**Professional Certificate in Machine Learning and Artificial Intelligence**

Required activity 25.3

Portfolio project on optimising a model for real-life data

**Motivation**

The dataset was created to build a machine learning model to accurately predict the direction of where the stock prices is heading.

Dataset are pulled directly from Yahoo! Finance API. Market data are downloaded from Yahoo! Finance’s API with the intention of personal usage only. Note that Yahoo, inc is not affiliated with this.

* License: Apache Software
* Author: Ran Aroussi
* <https://pypi.org/project/yfinance/>

Feature used to train the model is the closing price of the stock on the trading day.

**Composition**

The dataset consists of TESLA (‘TSLA’) stock closing price. Dataset is obtained from external resources, Yahoo! Finance API.

Consisting of up to 3347 and growing each trading day and features containing Open, High, Low, Close, Adjusted Close, Volume. There are also other data that could be imported as the dataset such as income statement, balance sheet, cash flow, news, and such. However, this model will be sticking with just the ‘Closing’ price for simplicity’s sake. Also noting, due to the inherently stochastic nature of the data, dataset will be using 3 months, which contains the latest 64 instances.

Data contains raw numeric format and historical price (‘Closing’) is used as input label for the model. No missing information from the input features and no outliers are detected.

Recommended data splits are 80% training, 5 are used for testing and the rest are for validation purposes. As the model is built to predict 5 days, it is not required to allocate more than required valuable resources to test the model.

Dataset is not considered confidential as information are readily available to be downloaded. Dataset was made public and are not considered sensitive information that will cause discomfort / harm to anybody.

**Collection process**

Over a time frame of 3 months are used and data is a subset of the original dataset.

Dataset are solely for intended for research and educational purposes only and are not affiliated, endorsed or vetted by Yahoo, inc. Dataset is an open-source tool and does not require consent or notification to the authors of the datasheet.

**Preprocessing/cleaning/labelling**

Dataset are filtered to have the latest 3 months data and are normalized by a min/max scaling.

**Uses**

Dataset are not susceptible to unfair treatment of individuals or groups. Note that the model is not intended to provide financial advice and trades made upon the model’s prediction are not liable to monetary loss/gain. User are advised to conduct their own due diligence when making trades.

**Distribution**

As an open-source tool, dataset is free to download and use for educational or research purposes. Dataset was made public.

* 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: Dataset license was made CC0 (Public Domain)

**Maintenance**

Maintenance is done by the Author Ran Aroussi.