CS 363D Data Selection

Group Members: Zhiyao Bao, Andy Li, Yuhan Zheng, Jesse Zou

1. Describe the dataset. (ex: The number of microaneurysms found in a patient's eye and whether or not they have diabetic retinopathy; or Lending Tree loan data, including who defaulted and who paid off their loan.)

Predicting the stock index movement (i.e., UP or DOWN?) using some useful features like opening price, closing price, and etc.

2. How many records does the dataset have?

7344

3. How many features does the dataset have? List or describe a few of them.

There are 13 features.

1) Date: The date the data is collected
2) Open: The price at which a security first trades upon the opening of an exchange that day
3) Close: The last price at which a security traded that day
4) High: The highest price that day
5) Low: The lowest price that day
6) Volume: The total number of shares that are actually traded that day
7) InterestRate: The interest rate that day
8) ExchangeRate: The exchange rate that day
9) VIX: A real-time index that represents the market's expectations for the relative strength of near-term price changes
10) Gold: The gold price that day
11) Oil: The oil price that day
12) TEDSpread: The difference between the three-month Treasury bill and the three-month LIBOR based in U.S. dollars
13) EFFR: A volume-weighted median of overnight federal funds transactions

4. What can you try to predict in this dataset? (ex: We can use the number of microaneurysms measured in the patient's eye to predict whether or not they have diabetic retinopathy; or We can try using the features, including age, income, home ownership status, etc, to predict whether or not someone will default on their loan.)

We can use the features, including time(date), open price, close price, the highest price, lowest price, etc., to try to predict whether or not the stock index movement will go up or down the next day.

5. Is this a **labeled** dataset, appropriate for a supervised learning classification problem? (In other words, if you are trying to predict whether or not someone has a disease, does your dataset contain whether or not each record has the disease?)

The stock index prediction dataset is a labeled dataset.

6. Provide a link to the dataset, if there is one. If you are getting your data from somewhere other than a link, where are you getting it from?

https://www.kaggle.com/jmq19950824/stock-index-prediction-both-labels-and-features?select=SP500.csv