Stock Price Prediction with Media Sentiment

sc5124 Shengqi Cao

jh4312 Jingchao Hu

wx2283 Wenshuo Xie

Goal & Novelty

Goal:

1. Our goal is to create a model that provides accurate and reliable predictions about stock prices to assist with investment decision-making.
2. Build a model that can predict stock prices using sentiment analysis on Twitter data.
3. Estimate the “market sentiment” based on the “public sentiment” and then predict the stock trend.

Novelty:

Our novelty is connecting the public sentiment with the stock price, and the use of different techniques to improve the accuracy and interpretability of the model.

Data

Tweets about the Top Companies from 2015 to 2020

- Volume: Over 3 million rows of unique tweets data
- Velocity: We performed a one-time scraping to collect the data, it contains tweets about top companies from 2015 to 2020
- Variety: 7 columns, including tweet id, author of the tweet, post date, the text body of the tweet, and the number of comments, likes, and retweets

Values of Top NASDAQ Companies from 2010 to 2020

- Volume: About 17,500 rows of stock price data
- Velocity: We performed a one-time scraping to collect the data, it contains stock price data of top companies from 2010 to 2020
- Variety: 7 columns, including ticker symbol, day date, close value, volume, open value, high value, low value

Twitter7 dataset(from Stanford Large Network Dataset Collection)

- Volume: 467 million Twitter posts from 20 million users
- Velocity: It's an existing dataset. It contains data covering a 7 month period from June 1 2009 to December 31 2009
- Variety: 7 columns, including tweet id, author of the tweet, post date, the text body of the tweet, and the number of comments, likes, and retweets

Historical stock price dataset

- Volume: About 15,000 rows of stock price data
- Velocity: It's an existing dataset. It contains stock price data of top companies in 2009
- Variety: 7 columns, including ticker symbol, day date, close value, volume, open value, high value, low value

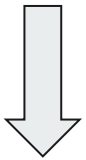
Methodology

1. **Data Pre-processing:**
   Load the data csv file into the Google Cloud storage bucket, and use Pyspark to read it, then we parse and filter the data.

1. **Sentiment Analysis:**
   Used the **vader_lexicon** in the NLTK package in Python to conduct sentiment analysis on the twitter data, obtaining the polarity of the tweets for each day.

tweet_id	writer	post_date	body	comment_num	retweet_num	like_num
550441509175443456	VisualStockRSRC	1420070457	lx21 made $10,008...	0	0	1
550441672312512512	KeralaGuy77	1420070496	Insanity of today...	0	0	0
550441732014223360	DozenStocks	1420070510	S&P100 #Stocks Pe...	0	0	0
550442977802207232	ShowDreamCar	1420070807	$GM $TSLA: Volksw...	0	0	1
550443807834402816	i_Know_First	1420071005	Swing Trading: Up...	0	0	1

ticker_symbol	day_date	close_value	volume	open_value	high_value	low_value
AAPL	2020-05-29	317.94	38399530	319.25	321.15	316.47
AAPL	2020-05-28	318.25	33449100	316.77	323.44	315.63
AAPL	2020-05-27	318.11	28236270	316.14	318.71	313.09
AAPL	2020-05-26	316.73	31380450	323.5	324.24	316.5



	Date	Tweet	Close
0	2015-01-08	\#Microsoft : \"\"Transparent Failover\"\" in ...	47.590
1	2015-01-13	#Dow #stocks $MSFT Microsoft Daily:-1.25% Wee...	46.355
2	2015-01-19	#Microsoft : Synect Create World's Largest Pho...	46.240
3	2015-01-21	How to make a 6% yield with Microsoft shares h...	45.920
4	2015-01-21	Get your #Windows10 fix here with @JoannaStern...	45.920
...	...	...	...

