DOW JONES 30

A dataset contains day-level price and volume information of 29 stocks from DJ 30 index with a total of 87609 data samples.

Dataset Snapshot

NATURE OF CONTENT

Date, trading price (open, high, low, close price), trading volume, stock ticker and day of the week (0, 1, 2, 3, 4 representing Monday to Friday).

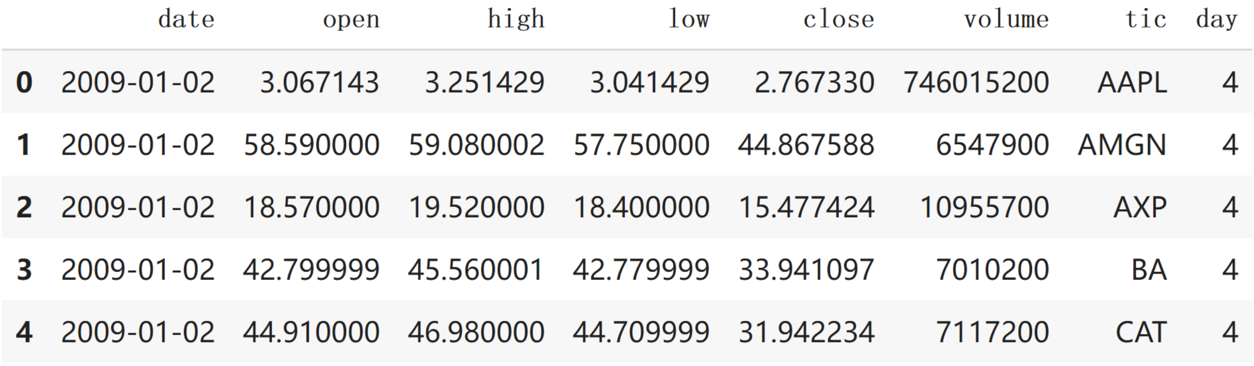
BREAKDOWN-BY INSTANCE

Total instances	87609
Training	70093
Validation	8758
Testing	8758
Total stocks	29
Instances per stock	3021

NOTES

Stock data is collected daily from 2009-01-01 to 2020-12-31 on all trading days. The recommend split is [0.8,0.1,0.1] for training, validation and testing respectively.

EXAMPLES OF ACTUAL DATA POINT



	date	open	high	low	close	volume	tic	day
0	2009-01-02	3.067143	3.251429	3.041429	2.767330	746015200	AAPL	4
1	2009-01-02	58.590000	59.080002	57.750000	44.867588	6547900	AMGN	4
2	2009-01-02	18.570000	19.520000	18.400000	15.477424	10955700	AXP	4
3	2009-01-02	42.799999	45.560001	42.779999	33.941097	7010200	BA	4
4	2009-01-02	44.910000	46.980000	44.709999	31.942234	7117200	CAT	4

Publisher & Licenses & Access

PUBLISHER

Yahoo Finance

DIRECT LINKS TO DATASET

GitHub link

ACCESS

Open Access

ACCESS COST

Free

LICENSE PERMISSIONS

You are free to share and adapt.
Attribution required.
You cannot apply any additional restrictions.

LICENSE TYPE(S)

CC-BY-4.0

Motivations & Use

DATASET PURPOSE

The dataset was created to provide data of US large companies' stock trading for research in various quantitative trading tasks by selecting the most prominent stocks at New York Stock Exchange and NASDAQ.

INTENDED USE CASES

- Algorithmic trading
- Portfolio Management

EXTENDED USE

- Intraday trading
- High frequency trading

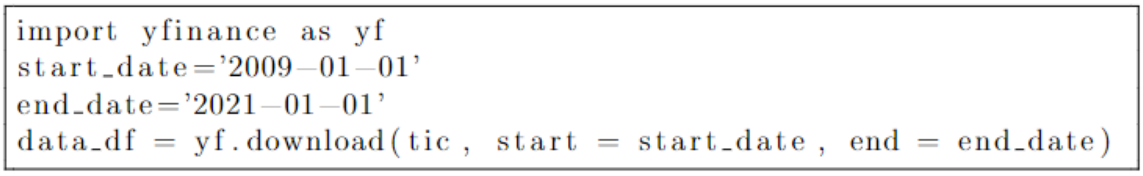
Collection

DATA SOURCE

Trading information of stocks at Nasdaq and New York Stock Exchange, retrieved from Yahoo! Finance API

DATA COLLECTION

Retrieved from Yahoo! Finance API using the following code, where tic is the stock ticker from Dow Jones 30 stock list:



```
import yfinance as yf
start_date=’2009-01-01’
end_date=’2021-01-01’
data_df = yf.download(tic , start = start_date , end = end_date)
```

Preprocessing

INDICATOR ADJUSTMENT

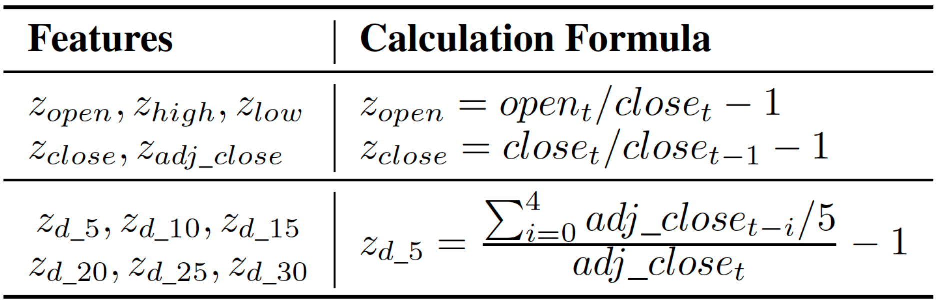
The raw data consists of 8 indicators, which are date, open, high, low, close, adjcp, volume and tic. Our dataset uses adjusted close price (adjcp) to replace original close price because it is considered as a more accurate measure of stock’s value.

DATA CLEANING

Firstly, all the NaN terms are dropped, and it is observed that some of the stocks are lack of data (number of instances less than 3021). In order to maintain consistency, data of these stocks are filtered out.

FEATURE GENERATION

We generate 11 temporal features to describe the financial markets. z_{open}, z_{high}, z_{low} represent the relative values of the open, high, low prices compared with the close price at current time step, respectively. z_{close} represents the relative values of the closing prices compared with time step t-1. z_{dk} represents a long-term moving average of the adjusted close prices during the last k time steps compared to the current close prices. The detailed calculation formulas are as follow:



Features	Calculation Formula
$z_{open}, z_{high}, z_{low}$ z_{close}, z_{adj_close}	$z_{open} = open_t / close_t - 1$ $z_{close} = close_t / close_{t-1} - 1$
$z_{d\_5}, z_{d\_10}, z_{d\_15}$ $z_{d\_20}, z_{d\_25}, z_{d\_30}$	$z_{d\_5} = \frac{\sum_{i=0}^{4} adj\_close_{t-i}/5}{adj\_close_t} - 1$

Maintenance & Status

STATUS	FIRST RELEASE	CURRENT VERSION
Actively Maintained	08/2022	1.0