UK Department for Transport Road Traffic Collision Dataset

Model datasheet

Motivation

Answer the following questions:

- 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.
    - Answer:
    - UK police forces collect data on every vehicle collision in the uk on a form called Stats19. Data from this form ends up at the DfT
    - This dataset provides a better understanding of the factors associated with road traffic collisions
- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? Who funded the creation of the dataset?
    - Answer:
    - This dataset is provided by the united Kingdom's Department for Transport based on information recorded by UK Police forces
- Any other comments?
    - Answer:
    - There are 3 CSVs in this set. Accidents is the primary one and has references by Accident_Index to the casualties and vehicles tables. This might be better done as a database.

Composition

Answer the following questions:

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Please provide details.
    - Answer:
    - There are 3 CSVs in this set. Accidents is the primary one and has references by Accident_Index to the casualties and vehicles tables. This might be better done as a database.
    - There are a number of instances rallying to many factors relating to the casualties, vehicles involved and environmental conditions.
    - Crucially, it includes data on the severity of the injuries sustained.
- How many instances of each type are in total?
    - Answer:
    - The dataset included a total to 4287593 instances
- 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 (e.g., 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 (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
    - Answer:
    - This data includes instances representing everything recorded by the Police forces
- What does each instance consist of? Raw data? Unprocessed? Text, images?
    - Answer:
    - Each instance contains tex (numerical) data.
    - The data has been assembled by Police Forces as part of an ongoing study running between 2005 and 2015
- Are there any labels to the data?
    - Answer:
    - The data is labelled according to the predictor
- Is there any missing information from individual instances?
    - Answer:
    - There is no missing or mismatched data
- Are relationships between individual instances made explicit?
    - Answer:
    - There are no relationships between the instances other than they have all been involved in unrelated accidents.
- Are there recommended data splits (e.g. train / test)? Provide a description of the splits, and the rationale behind them.
    - Answer:
    - There is no recommended data split
- Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
    - Answer:
    - The dataset is self-contained
    - It originally comes in three css files that can be merged but does not link to other data
- Does the dataset contain data that might be considered confidential (e.g., 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.
    - Answer:
    - This includes instances of sex and age group but this is not linked to a named individual.
- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
    - Answer:
    - This dataset does not include data that could induce anxiety or create insult or threaten when viewed directly
- Does the dataset identify any subpopulations (e.g., by age, gender)?
    - Answer:
    - This dataset identifies sub-populations including gender (female) and age (>21)
- Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
    - Answer:
    - It is not possible to identify individuals
- Does the dataset contain data that might be considered sensitive in any way (e.g., 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.
    - Answer:
    - The data does not include anything that is sensitive in terms of identifying beliefs or race.
    - It does not identify any governmental identification means nor personal beliefs.
- Any other comments?
    - Answer:
    - The dataset does not provide any data that would be considered sensitive

Collection process

Answer the following questions:

- How was the data associated with each instance acquired? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
    - Answer:
    - The data was acquired though Police Forces recording data following every RTC in the UK
- If the data is a sample of a larger subset, what was the sampling strategy? Deterministic, random, etc...?
    - Answer:
    - The dataset is not part of a larger dataset.
    - Everyone involved in an RTC was tested.
- Over what time frame was the data collected?
    - Answer:
    - Data was recorded between 2005 and 2015
- Were there any ethical review processes conducted (e.g. by an institutional reviewing board?)
    - Answer:
    - No there are no recorded ethical reviews recorded
- Were the individuals notified of the collection of the data?
    - Answer:
    - Not known
- Did the individuals consent to their data being collected?
    - Answer:
    - Not known
- If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
    - Answer:
    - This is not known
- Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
    - Answer:
    - This dataset has provided a well validated data resource in which to explore prediction of the severity of the accident
- Any other comments?
    - Answer:

Preprocessing/cleaning/labelling

Answer the following questions:

- Was any preprocessing/cleaning/labeling of the data done (e.g., 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.
    - Answer:
    - The data was not pre-processed prior to publication on Kaggle
- Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
    - Answer:
    - The raw data remains available
- Any other comments?
    - Answer:

Uses

Answer the following questions:

- What other tasks could the dataset be used for?
    - Answer:
    - The dataset could be used for predictions about location of accidents, lighting conditions and so on.
- 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 (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
    - Answer:
    - It is unlikely that the selection of certain instances would cause harm
- Are there tasks for which the dataset should not be used? If so, please provide a description.
    - Answer:
    - Not likely
- Any other comments?
    - Answer:

Distribution

Answer the following questions:

- Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
    - Answer:
    - The dataset is widely and openly available on the internet
- How will the dataset be distributed?
    - Answer:
    - The dataset is widely and openly available on the internet
- When will the dataset be distributed?
    - Answer:

    - The dataset is widely and openly available on the internet
- 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.
    - Answer:
    - The dataset is widely and openly available on the internet
    - There are no licensing conditions
- Any other comments?
    - Answer:

Maintenance

Answer the following questions:

- Who will be maintaining the dataset?
    - Answer:
    - There is no existing schedule for maintenance or update of this data.
- Any other comments?
    - Answer: