Datasheet for NFLFastR*

Vandan Patel

2024-12-03

This datasheet provides a comprehensive overview of the dataset used to analyze the relationship between quarterback performance metrics and team success in the NFL's 2023 season. It details data provenance, variable descriptions, processing steps, and potential limitations to ensure transparency and reproducibility for future research and applications.

1 Motivation

Why was the dataset created? The dataset was created to analyze the relationship between quarterback (QB) performance and team success during the 2023 NFL season. This includes exploring metrics such as passing yards, touchdowns, interceptions, wins, and playoff qualifications. It aims to provide a comprehensive resource for understanding the impact of QBs on team outcomes while recognizing football as a team sport.

Who created the dataset? The dataset was derived from the NFL's publicly available play-by-play data using the `nflfastR` package (Carl and Sharpe 2023). Additional processing and aggregation were performed using R's `tidyverse` (Wickham et al. 2019).

Who funded the dataset creation? The dataset creation was part of an independent research initiative for academic purposes.

2 Composition

What does the dataset contain? The dataset includes:

- **Quarterback Metrics:**
  - Passing yards

*Code and data are available at: https://github.com/vandanppatel/impact_qb_nfl/.

  - Touchdowns
  - Interceptions
  - Average per-game performance metrics

- **Team Metrics:**
  - Total wins
  - Average score differential
  - Playoff qualification **How was the data collected?** The data was collected from publicly available NFL play-by-play data using tools like `nflfastR`. Advanced statistics and aggregation were performed using R packages, including `janitor` for data cleaning and `arrow` for efficient data storage.

Does the dataset contain all possible data points? No, it excludes preseason and playoff data to focus on regular season performance.

Are there known errors or gaps? Possible limitations include: - Exclusion of contextual variables like weather or injuries. - Potential inconsistencies in manually recorded NFL play-by-play data (**wirednfl?**).

3 Use and Distribution

How is the dataset distributed? The dataset is distributed in a Parquet format to ensure efficient storage and faster I/O operations.

Who can use this dataset? The dataset is open to researchers, analysts, and students interested in exploring NFL analytics. Ethical use is encouraged to respect players' privacy and the integrity of the sport.

Are there licensing or usage restrictions? The dataset adheres to MIT licensing. Users must credit the source and ensure the data is used responsibly.

What are potential ethical issues? While the dataset anonymizes individual player actions, ethical concerns arise when analyzing performance data. For instance, overemphasis on metrics like interceptions may lead to undue criticism of players (**pmcnfl?**).

4 Methodology

How was the data processed?

1. **Cleaning and Standardization:** Using `janitor` to clean raw play-by-play data.
2. **Aggregation:** Summarizing QB and team-level metrics with filtering for starting QBs.

3. **Storage:** Saving processed data in Parquet format using the `arrow` package for reproducibility and efficiency. **Were there any transformations?** Yes, raw metrics like passing yards were aggregated at a season level, and team metrics such as wins and playoff qualification were computed from game results.

What limitations exist in the methodology? The methodology does not account for external variables such as: - Schedule difficulty - Injuries - Coaching strategies

5 Maintenance

Will the dataset be updated? No, this dataset focuses exclusively on the 2023 season. Future datasets may include subsequent seasons.

Is there a point of contact? Yes, you can reach out to vandanp.patel@mail.utoronto.ca for further inquiries.

References

Carl, Sebastian, and Lee Sharpe. 2023. *nflfastR: Functions to Efficiently Access and Analyze NFL Play-by-Play Data*. https://CRAN.R-project.org/package=nflfastR.

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.