Professional Certificate in Machine Learning and Artificial Intelligence

Datasheet for Dataset:

Capstone project – Kenneth French Data Library

Motivation

- 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 for research on asset allocation purposes.
The underline task in mind was to be able to explain the market behaviour according to different factors (for the famous CAPM model and 3-5 factor model).
Kenneth French collected historical stock prices for individual stock and then grouped them according to industrial or factor criteria like company, earning, book size etc …

- 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?

The dataset was created by Kenneth French for his research at the Chicago Booth School of Business and at the university of Tuck School of Business at Dartmouth College now.

- Any other comments?

N/A

Composition

Answer the following questions:

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Please provide details.

The dataset represents a collection of market return time series for different types of industries. Those industries are classified by Compustat or CRSP SIC codes.

- How many instances of each type are in total?

As of September 2024, there is 1177 monthly returns from July 1, 1926, to July 31, 2024 for 49 industry features.

- 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)

By construction it contains a very large size of US stock of all US stocks because the data come from NYSE, AMEX and NASDAQ and is categorized with the Compustat SIC code for the fiscal year ending in calendar year t-1 if any else with the Compustat SIC code is not available the CRSP SIC code is used instead.

- What does each instance consist of? Raw data? Unprocessed? Text, images?

Each instance consists of raw data of market returns based on single stock returns over a month and aggregated according to his market capitalization.

- Are there any labels to the data?

The data is a scalar number so not labelled data.

- Is there any missing information from individual instances?

Sometime some data is missing because either no stock performance was found (during war or market crash or not publicly disclosed) or more basically when the sector did not exist.

- Are relationships between individual instances made explicit?

There are no relationships between individual instances.

- Are there recommended data splits (e.g. train / test)? Provide a description of the splits, and the rationale behind them.

No recommendation is made between train and test split.

- Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The data is self-contained. However, it is an aggregate of stock performances, so we do not have vision from which stocks the industry performance comes from.

- 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.

The dataset does not contain data that might be considered confidential. The data protected by privilege is in general stock prices because you do not have access in general to all the history of stock data. This layer is hidden through the work of Kenneth French.

- 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 data is offensive, insulting or threatening.

- Does the dataset identify any subpopulations (e.g., by age, gender)?

No subpopulations are identified.

- 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.

N/A

- 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.

No

- Any other comments?

No

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.

The data associated with each instance has been collected by Kenneth French using CRSP, AMEX, NYSE and NASDAQ data. For very old stock data CRSP is the main provider. CRSP (Center for Research in Security Prices affiliated to the University of Chicago) has collected very large among of stock data since 1926. For more recent data market providers like NYSE provide data.
The main purpose of the collection is to provide raw data for financial study and modelling.

If the data is a sample of a larger subset, what was the sampling strategy? Deterministic, random, etc...?

- Over what time frame was the data collected?

N/A

- Were there any ethical review processes conducted (e.g. by an institutional reviewing board?)

N/A

- Were the individuals notified of the collection of the data?

N/A

- Did the individuals consent to their data being collected?

N/A

- If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

N/A

- 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?

N/A

- Any other comments?

No

Preprocessing/cleaning/labelling

- Was any preprocessing/cleaning/labelling 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.

Some preprocessing is done before providing the data like the clearing of stocks delisted during observation period or missing returns for a period.

- Was the “raw” data saved in addition to the pre-processed/cleaned/labelled data (e.g., to support unanticipated future uses)?

No raw data is provided.

- Any other comments?

No

Uses

Answer the following questions:

- What other tasks could the dataset be used for?

This dataset is made for being used mainly for financial studies and modelling.

- Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labelled 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?

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

No

- Any other comments?

No

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.

The data is distributed online through the Kenneth French website as CSV.

- How will the dataset be distributed?

N/A

- When will the dataset be distributed?

N/A

- 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.

The dataset seems not be distributed under a copyright or IP license because nothing is mentioned on the Kenneth French website.

- Any other comments?

No

Maintenance

Answer the following questions:

- Who will be maintaining the dataset?

The dataset is still maintained by Kenneth French monthly. He provides monthly performance for each industry.
He also provides historical adjustments when some provider made adjustment on past methodology or include more historical data. The full history of returns is reconstructed monthly.
Those changes happen for example when CRSP revises its database or complete the data about outstanding shares data.
Some methodological changes could be also made for example on criteria when excluding a stock (no more quoted for example) or the consistency of the referential for sectorial classification (provided by CRSP or Compustat).

- Any other comments?