Stock Price Prediction Using Twitter Sentiment Analysis

Group 16
Yuqing Jin, Hao Zeng, Shuai Ren

Body of the project

1. Problem Statement
2. Sentiment Analysis & Score
3. LSTM & Linear Regression
4. Experimental results
5. Reference

Problem Statement

Problem Statement

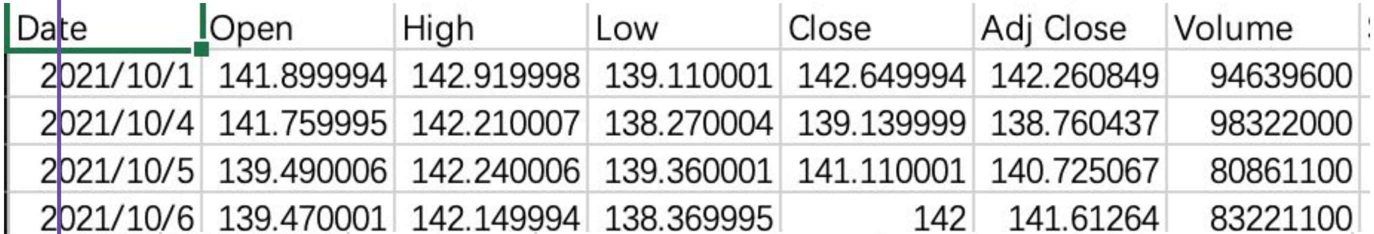
Background: Stock market prediction has been an active area of research for a long time. The Efficient Market Hypothesis (EMH) states that stock market prices are largely driven by new information and follow a random walk pattern.

Objective of project: to predict stock price of Apple Technology company during covid-19 period (2020-2021) using historical data and twitter sentiment analysis.

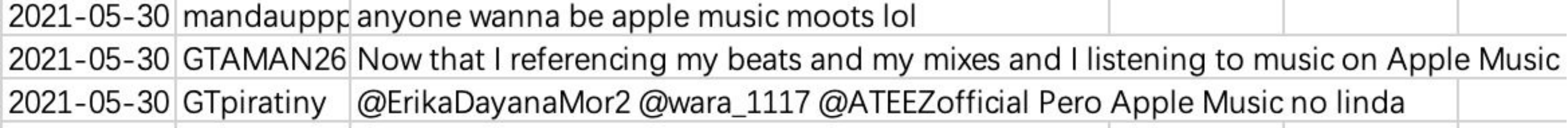
Data Sources

Historical Stock Price: Yahoo Finance API

Twitter Sentiment Analysis: Twitter API



Date	Open	High	Low	Close	Adj Close	Volume
2021/10/1	141.899994	142.919998	139.110001	142.649994	142.260849	94639600
2021/10/4	141.759995	142.210007	138.270004	139.139999	138.760437	98322000
2021/10/5	139.490006	142.240006	139.360001	141.110001	140.725067	80861100
2021/10/6	139.470001	142.149994	138.369995	142	141.61264	83221100
2021/10						

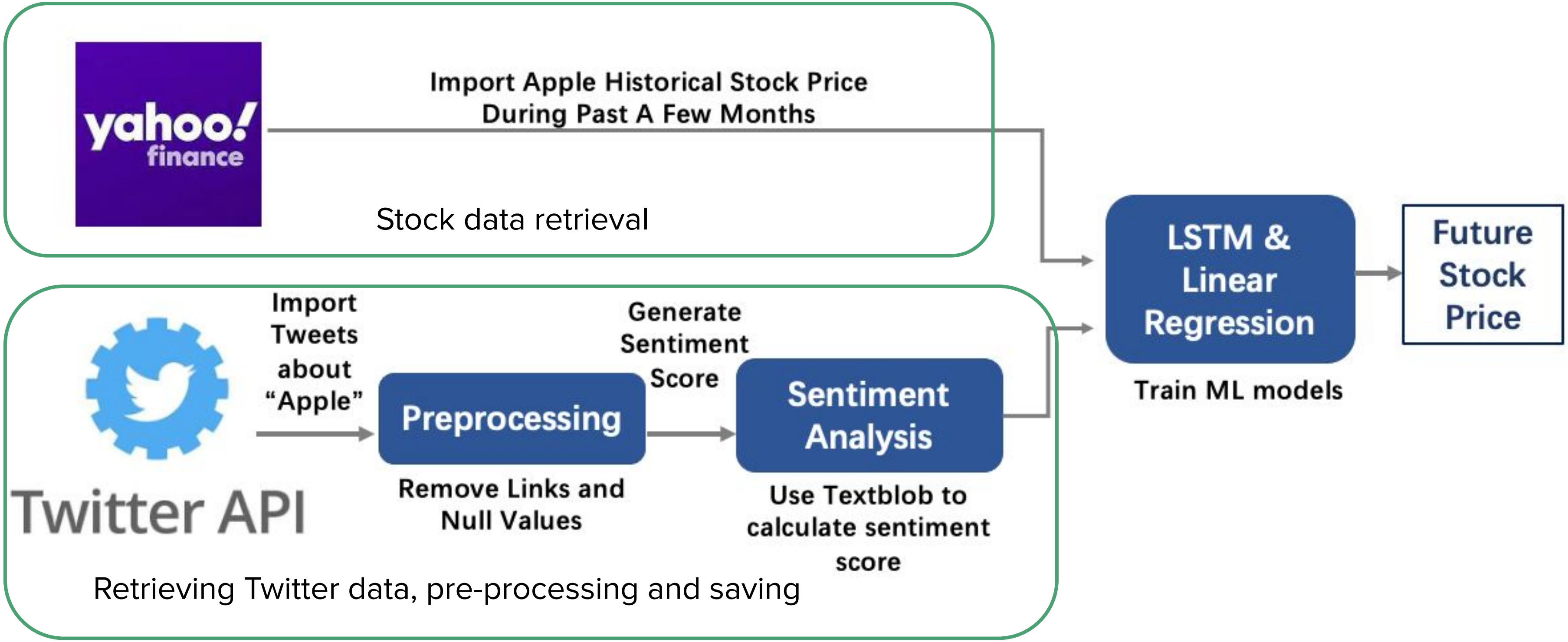


2021-05-30	mandauppp	anyone wanna be apple music moots lol
2021-05-30	GTAMAN26	Now that I referencing my beats and my mixes and I listening to music on Apple Music
2021-05-30	GTpiratiny	@ErikaDayanaMor2 @wara_1117 @ATEEZofficial Pero Apple Music no linda

We discovered the relationship of historical price and twitter sentiment score through 3/6/12/24 months data.

Also, we trained models using LSTM and Linear Regression(spark library) models.

Overall process



yahoo! finance
Import Apple Historical Stock Price During Past A Few Months
Stock data retrieval
Twitter API
Import Tweets about "Apple"
Preprocessing
Remove Links and Null Values
Generate Sentiment Score
Sentiment Analysis
Use Textblob to calculate sentiment score
Retrieving Twitter data, pre-processing and saving
LSTM & Linear Regression
Train ML models
Future Stock Price

Sentiment Analysis

Overview of Sentiment Analysis Process

1. Data Preprocessing & Cleaning

   Remove links, user id, null values, etc.

