I have used the netflix stock price from 2018-2022 as our dataset

In our dataset we have the following features: Open price, close price, highest and lowest prices, sentiment, no of tweets and volume.

Classified the tweets as positive, negative or neutral based on whether the P\_mean value is above, below or equal to 0 respectively.

Assigned pos the value +1, neg -1 and neu 0 values.

Multiplied the sentiment values by the corresponding no. of tweets to get the tweet factor which we have used, replacing the 2 features.

Y is the adjusted closing price of the stock.

Applied models such as : linear regression, neural networks and CNN.

Particularly focussed on linear regression to analyze the effects of the features independently and also in conjunction with other feature(s) to identify more relevant features.

Trimmed the dataset to 10 data points - using 8 data points to predict the other 2 points.

On doing the feature-wise analysis, I found the following:

Feature MSE

**All 0.00021500089915039378**

Volume  **1.6640217319371298**

Open 0.04747987034452305

Close **0.00021500089914298356**

High 0.024203024075131828

Low 0.027122174068776254

Twt\_factor  **0.8303625167069273**

Next let's consider combinations of features.

Close + High **0.00021500089914298353**

Close + twt\_factor  0.00021500089914298356

Volume + twt\_factor **1.0801152523891053**

Close + twt\_factor + High 0.0002150008991468026

Close + twt\_factor + Low 0.00021500089914734308

Close + twt\_factor + Open 0.00021500089914568263

Close + twt\_factor + Volume **0.00021500089914445517**

ALL - High 0.00021500089914744077

All-Low 0.00021500089914675214

All- volume 0.0002150008991503417

All- open 0.000215000899150384

All -close **0.04045901148445122**

#Conclusion: close, Twt factor and Volume are more important features compared to others.

We can also give importance to (close + twt) factor only.

Twt factor is necessary because it is the basis of the research work, otherwise observation is that close + high is the best combination for our set of data points.

COnsidering data points :

Data Point MSE

X[0] 1.970889182980934

X[1] 1.274171233555588

X[2] 1.1852751247497808

X[3] 0.6199280243063392

X[4] 0.4160007315182804

X[5] 0.6961465106329637

X[6] 0.37515466210371445

X[7] **0.08199510845382817**

X[7] is most near to our test cases of x[8] and x[9].

All+x[7] 0.00021500089917462687

ALl+2\*x[7] 0.00021500089914298028

All+3\*x[7] **0.00021500089914296562**

All+4\*x[7] 0.00021500089914299332

ALl+5\*x[7] 0.00021500089914300795

All+6\*x[7] 0.00021500089914298679

All+7\*x[7] 0.0002150008991429982

All+10\*x[7] 0.00021500089914299004

For our small dataset, taking x[7] 3 times seems to be the optimal way.

Now, let's take 8 data points to just predict the 9th point i.e. x[8].

Data points MSE

**All 1.1122958485138897e-25**

X[0] 1.0

X[1] 0.6441337828281153

X[2] 0.5987811416687832

X[3] 0.3108503390248692

X[4] 0.20735003760568363

X[5] 0.34960028938797644

X[6] 0.1866615769975712

**X[7] 0.03917087297931499**

**The previous day is very much important for predicting the next day.**

**Still combined effect of last 8 days is far better.**

Now, lets extend the no. of days

Lets see for x[29]

No. of prev days MSE

**1 1.467765473992085e-05**

2 0.010118669101111154

3 0.010504769686189646

4 0.0007542688470444694

5 0.00010121161318663264

6 0.0019825245841761075

**7 3.155443620884047e-28**

8 2.8104402343662953e-28

**9 9.98402083170343e-31**

10 3.993608332681372e-30

11 6.520428419717426e-30

12 1.1845239529959255e-29

13 3.0814879110195774e-29

14 4.496630419276208e-28

15 5.320484701214758e-27

16 5.919525951116964e-27

17 1.9760985397308976e-25

**18 1.1093356479670479e-31**

19 3.0814879110195774e-31

20 1.4148397828842521e-25

Taking 18 days data gives highest accuracy

While generally, the maxima of accuracy is obtained at last 9 days.