Proposal: March Madness Predictor

DATA 450 Capstone

Joseph Bellani

2/5/23

1 Introduction

For my project I want to explore the world of predicting sports, specifically the NCAA Men's basketball Tournament. To do so my goal is to create an algorithm that can predict the chances that a team has to win against any given team. Using all sorts of metrics and data about tournaments past. I want to see if I can build a predicted model that will perform better than average human being, and would like to put it to the test in a month and a half, given all goes well.

2 Dataset

ESPN: ESPN Stats - Accessed 2/5/2023

ESPN: ESPN BPI - Accessed 2/5/2023

kenpom: kenpom ratings - Accessed 2/5/2023

- oBPI: The team's offensive Basket Power Index rating
- dBPI: The team's defensive Basket Power Index rating
- BPI: Summation of oBPI and dBPI
- wins: The team's wins in the season
- losses: The team's losses in the season
- SOS: The team's Strength of Schedule in the season
- ppg: The team's points scored per game
- papg: The team's points allowed per game
- to: The team's turnovers per game
- toForced: The team's turnovers forced per game
- Qwins: The team's quality wins in the season
- Qlosees: The team's quality losses in the season

- AdjEM: The team's Adjusted Efficiency Margin
- AdjO: The team's Adjusted Offensive Efficiency
- AdjD: The team's Adjusted Defensive Efficiency
- AdjT: The team's Adjusted Tempo
- Conf: The team's Conference]

3 Data Acquisition and Processing

The way I will be collecting the data will be through WebScraping from both ESPN and KenPom. After scraping is complete I will need to go through the data and see if anything must be cleaned before proceeding. It is likely that some variable names will need to be adjusted before use.

4 Research Questions and Methodology

1. Can a model based on data predict the March Madness tournament better than the average person? For this I will create a logistic regression model that will predict singular games which then will be combined to form a whole bracket. I will compare the results of the generated bracket to that of the ESPN "The People's Bracket" which is a culmination of every bracket submitted on ESPN, whichever bracket scores more points higher will be deemed the better bracket.

2. What statistics does a model consider to be the most important when it comes to a team's chances of winning, and after the results are shown did the model overestimate or underestimate certain statistics. For this question I will analyze the equation of the logistic regression taking notes of which variables are kept, which are used and the amount of weight the variables have. Once the results are done I can look at which variables the winning teams had and see if the model accurately projected their chances based on those variables.

3. For the last question I would like to take a look at trends of teams making the tournaments. This will be a more historic view as opposed to a predictive model. I can make graphs to lay out trends of teams such as a line graph of points per game to see if that has changed in the past 10 years or so. A pie graph of which conferences teams are coming from. And more charts like those to analyze what teams that get into the tournament look like.

5 Work plan

[Fill in the list below with a plan for what you will do each week. You should have around 7 hours worth of work each week. Writing work counts. Several tasks have already been filled in for you.]

Week 4 (2/6 - 2/12): [Just an example:

- Webscraping Data (4 hours)

Week 5 (2/13 - 2/19):

- Building model for Question 1 (4 hours)
- Analyze model equation for Question 2 (1 hour)

Week 6 (2/20 - 2/26):

- Test model for 2022 Tournament (2 hours)
- Create visualizations for Question 3 (4 hours)

Week 7 (2/27 - 3/5):

- Presentation prep and practice (4 hours)

Week 8 (3/6 - 3/12): *Presentations in class on Thurs 3/9.*

- Presentation peer review (1.5 hours)
- Insert Data for 2023 Tournament (1 hour)

Week 9 (3/20 - 3/26):

- Poster prep (4 hours)
- Compare Results for Question 2 (2 hours)

Week 10 (3/27 - 4/2): *Poster Draft 1 due Monday 3/27. Peer feedback due Thursday 3/30.*

- Peer feedback (2.5 hours)

- Poster revisions (2 hours)

Week 11 (4/3 - 4/9): *Poster Draft 2 due Monday 4/3. Final Poster due Saturday 4/8.*

- Poster revisions (1 hour).

Week 12 (4/10 - 4/16):

Week 13 (4/17 - 4/23): [All project work should be done by the end of this week. The remaining time will be used for writing up and presenting your results.]

- Draft blog post (5 hours).

Week 14 (4/24 - 4/30): *Blog post draft 1 due Monday 4/24. Peer feedback due Thursday 4/27. Blog post draft 2 due Sunday 4/30.*

- Peer feedback (2.5 hours)
- Blog post revisions (2 hours)
- [Do not schedule any other tasks for this week.]

Week 15 (5/1 - 5/7): *Final blog post due Tuesday 5/2.*

- Final presentation prep and practice.
- [Do not schedule any other tasks for this week.]

Final Exam Week (5/8): *Final Presentations during final exam slot, Monday May 9th 3:20-6:40pm.* [Do not schedule any other tasks for this week.]

6 References

I'm not sure if we are meant to repost the same references from earlier, the description said it would autogenerate a list, but it did not so I put the websites I will be scraping from down here as well.

ESPN: ESPN Stats - Accessed 2/5/2023

ESPN: ESPN BPI - Accessed 2/5/2023

kenpom: kenpom ratings - Accessed 2/5/2023