   Tokenize tweet content

2. Sentiment Analysis Score Computation

   Use **textblob** package to compute polarity

3. Combine sentiment score with historical stock price via **Structured Streaming**

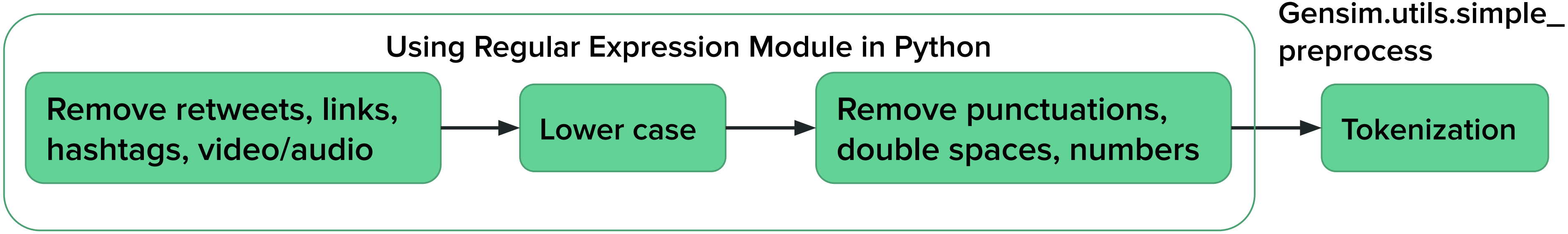
   Groupby data of each day and sum the score,

   combine score with the historical stock price according to the date

Data Preprocessing & Sentiment Analysis Score

1. Data Preprocessing

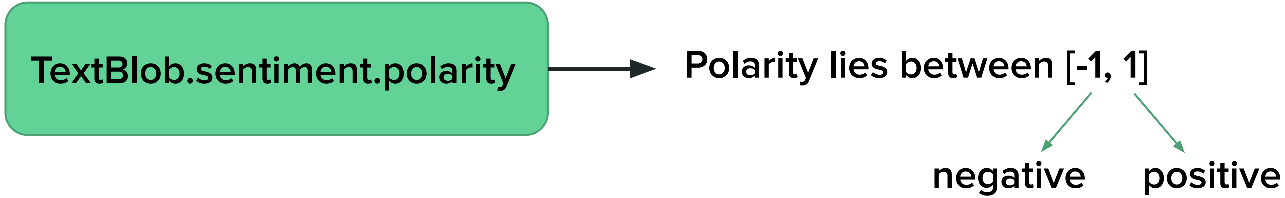
WordNetLemmatizer()
Gensim.utils.simple_
preprocess



Using Regular Expression Module in Python

Remove retweets, links, hashtags, video/audio → Lower case → Remove punctuations, double spaces, numbers → Tokenization

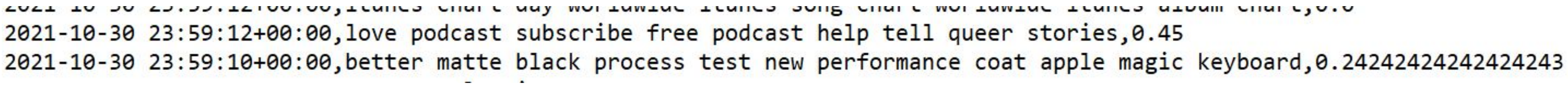
2. Sentiment Analysis Score Calculation



TextBlob.sentiment.polarity → Polarity lies between [-1, 1]

negative

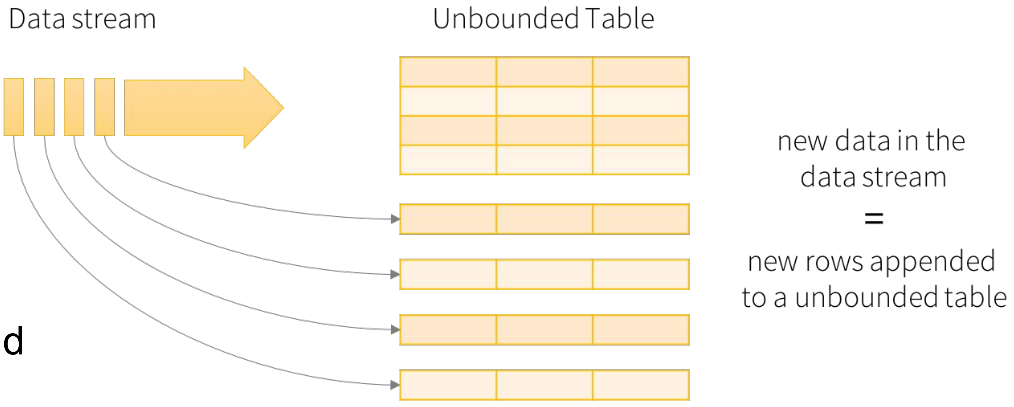
positive



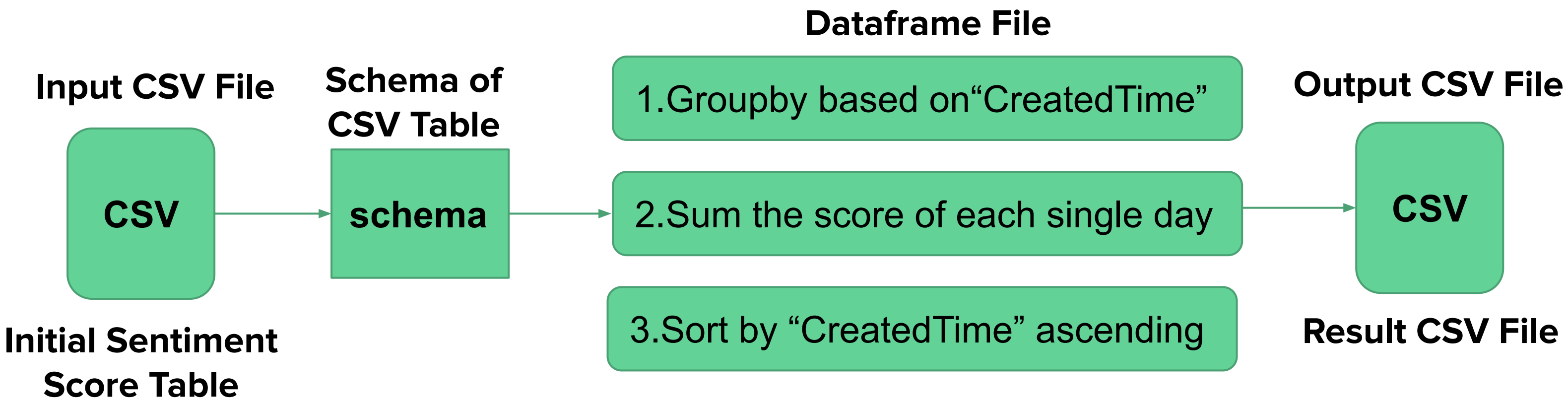
```
2021-10-30 23:59:12+00:00,itunes chart day worldwide itunes song chart worldwide itunes album chart,0.0
2021-10-30 23:59:12+00:00,love podcast subscribe free podcast help tell queer stories,0.45
2021-10-30 23:59:10+00:00,better matte black process test new performance coat apple magic keyboard,0.24242424242424243
```

3.Structured Streaming

The processed data is appended to the continuously flowing data stream.
Each row of the data stream is processed and the result is updated into the unbounded result table.

