Datasheet for dataset used in 'The Success of College Basketball Teams can be predicted using several advanced team statistics'[*]

Thomas D'Onofrio

26 April 2022

Extract of the questions from Gebru et al. (2021)

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The datasets used and combined in the paper were created by the site Sports Reference. Sports Reference provides data for the purpose of helping fans answer their many questions about sports data and statistics. Their objective is to help fans "grow their appreciation, understanding, and love of the game" by providing them with the informative datasets and resources free of cost (Sports Reference 2022).
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The website and its many datasets were first founded by Sean Forman who was a PhD math student at the University of Iowa (Cannella 2002), and the basketball section in particular was created by Justin Kubatko. There is now a large team of workers who help create, update, and maintain the datasets on the website.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - The websites and datasets were self-funded by the creator.
4. *Any other comments?*
   - No.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - Every instance or observation in the dataset contains team statistics of an NCAA Divison 1 men's basketball school in a given season. Since multiple seasons were used in the dataset, schools may appear up to 5 times in the dataset, with different team statistic values depending on the specific season the observation belongs to.
2. *How many instances are there in total (of each type, if appropriate)?*
   - 1776
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - The dataset contains every instance of a school participating in a college basketball regular season anywhere from 2017-22. Since seasons before 2017 were not used, the dataset is a sample of all

[*] Code and data are available at: https://github.com/TDonofrio62/Predicting-College-Basketball-Success.

team statistics in the history of NCAA basketball. Basketball has experienced many changes over time, and the sport is played much differently than it used to be, so statistics in the sample likely vary from the seasons not taken into account. Since the goal of the model created was predicting future results, using only recent seasons would make our estimate more accurate relative to upcoming seasons as opposed to if the entire population of team data, dating back to the 1800s, was used.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - Each instance consists of a variety of advanced team statistics of a certain school in a season, like for example how many points were scored, shooting rates, percentages, and more. Simpler statistics such as how many games each team won and lost were also included.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - Each instance of team statistics is labeled by two variables. One stating the school of the team, and the other stating which season is being looked at in the observation.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
   - No, all team statistic values are filled out.
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - No, as there are no relationships between instances
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - No split was recommended, but the dataset was eventually split into two datasets when creating the linear model, one to train the model and make estimates, and the other to test and validate the model.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - No.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The datasets rely on taking data from the box scores, tracked by referees and scorekeepers, of each college basketball game. These will always exist as every game played must have one. The gamesheets are all held on to and archived by the NCAA, and provided to the public, so there are no restrictions.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - No.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - No.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - No.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - Yes, as the public could easily find the roster for each school team in a specific season for every instance of the dataset.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - No.
16. *Any other comments?*
    - No.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
    - The data used to compile the team statistics in the dataset was directly observable as the statistics are provided to the public by the NCAA, and then by Sports References in a clear and easily accessible way.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
    - All statistics for each college basketball game are systematically tracked and recorded by referees and scorekeepers. For use in this paper, the data was scraped off of the Sports Reference website by exporting files into excel and reading them into R.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - A cluster sampling strategy was used where seasons were grouped into clusters of 5, and the cluster containing the most recent 5 seasons was used in order to use data that was would be the most representative of games being played in the near future.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    - Thomas D'Onofrio, with no compensation as the project was completed for himself.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
    - The data has been collected by Sports Reference from 2017 to 2022, and the data was simply collected quickly all at once thanks to the work done by the website over time.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
    - The data was obtained via Sports Reference, who collected the data from the NCAA, who initially tracked and collected the data.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
    - All NCAA teams and players are well aware that the statistics of the games they play are being tracked and distributed. No notice was provided, but it is common knowledge among all athletes at all levels that this occurs.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
    - No.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - Consent was not obtained.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No.
12. *Any other comments?*
    - No.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - The data was cleaned where variables were removed, renamed, and also created.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - Yes, the raw datasets exported straight from Sports Reference can be found in the inputs folder at the GitHub page linked in the paper.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R was used, as well as the R packages tidyverse (Wickham et al. 2019) and janitor (Firke 2021).
4. *Any other comments?*
    - No.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Considering the popularity of the Sports Reference website, the datasets used to create the full dataset used in this paper have been used by many. One example of a paper using datasets from Sports Reference can be found at https://arxiv.org/pdf/2007.10550.pdf.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - No.
3. *What (other) tasks could the dataset be used for?*
    - The dataset has little use for tasks outside of analyzing the sport of basketball in many different ways.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - No.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No.
6. *Any other comments?*
    - No.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - The complete dataset used in the paper can be found on GitHub in the inputs folder.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - Apart from being on GitHub, the dataset will not be distributed in any other way.
3. *When will the dataset be distributed?*
    - The dataset is already available on GitHub
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - No. MIT License.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - No.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - No.
7. *Any other comments?*
    - No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Thomas D'Onofrio
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - thomas.donofrio@mail.utoronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - Not as of right now, but in the future instances from future seasons may potentially be added to the dataset to increase the sample size and have updated information. This would be done yearly at the end of every season, which is in April, and the updated datasets will be on GitHub.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - No.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
    - The older versions will not be maintained, as the dataset will just be updated over time. Previous versions of the dataset can be found on GitHub.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
    - Pull Request on Github.
8. *Any other comments?*
    - No.

References

Cannella, Stephen. 2002. *Seamheads, Rejoice Baseball-Reference.com Is the Ultimate Online Statistical Source for Fans with a Sense of History*. Sports Illustrated. https://vault.si.com/vault/2002/12/16/seamheads-rejoice-baseball-referencecom-is-the-ultimate-online-statistical-source-for-fans-with-a-sense-of-history.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Sports Reference. 2022. *Sports Reference | Sports Stats, Fast, Easy, and up-to-Date*. https://www.sports-reference.com/cbb/seasons/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.