Datasheet for 'NFLVerse Play-by-Play Data from 2023'*

Alexander Guarasci

2024-04-18

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 dataset was created to analyze play-by-play data of NFL teams.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

The nflverse authors created the dataset on there own behalf

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.*

Could not find information on this, the data was taken from the NFL.

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.*

The instances that comprise the dataset primarily represent an individual NFL snap, however, there are data on overall game stats (who won, etc). For each instance, there are multiple types, for example, there are stats on points scored, yards gained, where the ball is on the field, the expected points added of a play, and lots more.

2. *How many instances are there in total (of each type, if appropriate)?*

[*] Code and data are available at: https://github.com/AlexanderG123/nfl_analysis.

Just under 19 million

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).*

It contains all possible instances. 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 numeric and text data, all of which has been formatted apropriately

5. *Is there a label or target associated with each instance? If so, please provide a description.* No

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

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, the relationships are not made explicit

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.*

There is no recommended data splits

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

There are some redundancies, for example they have data on how many rushing yards and passing yards were gained on a specific play, however, if the ball was run it could not have been passed as well.

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 dataset is self-contained.

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, this dataset contains the names of individuals who were involved in certain plays.

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 is directly observable and reported by NFL referees and broadcasters. They also use technology, according to the NFL, they use “20–30 ultra-wide band receiversm, 2–3 radio-frequency identification (RFID) tags installed into the players’ shoulder pads, RFID tags on officials, pylons, sticks, chains, and in the ball. Altogether, an estimated 250 devices are in a venue for any given game.”

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?*

As stated above, the NFL uses an estimated 250 devices each game that they play, to validate these procedures, “a team of three operators is required at every game to confirm that all tracking systems are functioning properly.”

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

NA

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)?*

Thousands of people were involved in the data collection process, all of whom were compensated differently, for example, the referees make approximately $12000 for each game they participate in.

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.*

This data was collected over the 2023 NFL season, the time frame does match the creation timeframe of the data associated with the instances.

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)?*

No

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.* NA, working as a professional athlete requires that stats are kept while you are employed

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.*

NA

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).*

NA

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.*

NA

12. *Any other comments?*

No

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

This is a well-known dataset that has likely been used for a wealth of tasks

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?*

An analysis of how EPA impacts a team's likelihood of winning could be interesting. There are literally an infinite number of things that can be done with this data. 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.*

Because the data is play-by-play, in other words, it documents these data points for each snap, it would be unwise to use it for macro NFL trends, unless they require this granularity.

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.*

No

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

NA

3. *When will the dataset be distributed?*

NA

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.*

NA 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.* NA

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?*

The kind folks at NFLVerse

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

https://twitter.com/mrcaseb

3. *Is there an erratum? If so, please provide a link or other access point.*

NA

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)?*

NA

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.*

NA 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.*

Yes, they are all hosted by NFLVerse

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 consumewrs? If so, please provide a description.*

NA

8. *Any other comments?*

No

1 References