Datasheet*

F1 World Championship (1950 - 2023) reliability

Zhijun Zhong

April 12, 2024

This datasheet is used to analyze the lap time in qualifying for F1 World Championship from 2006 to 2023. It aim to explore the strategic optimization and estimate the fastest lap time while maintaining minimum tyre wear. In the dataset, it contained historical data from official Formula 1 records, including lap times, and driver information. This datasheet provides detailed descriptions of the motivation, composition, collection methodology, preprocessing steps, uses, distribution and maintance which aim to focus on reproducibility and mutiple usage of the dataset for further research in sports analyze and strategy optimization in F1

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.*
    - This dataset is created by Rohan Rao aiming to provide general public a readable dataset about the past races which F1 offical website does not provide where people can analyze driver's data after every race.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - This dataset is created by an individual called Rohan Rao who is on behalf of himself.
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.*
    - No one funded this dataset. Rohan Rao imported the dataset from Ergast Developer API which is for non-commercial purpose.

*Code and data are available at: https://github.com/JerrZzzz/Strategic-Optimization-in-F1.

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.*

    - The original dataset contained a lot of different files where every one of the file represents various thing. I will talk about the files that I used in my paper. All files displayed are inter-connected. For the qualifying file, circuitId is inter-connected with specific circuit information in the circuit file. raceId is inter-connected with specific race information in the race file. Every row represents a driver's lap time in a specific race. For example, Max verstappen achieved a final grid position of 1 in raceId 1230 with Q1, Q2, Q3 fastest lap time displayed.

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

    - There are raceId going from 1 to 1100. driverId going from 1 to 858. Position going from 1 to 28. and lap times below 3 minutes per lap.

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 does not contained all possible instances which more recent races are not included. However, the dataset had contained all possible qualifying results in the history. I can say that the dataset we used can represent almost all population in history.

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.*

    - It consist of raceId which is a number represent each race. DriverId used a number to represent all drivers. ConstructorId used numbers to represent the team that the driver is in at that race. Position represents the final grid position the driver is in. Q1, q2, q3 represents the driver's fastest lap time of each qualifying sections while those eliminated driver's lap time in further qualifying will represent with NA values.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

- The label or target in my dataset is the lap time for q1, q2 and q3. Each of them containe a lap time each driver achieved in that specific qualifying session.

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.*

    - For those driver being eliminated from q1 and q2. They will have missing value to q2, q3 lap time and q3 lap time respectively. It is not intentionally removed but no value can appear in those areas because driver are not allow to take part in those qualifying sessions.

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.*

    - RaceId are interconnected in datasets. Race dataset described all races in F1 history. Every race had being addressed to a raceId where other datasets like qualifying will be using these raceId to identify the qualifying for that race.

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.*

    - Yes, we can split data for different circuits to validate our model and use a more recent data to test the fit of our model.

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

    - By looking at the data and modeling them, we found that there is no error of measurement and redundancies in the dataset. There are a lot of factors can influence our model where it made the uncertainty of our model a bit high. But we expect the noises because of real life factors.

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.*

    - Our dataset is partly external. The circuit information, team information, Driver information are dependent on wikipedia. However, other information about races

and the lap time for example that we used do not depend on ourside sources. The outside sources will not influence our model and the whole paper. The term and conditions are stated in the API website where we must not create any application which polls the API more than four times per second or more than 200 times per hour.

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, all the data are considered public. Even though that the lap times are not published by the offical website of F1. But it does not offend any personal information. We can find all possiable information in F1tv.

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.*

    - The dataset is not offernsive to any individual while not cause any public anxity. The dataset can be considered as recreational.

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. There is no specific sub-populations being identified in dataset.

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, we are able to identify individual person in our dataset with the name presented and the sport itself. However, no personal information had revealed in the dataset. All information in the dataset are strongly relate to F1 motorsport.

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, races, sexual information, belief systems and so on are not stated in the dataset. Only the name, Birth date and nationality of the individual are stated.

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.*

    - Our dataset is a direct observation while taking the data from its source directly without transformation. The author of our dataset extract the file from a API website where it is a direct observation.

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?*

    - Authors used python programs to create the dataset where we can found the python code on the API website here: https://ergast.com/mrd/. There are workers validating the dataset using python and being stated in the API website here: https://ergast.com/mrd/development-tools/

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

    - The dataset is not a sample from a larger set.

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)?*

    - API workers are involved in the collection process while there are also volunteers who created tools to help validate the dataset.

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.*

    - We expect the dataset to update every half a year. The creation of the data time is not stated. The inital release of this dataset is 4 years ago.

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 ethical review had conducted.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

- I collect the data from third party called kaggle. The dataset information can be found here: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020/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.*

    - No, it is a public dataset where individuals can download the data freely.

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.*

    - The term of usage in Kaggle provide user with right to use collect the data published by authors.

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).*

    - No, only if the author delete the dataset on Kaggle or Kaggle does not allow users to download the datasets.

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.*

    - There is no analyze on the potential impact of the dataset.

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.*

    - Preprocessing only happened when the data is first downloaded by API workers. However, their aim is to provide public with true and honest data instead of fabricate information.

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.*

    - No, the raw data is the collected data. There is no raw data which had not "cleaned" or "preprocessed".

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - No.

4. *Any other comments?*

    - No.

Uses

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

    - The dataset that I downloaded from Kaggle is not intended for any tasks. However, there are other users in Kaggle provide their own task using the raw data provided by Rohan Rao.

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.*

    - Yes, we can find all other users' work here: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020/code

3. *What (other) tasks could the dataset be used for?*

    - This dataset had being used for finding out the relationship between altitude and engine failures and so on.

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?*

    - We can say that the dataset represents almost population. There is only a few data missing from all population which is because of the update speed of the dataset.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

- The dataset is open to public.

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, the dataset is open to public extracted from ergast API.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - No, the dataset does not have a DOI.
3. *When will the dataset be distributed?*
    - No
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.*
    - It is published under the CC0: Public Domain 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.*
    - Yes, the Ergast API restrict users to create any application which polls the API more than four times per second or more than 200 times per hour.
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 author - Rohan Rao.

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

    - We can contract the owner of the dataset using twitter, instagram. Here (**vopani?**), LinkedIn, https://www.linkedin.com/in/vopani.

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)?*

    - Yes, The dataset will be updated about every half a year. However, the Ergast API stated that they will not further update at the end of 2024.

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.*

    - The dataset will be retained unless Author delete it or kaggle delete it.

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.*

    - Older versions can be found in the version buttom. However, the dataset will not continue maintaining at the end of 2024.

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.*

    - Yes. Other authors can send notes under the dataset in kaggle here: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020/code.

8. *Any other comments?*

    - No

References

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.