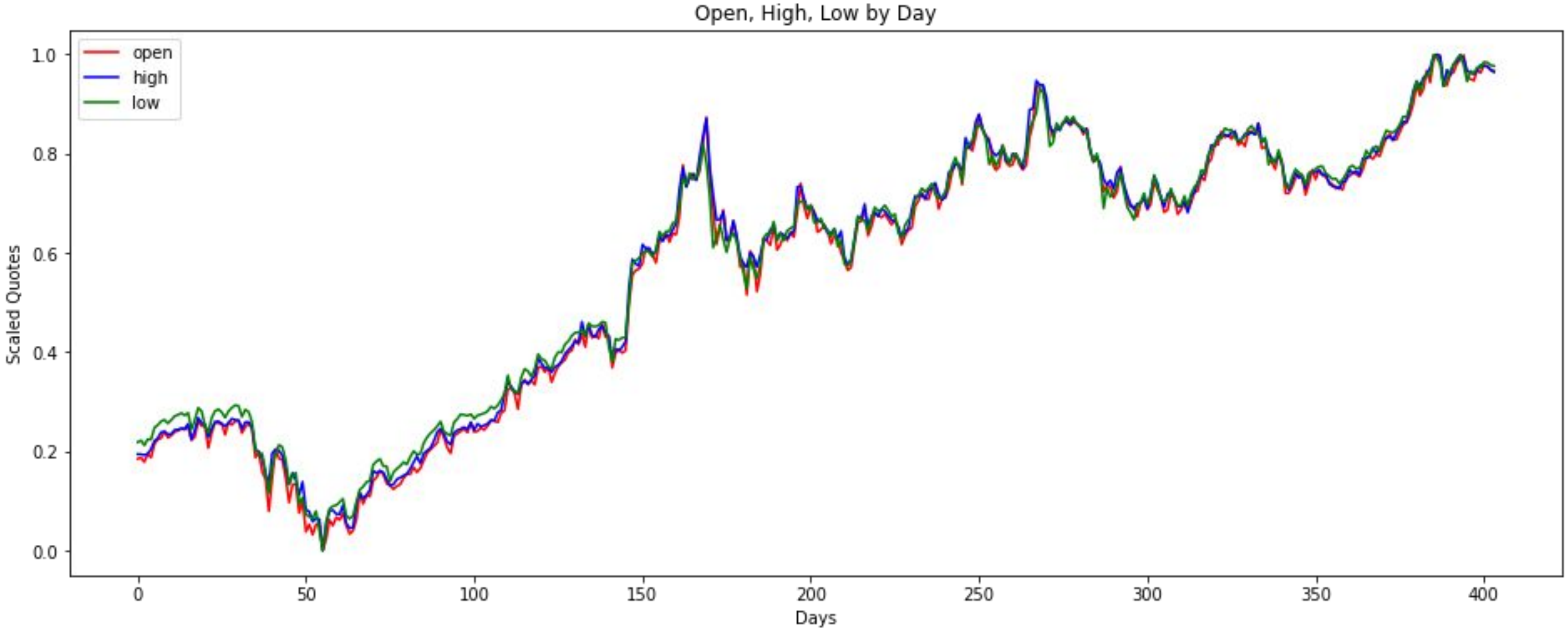

Data stream
Unbounded Table
new data in the data stream
=
new rows appended to a unbounded table

Data stream as an unbounded table


Input CSV File
CSV
Initial Sentiment Score Table
Schema of CSV Table
schema
Dataframe File
1.Groupby based on“CreatedTime”
2.Sum the score of each single day
3.Sort by “CreatedTime” ascending
Output CSV File
CSV
Result CSV File

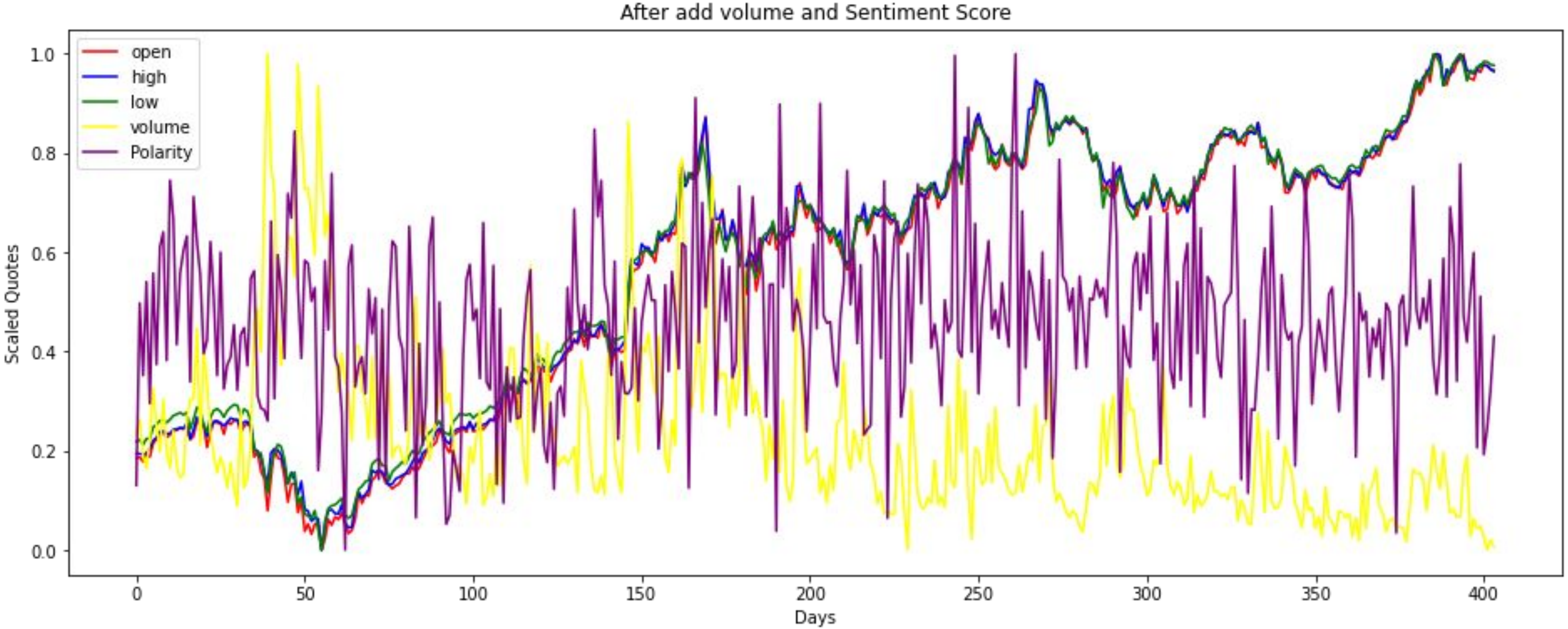
LSTM & Linear Regression

The overall stock price trendline of past 2 years



Stock price of Apple Company in past 2 years. The open price, high and low of each day are shown on the graph.

