Data Sheet for Machine Learning Project

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: The dataset was created to analyze the relationship between various economic indicators and the price of Gold. The goal is to build a predictive model that can forecast Gold prices based on the provided economic variables.

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: The dataset was gathered from yahoo finance and various sources from the internet.

Any other comments?

Answer: None.

Composition

Answer the following questions:

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

Answer: Each instance represents a monthly observation of economic indicators and Gold price.

How many instances of each type are in total?

Answer: Data is daily timeseries from 1996 to 2023.

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: Dataset contains daily data from 1996 onwards.

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

Answer: Each instance consists of numerical values representing economic indicators (Gold, US Inflation index, Dollar Index, Fed Fund Rate, SPX Index, Vix, 10-year Government Bond Yield, 2-year Government Bond Yield, Unemployment) and the corresponding Gold price.

Are there any labels to the data?

Answer: Yes, the Gold price is the target variable (label).

Is there any missing information from individual instances?

Answer: Yes, there are a few missing values in the 'Unemployment' variable, which were handled during preprocessing.

Are relationships between individual instances made explicit?

Answer: No, the relationships between individual instances are not explicitly defined.

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

Answer: Yes, the dataset is split into a training set (66% of the data) and a test set (33% of the data) to train and evaluate the machine learning model's performance effectively.

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 and does not rely on external resources.

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: No, the dataset does not contain any confidential or sensitive information.

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: No, the dataset does not contain any offensive or harmful content.

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

Answer: No, the dataset does not identify subpopulations.

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: No, the dataset is aggregated and anonymized, preventing the identification of 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: No, the dataset does not contain sensitive information of any kind.

Any other comments?

Answer: None.

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 associated with each instance was acquired through a combination of public data sources, financial databases, and government reports. The specific task was to gather economic indicators and Gold price data for analysis.

Over what time frame was the data collected?

Answer: The data was collected on one fine day from multiple websites, it’s a historical timeseries data.

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

Answer: No

Were the individuals notified of the collection of the data?

Answer: Not applicable, as the data represents economic indicators and market prices, not individual-level data.

Did the individuals consent to their data being collected?

Answer: Not applicable, as the data does not involve individual-level information.

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

Answer: Not applicable, as the data does not involve individual-level information.

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: Not applicable, as the data does not involve individual-level information.

Any other comments?

Answer: None.

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: Yes, preprocessing was performed to handle missing values in the 'Unemployment' variable using mean imputation. The data was also normalized to bring all features to a similar scale.

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

Answer: Not applicable.

Any other comments?

Answer: None.