a NxD matrix with N individuals and D features? If not, why?

The data is in a table format and can be represented as a NxD matrix with N being players and D being statistics.

The draft data is a similar matrix format, with N players and D features, those being related to the player's attributes and draft position.

(d) Load the data as a pandas dataframe (e.g. `pd.read_csv`) and provide a printout of at least ten rows.

	player_name	team	...	Unnamed: 64	Unnamed: 65
0	DeAndrae Ross	South Alabama	...	NaN	6.22026
1	Pooh Williams	Utah St.	...	NaN	3.94375
2	Jesus Verdejo	South Florida	...	NaN	10.92680
3	Mike Hornbuckle	Pepperdine	...	NaN	6.77427
4	Anthony Brown	Pacific	...	NaN	0.00000
...	...	...	...	...	...
61056	Trey Patterson	Villanova	...	Pure PG	0.00000
61057	Stavros Polatoglou	Northwestern St.	...	C	0.00000
61058	Sandy Ryan	Tulane	...	PF/C	0.00000
61059	Ty Larson	Texas Tech	...	PF/C	0.00000
61060	Jaden Jones	Rutgers	...	Pure PG	10.43920

(e) For at least six variables (columns) in your dataset:

i. Describe that variable: what is it measuring? Is it a discrete or continuous variable? Does it have any missing values? What is its mean and standard deviation?

ii. What are the possible[1] causal relationships between this variable and the other variables (in part e)?

- pts, measures the amount of points scored by a player per game they played. It is a continuous variable. It has no missing values. Mean: 5.77, SD: 4.95. A higher points value could be caused by a higher minutes value.

- ast, measures the amount of assists a player had per game played. It is a continuous variable. It has no missing values .Mean: 1.07, SD: 1.17. A higher assist value could be caused by a higher minutes value.

- blks, measures the amount of blocks a player has per game played. It is a continuous variable. It has no missing values. Mean: 1.87, SD: 5.70. A higher blocks could be cause by a higher min_per value.

- rbs, measures the number of rebounds per game of a player over the course of the season. It is a continuous variable. It has no missing values. Mean: 2.66, SD: 2.10. A higher rbs could be cause by more minutes.

- TS_per, measures the true shooting percentage of a player over the course of a season. It is a continuous variable. It has no missing values. Mean: 47.58, SD: 17.64. A higher TS_per could cause a higher pts value.

- Rec Rank, measures the average recruting rank of college basketball players coming out of high school. It is a continuous variable. It has no missing values. Mean: 53.56, SD: 27.56. A higher Rec Rank could cause more pts, ast, blks, and rbs as they are projected to be a better player.

(f) What is at least one variable that your dataset doesn't contain but might be a causal factor? How might such a variable complicate your causal question(s)?

None of the NBA combine data is included with stats such as bench press and vertical jump. These stats do play a role in players' evaluation pre-draft.

5 Expectations and Concerns

Write a few sentences about what you hope to learn during this project. Are there concepts from the class that you hope to explore with this particular dataset? Are there any challenges you expect to encounter while working on this project?

We hope to learn an overall application of causal inference topics on a real data set, gaining experience as a computer scientist working with and analyzing big data. One concept we hope to possibly explore more with this data set is how to handle missing data. We hope to eventually explore a predictive aspect where we can more accurately predict the order players get selected in the NBA Draft. One challenge that could arise is the fact that there are only about 45-50 college basketball players drafted each year, which could lead to a lack of data over the time span. Thankfully, the data has enough years that there are still over 700 rows in the draft data.

[1] If you don't have the domain knowledge to answer this question, that's okay. Just focus on at least a few variables where the way the data was measured makes it clear. For example, someone's age cannot be caused by any other variables.