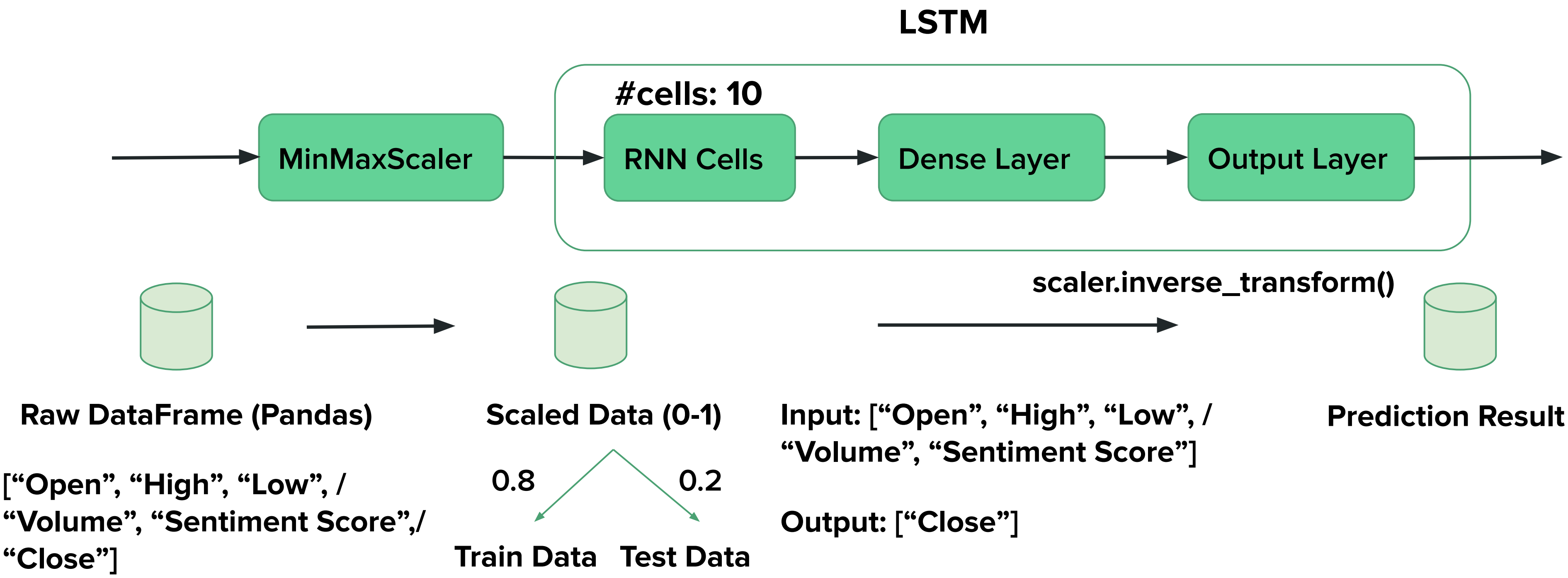
Add sentiment score and volume into the graph



Before training, twitter sentiment score is noisy and have a large standard deviation.
It is difficult to predict stock price by using polarity as single input.

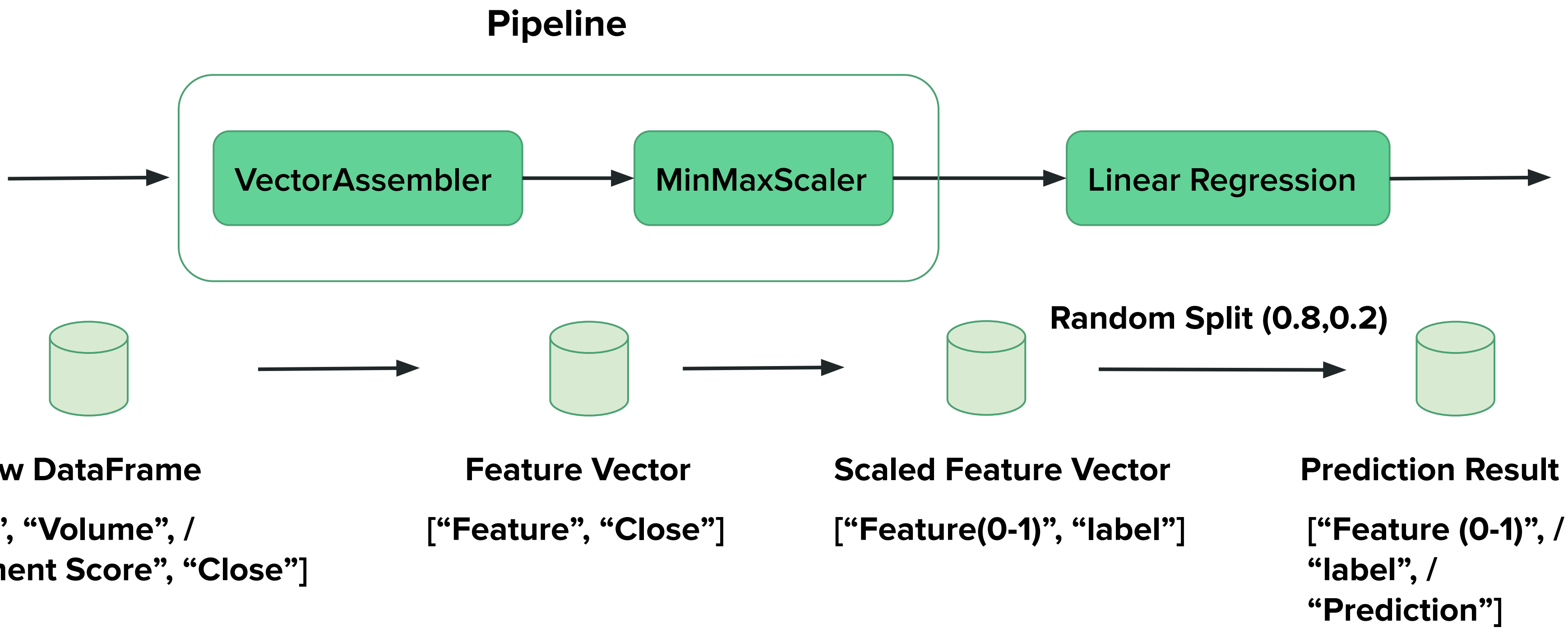
LSTM

We used tensorflow keras package to implement the LSTM structure.