Methodology

3. **Model Training and evaluation:**
Trained and evaluated machine learning models (linear regression, random forest, and multilayer perceptron) on the processed data

	Date	Comp	Negative	Neutral	Positive	Prices	Prices2	Prices3	Prices4
80	20090830	-1.0	0.120	0.764	0.116	24.680000	24.68	24.68	24.690001
81	20090831	1.0	0.075	0.792	0.133	24.650000	24.68	24.68	24.68
82	20090901	1.0	0.057	0.787	0.156	24.000000	24.65	24.68	24.68
83	20090902	1.0	0.066	0.804	0.130	23.860001	24.0	24.65	24.68
84	20090903	1.0	0.067	0.789	0.144	24.110001	23.860001	24.0	24.65
...	...	...	...	...	...	...	...	...	...
160	20091125	1.0	0.053	0.797	0.150	29.790001	29.940001	29.620001	29.620001
161	20091126	1.0	0.042	0.812	0.145	29.790001	29.790001	29.940001	29.620001
162	20091128	1.0	0.056	0.831	0.113	29.219999	29.790001	29.790001	29.940001
163	20091129	1.0	0.061	0.801	0.138	29.219999	29.219999	29.790001	29.790001
164	20091130	1.0	0.065	0.797	0.138	29.410000	29.219999	29.219999	29.790001

Training Sets

Algorithm

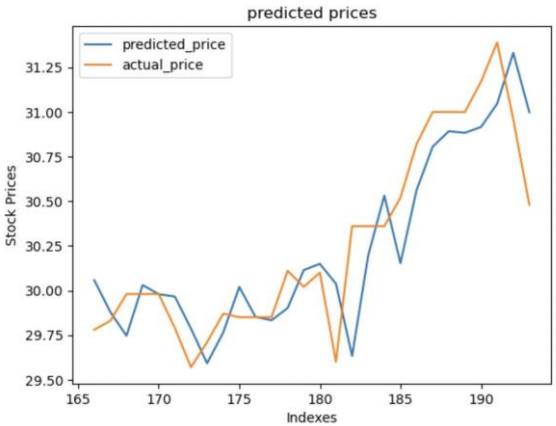
1. **Linear regression:**
   Modeled the relationship between sentiment data from twitter and stock prices.

1. **Random forest:**
   Used an ensemble learning method to predict stock prices based on twitter sentiment data

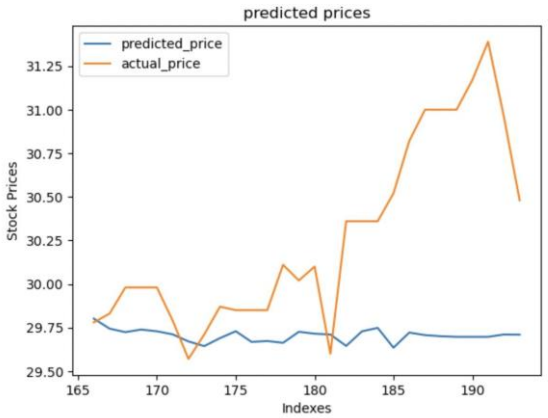
1. **Multilayer perceptron (MLP)**
   Used an artificial neural network to predict stock prices based on twitter sentiment data

