Datasheet for Investment Strategy Final Project

**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.

*This dataset was created to explore how different financial, macro-economic and behavioural features of equity markets might influence the relative performance of individual equity markets vs. the average. In other words, do certain equity market features bear information about the future performance of that equity market?*

* 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?

*This dataset was created by the Strategy team at Commonwealth Superannuation Corporation (’CSC’), a large pension fund based in Sydney, Australia. The dataset was created in 2022 and was funded by CSC.*

**Composition**

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

*The dataset consists of 19 feature datasets and 3 excess return datasets. Each dataset contains monthly information on one feature for each of 14 different equity markets.*

*These markets are – Australia, Canada, France, Germany, Hong Kong, Italy, Japan, Netherlands, Singapore, Spain, Sweden, Switzerland, UK and the US.*

*The features include financial features (‘Earnings yield’, ‘Book/Price’), macro-economic features (‘Leading indicators’) and behavioural features (‘1 Month Momentum’).*

*Other than the 3 character codes used to identify each country and a date field, all the feature data is purely numeric.*

* How many instances of each type are in total?

*Each feature dataset contains one row per month per country. The feature history runs from Jan 1991 to March 2021 comprising 363 rows in total per feature x 14 countries. As noted above, there are 19 feature.*

* Are there any labels to the data?

*In each dataset the countries are labelled: Australia (‘AUD’), Canada (‘CAD’), France (‘FRF’), Germany (‘DEM’), Hong Kong (‘HKD’), Italy (‘ITL’), Japan (‘JPY’), Netherlands (‘NLD’), Singapore (‘SGD’), Spain (“ESP’), Sweden (“SEK’), Switzerland (‘CHF’), UK (‘GBP’) and the US (‘USD’).*

*Each feature is tagged on the spreadsheet tab name.*

* Is there any missing information from individual instances?

*There are no NULL values but some missing data is represented as ‘0’.*

* Is there any missing information from individual instances?

*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.

*As noted above the dataset runs from Jan 1991 to March 2021.*

*The training dataset (used in the model) runs from 1991 to 2010*

*The validation dataset runs from 2011 to 2015*

*The test dataset runs from 2016 to 2021*

*The rationale for these splits is:*

1. *The training set needs to comprise at least 60% of the dataset*
2. *The validation set and the testing set each comprise c. 20% of the dataset*

*Firstly, since this is financial data that exhibits strong autocorrelations, randomly splitting the data would an error so we preserve the autocorrelation by having temporarily contiguous sections.*

*Secondly, we want the validation and testing dataset to be as close as possible to the current date so we can have more confidence in deploying the model that its out-of-sample dataset comes close to matching current conditions.*

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

*This dataset is self-contained.*

* 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.

*There is no information in the dataset that would be considered confidential.*

* 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*

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

*No*

* 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.

*No.*

* 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.*

**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 is a time-series of general financial, macro-economic and behavioural data associated with each market over time.*

*The data comes from a variety of commercial sources including Bloomberg databases, MSCI databases and OECD databases.*

*It was constructed at CSC to assist in exploring whether this data had any predictive power for future returns in an individual market.*

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

*No it isn’t.*

* Over what time frame was the data collected?

*The feature time-series is a monthly time-series that runs from Jan 1991 to march 2021.*

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

*This is purely market-based information so no ethical review was needed.*

**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.

*Each data time-series was winsorized using a 3 standard deviation band. This effectively mitigated any large outliers.*

*Each row of the each feature was then cross-sectionally z-scored so we could compare the same feature across countries.*

*Finally, any NaNs were replaced with 0s, effectively making them an ‘average’ z-score observation.*

* Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

*Yes it was.*

**Uses**

Answer the following questions:

* What other tasks could the dataset be used for?

*Given the specificity of the dataset, this dataset is used primarily for asset allocation research.*

* 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?

*N/A*

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

*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.

*No it will not.*

* How will the dataset be distributed?

*The dataset will be posted to a private Github database.*

* When will the dataset be distributed?

*The dataset is available now.*

* 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 will not be distributed.*

**Maintenance**

Answer the following questions:

* Who will be maintaining the dataset?

*The database may be periodically updated by CSC.*