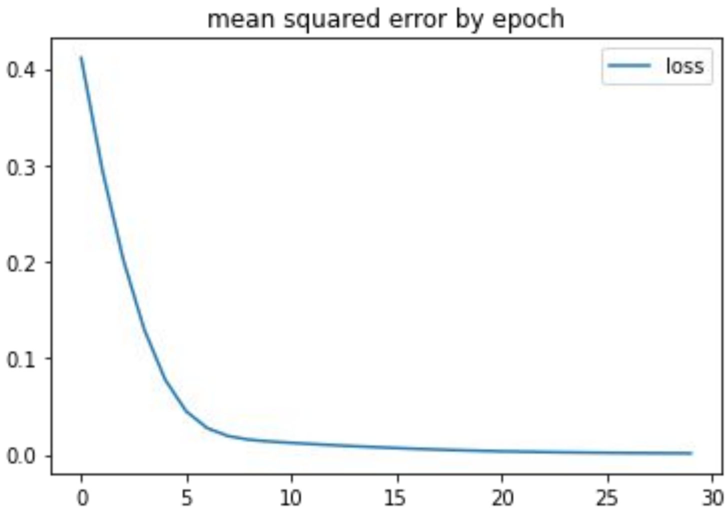
Linear Regression

We used spark ml package to implement the linear regression model



LSTM Results

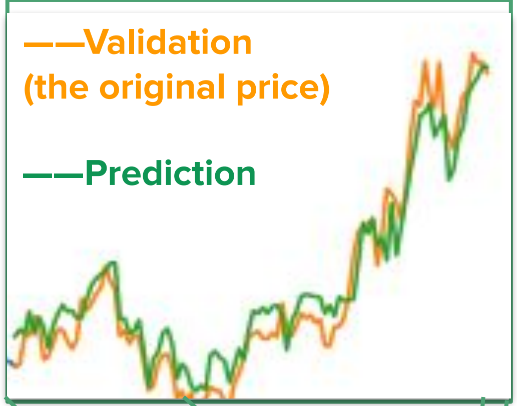
24 months data prediction result

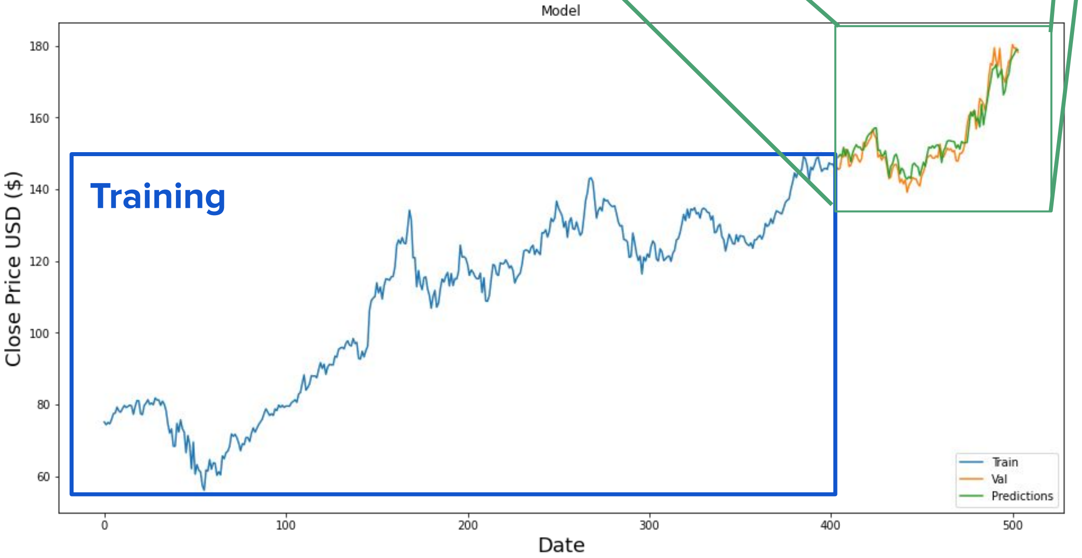


The training loss is stable at around 0.001 after 25 epoch

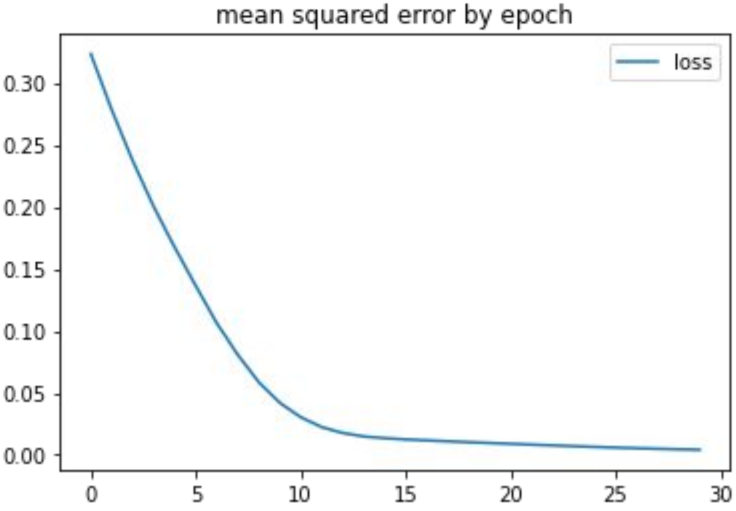
Validation root mean squared Error (RMSE): 0.4183

The prediction of past 2 years is quite similar as the original curve.





12 months data prediction result



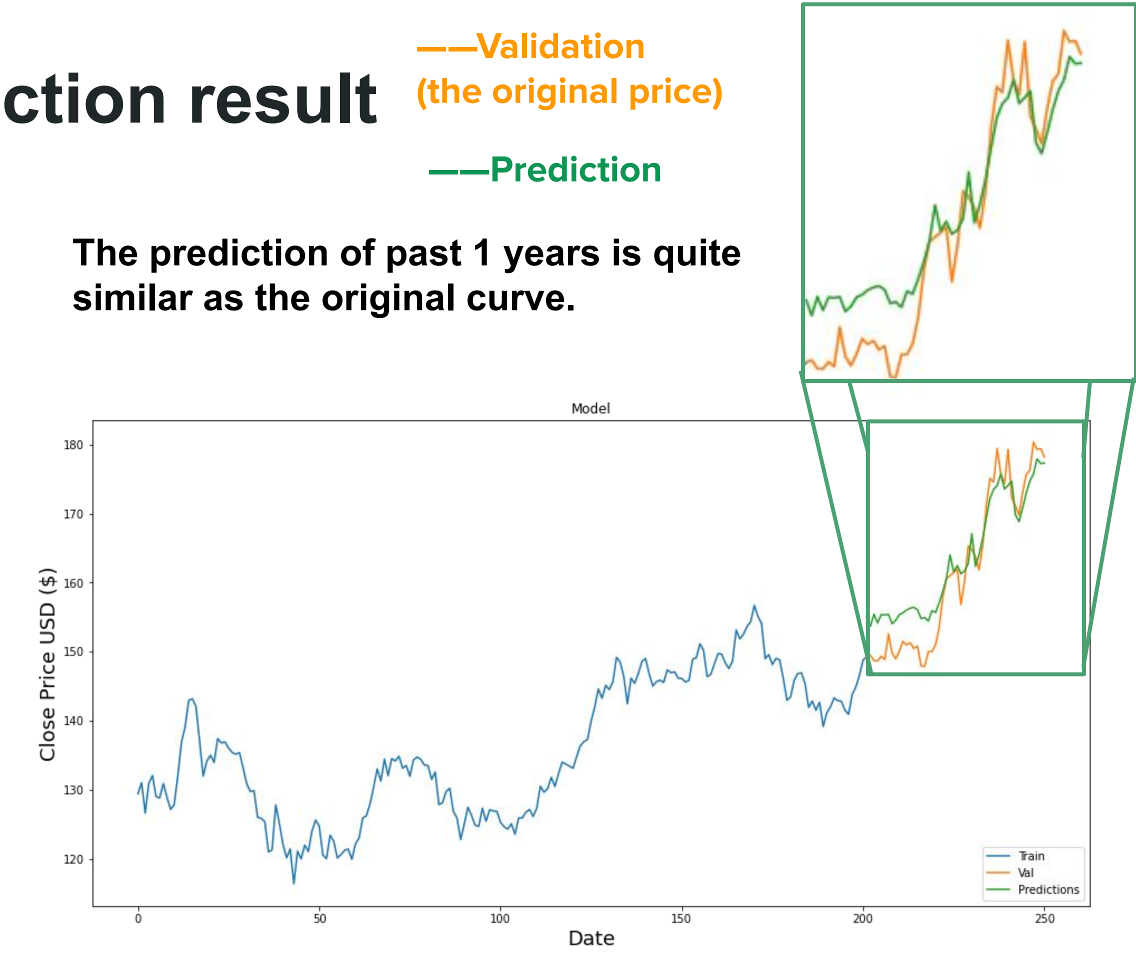
The training root mean squared is stable at around 0.001 after 28 epoch

Validation root mean squared Error (RMSE): 0.4185

——Validation (the original price)

——Prediction

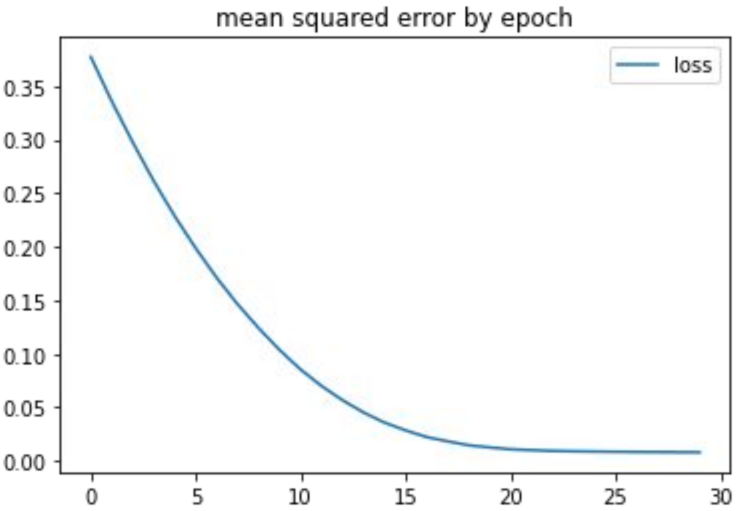
The prediction of past 1 years is quite similar as the original curve.