Test Result

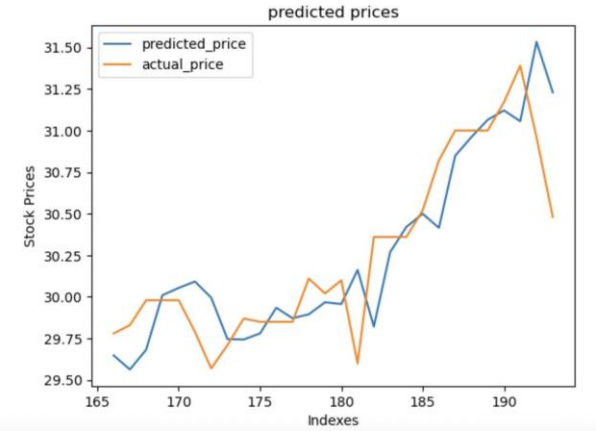
Tweet7 dataset



Linear Regression



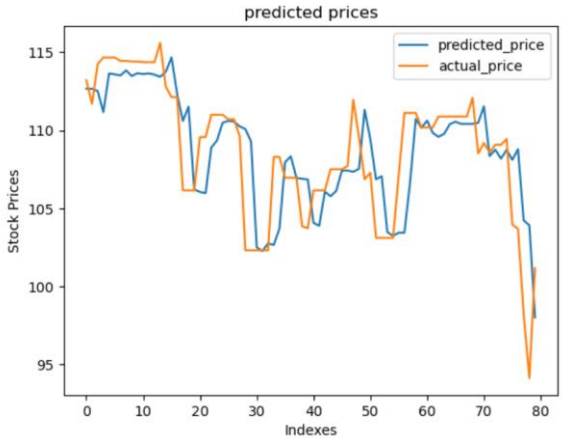
Random Forest



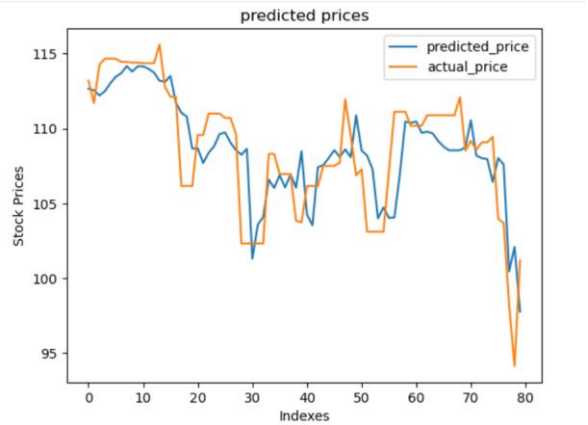
Multilayer Perceptron

Test Result

New dataset

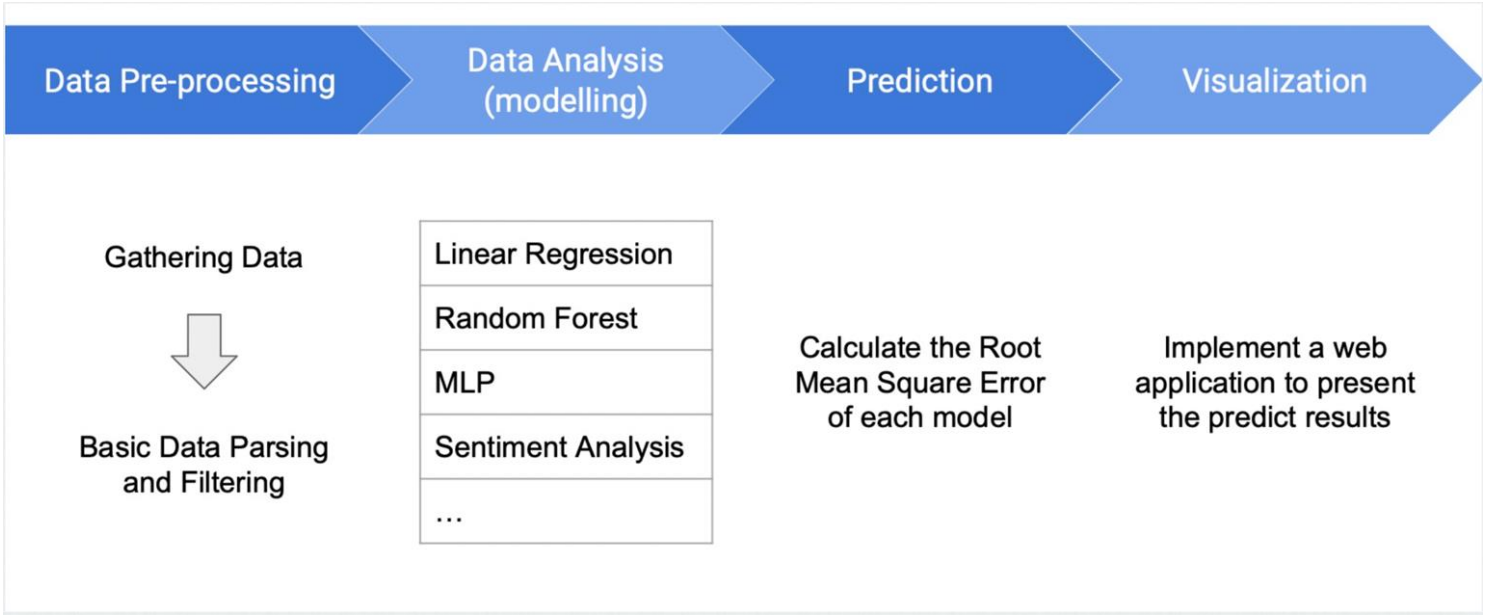


Linear Regression



Multilayer Perceptron

System


Data Pre-processing
Data Analysis (modelling)
Prediction
Visualization
Gathering Data
Basic Data Parsing and Filtering
Linear Regression
Random Forest
MLP
Sentiment Analysis
…
Calculate the Root Mean Square Error of each model
Implement a web application to present the predict results

Website Demo