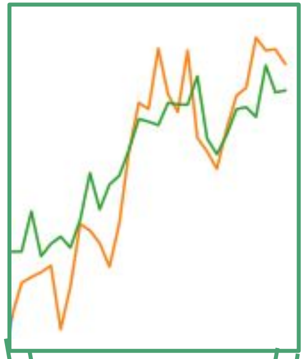
6 months data prediction result

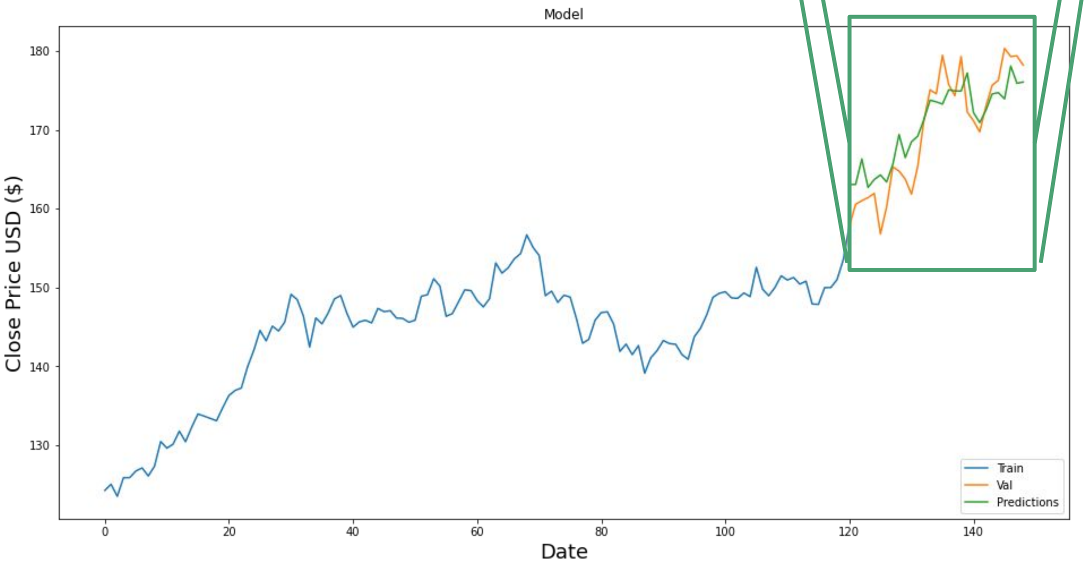
——Validation (the original price)

——Prediction



The predicted curve has the general trend of the original one, but less accurate than the 2-year model.

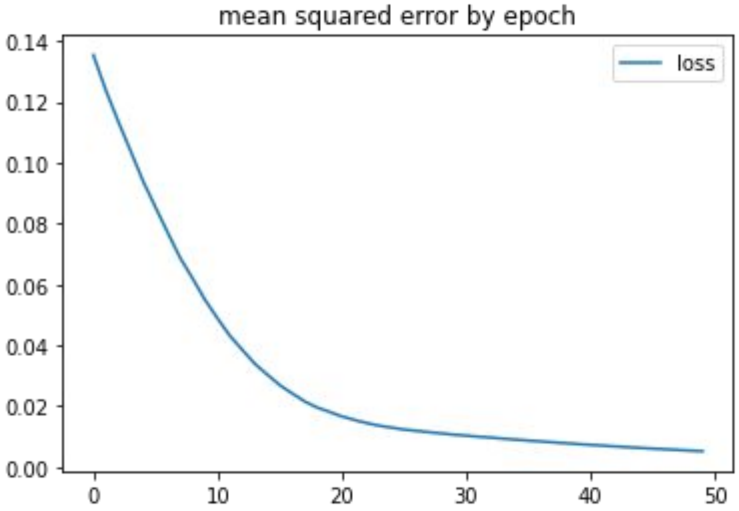




The training root mean squared is stable at around 0.001 after 30 epoch

Validation root mean squared Error (RMSE): 0.5876

3 months data prediction result

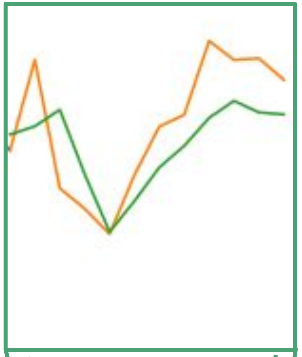


The training root mean squared is stable at around 0.005 after 45 epoch

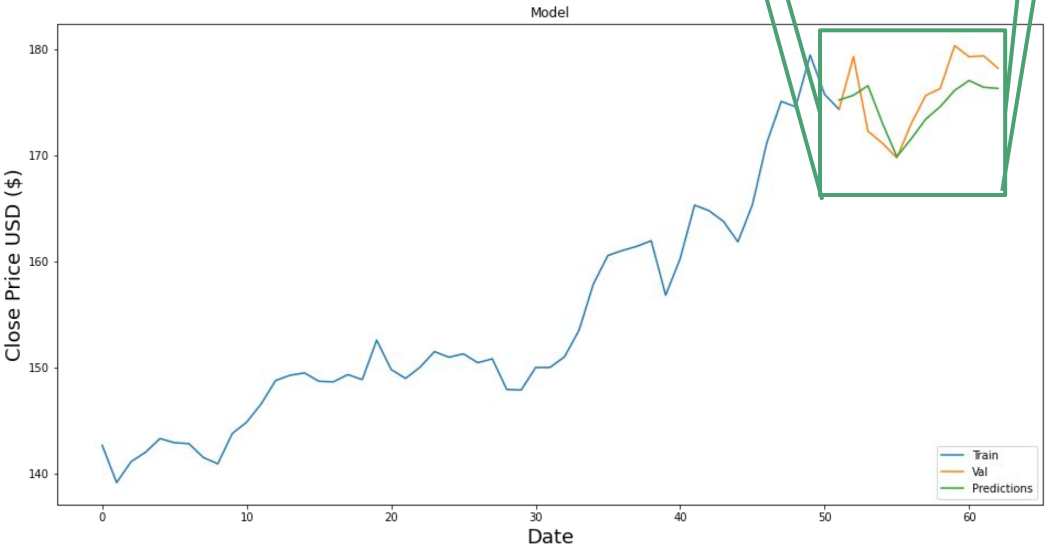
Validation root mean squared Error (RMSE): 0.6048

——Validation (the original price)

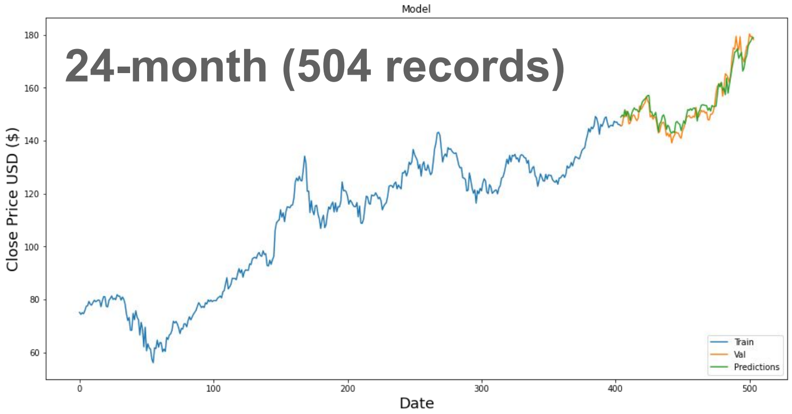
——Prediction



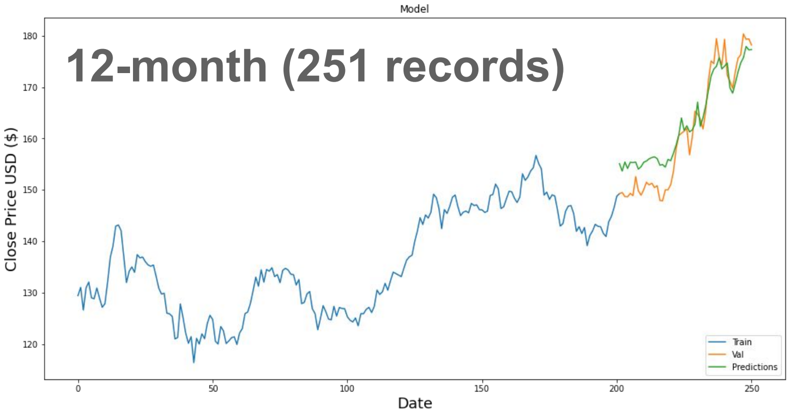
The 3-month result is less likely to predict an exact same value as the real one, but the general trend is almost the same.



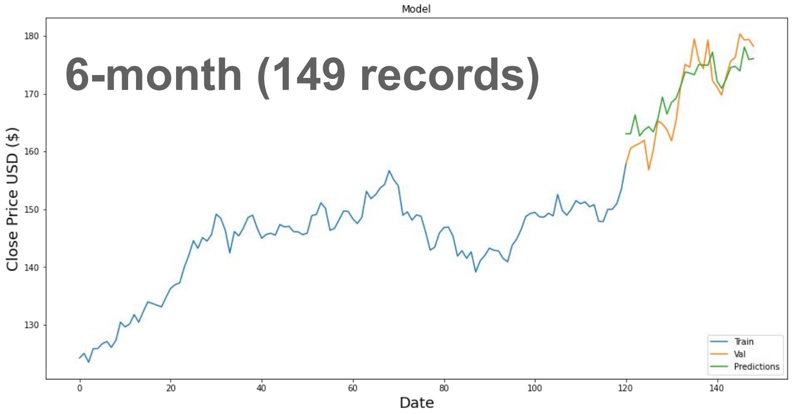
Comparison Between 3/6/12/24 months Prediction



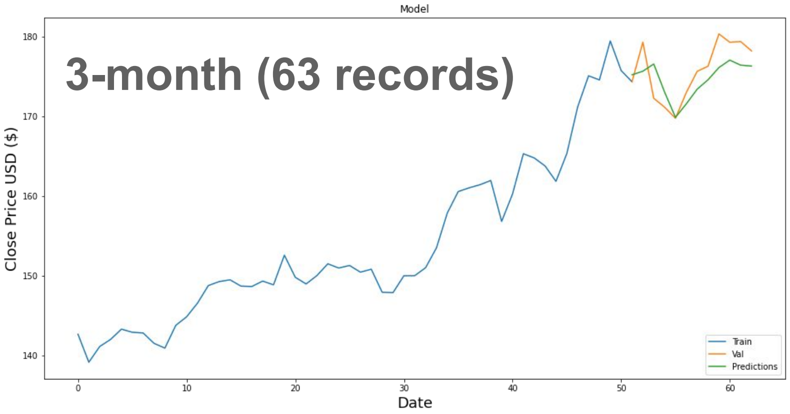
24-month (504 records)



12-month (251 records)



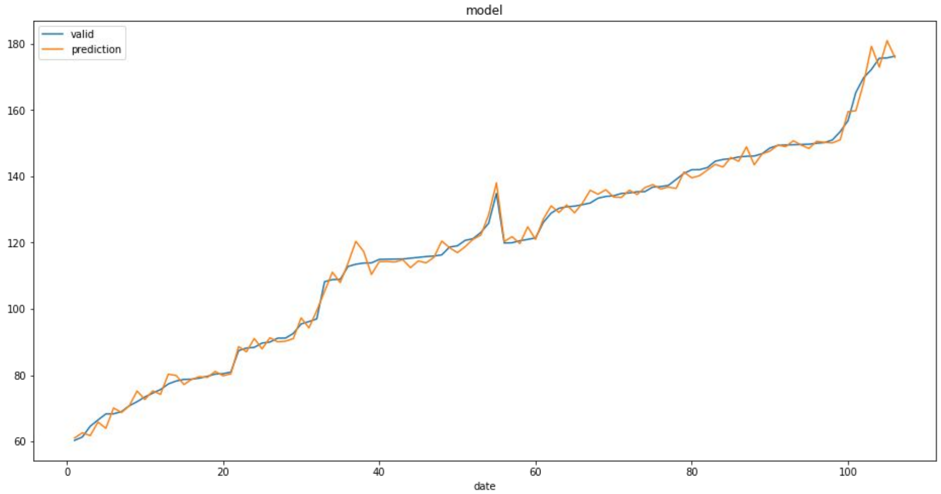
6-month (149 records)

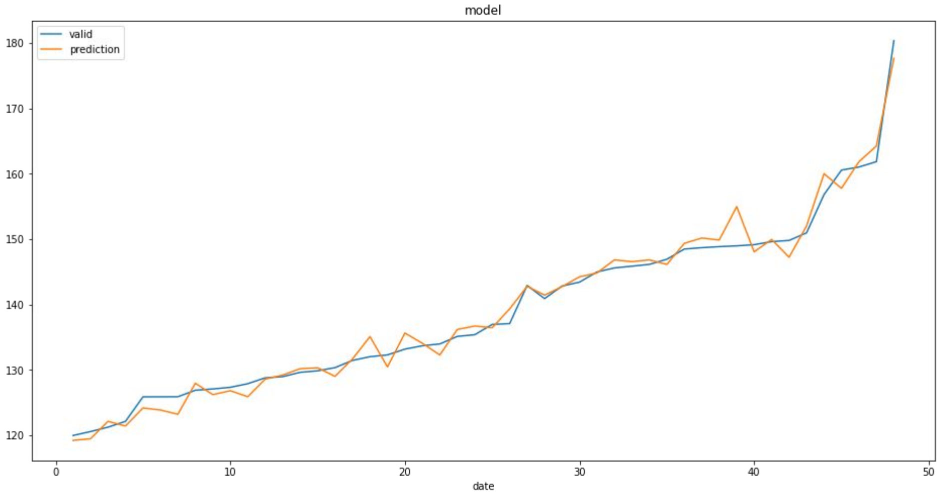


3-month (63 records)

Linear Regression Results

24 & 12 months-data prediction result



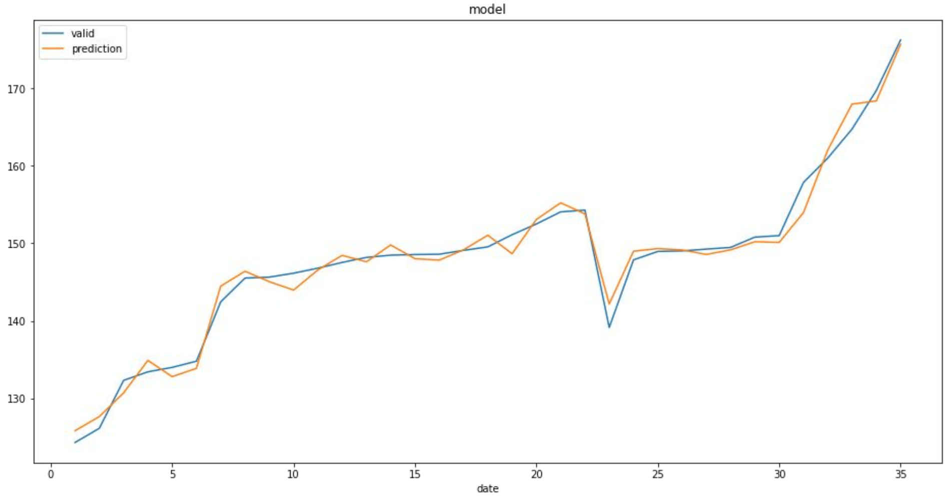


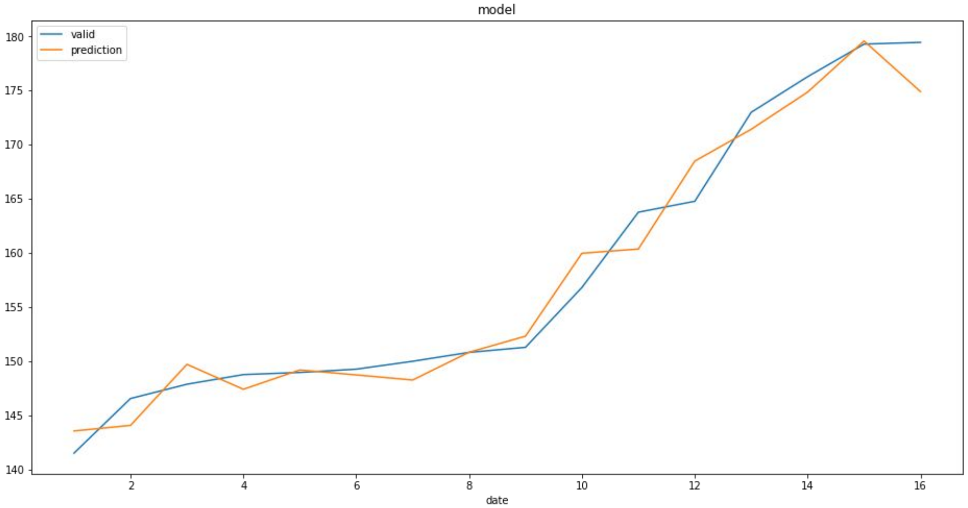
Validation RMSE：0.4078

Validation RMSE：0.4711

The predictions of past 2 years and 1 year are quite similar as the original curve.

6 & 3 months-data prediction result



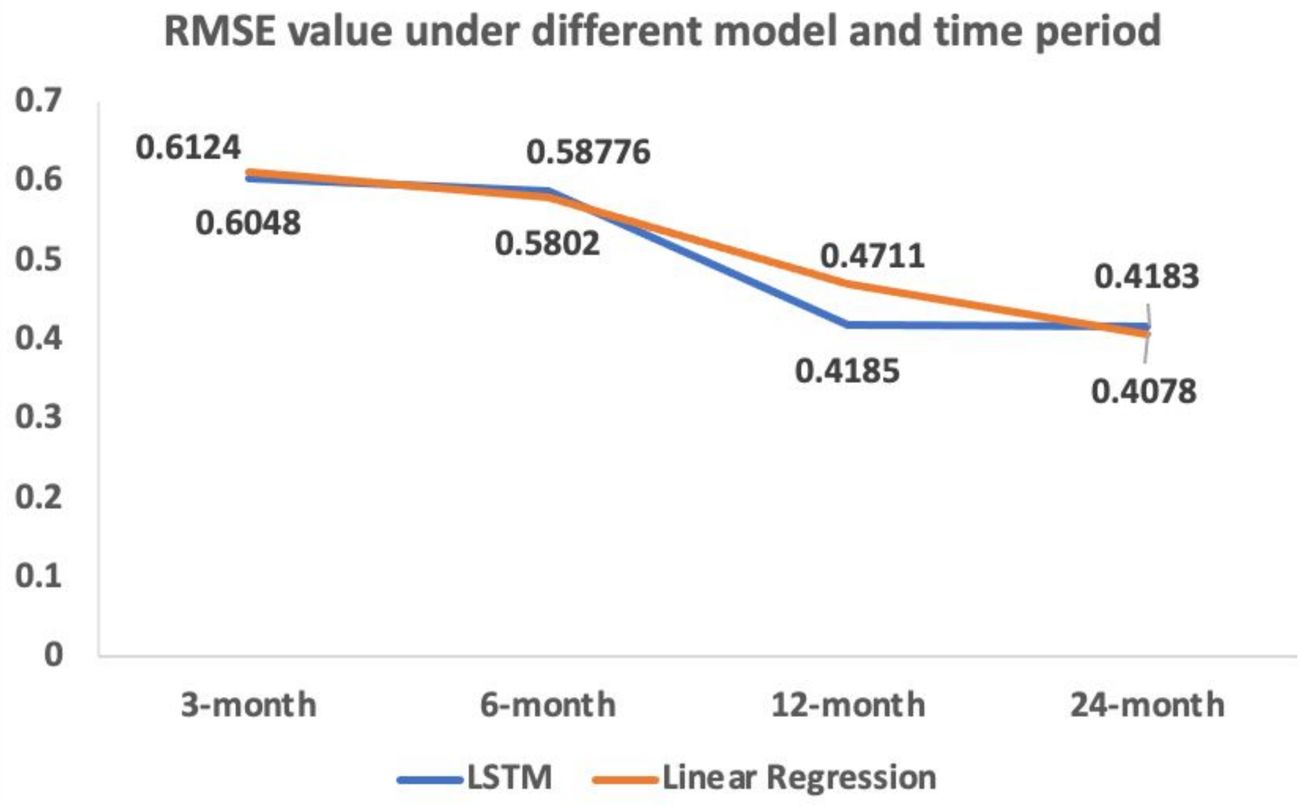


Validation RMSE：0.5802

Validation RMSE：0.6124

The predictions of past 6 months and 3 months are a little bit far from the original curve, but the trend is same.

Comparison between 3/6/12/24 months prediction



Related work

Reference

1. Kalyani, Joshi, Prof Bharathi, and Prof Jyothi. "Stock trend prediction using news sentiment analysis." *arXiv preprint arXiv:1607.01958* (2016).
2. Mehta, Pooja, Sharnil Pandya, and Ketan Kotecha. "Harvesting social media sentiment analysis to enhance stock market prediction using deep learning." *PeerJ Computer Science* 7 (2021): e476.
3. Xianya, Jiang, Hai Mo, and Li Haifeng. “Stock Classification Prediction Based on Spark.” *Procedia Computer Science* 162 (2019): 243–50. https://doi.org/10.1016/j.procs.2019.11.281.
4. Tiwari, Shweta, and Alka Gulati. "Prediction of Stock Market from Stream Data Time Series Pattern using Neural Network and Decision Tree 1." (2011).
5. Mittal, Anshul, and Arpit Goel. "Stock prediction using twitter sentiment analysis." Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf) 15 (2012): 2352.
6. Vu, Tien Thanh, et al. "An experiment in integrating sentiment features for tech stock prediction in twitter." Proceedings of the workshop on information extraction and entity analytics on social media data. 2012.

Thank you!