Daily News for Stock Market Prediction

王威斌 0316081 薛世恩 0316323 楊光傑 0316104 陳子軒 0316103

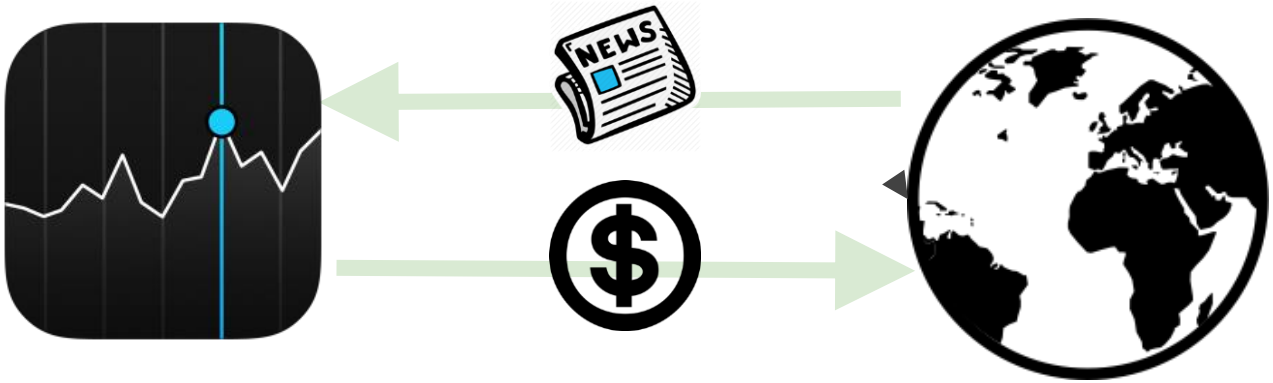
Introduction

Stock market is a sensitive and bumping market.

Any of news may influence or impact the stock price.

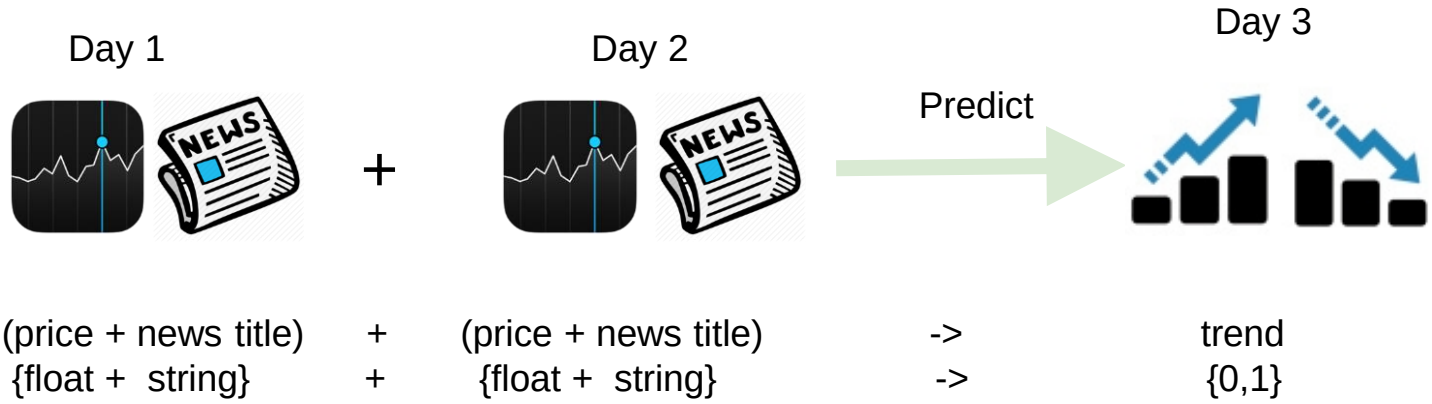
The economy is directly proportional to the stock market.

stock price drop, economy cool down.



Problem definition or formulation

What is the next day trend of the stock price by using two days news titles and prices?



Challenges or importance of this work

IMPORTANCE

- Successful prediction of a stock's future price could yield significant profit and prevent your loss

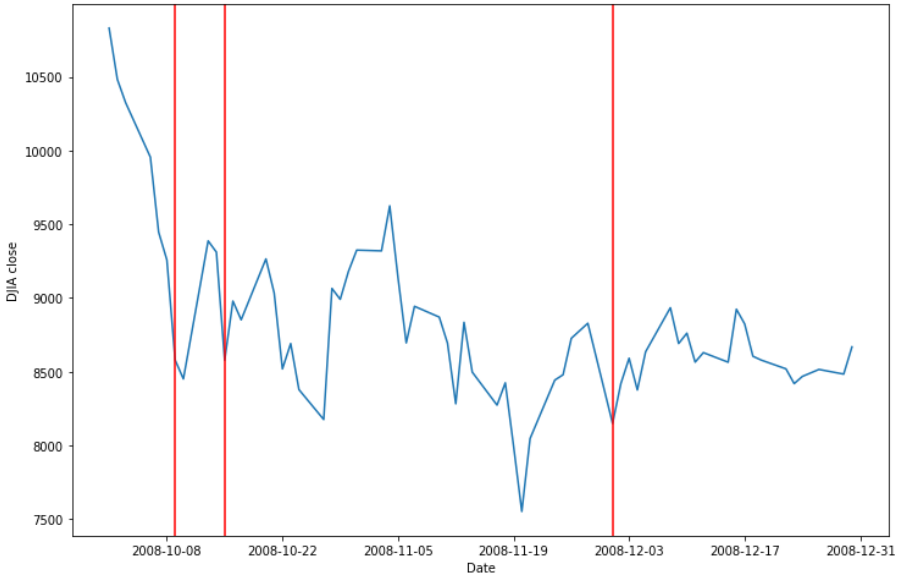
CHALLENGES

- The Future Might Not Be Like the Past, stocks will not perform in the future as they have in the past.
- Data Set Is Limited.
- Too much factors that affect the stock market:
    - -Internal Developments(a new innovative product, the hiring or firing of company executives …)
    - -World Events(war, natural disasters, terrorism...)

Global financial crisis in October 2008

On 15 October 2008, DJIA fell 733 points (0.078 % drop)

Date	Adj Close	diff	percent_change
2008-10-15	8577.910156	-733.080078	-0.078733
2008-12-01	8149.089844	-679.950195	-0.077013
2008-10-09	8579.190430	-678.909179	-0.073331



Past two days's news headlines from Reddit WorldNews Channel - 13/10/2008

Top1	EU Bans the Incandescent Light Bulb
Top2	AFP: Paul Krugman wins Nobel Economics Prize
Top3	Europe to U.S.: You messed up the rescue, too
Top4	Hindu Threat to Christians: Convert or Flee
Top5	Europe puts $2.3 trillion on line for banks, almost 3 times the USA bailout
Top6	Congratulations on your Nobel Prize in Economics Paul Krugman!
Top7	When can we get some damn privacy? Governments lose data! Up to 1.7m people's data missing
Top8	AFRICA: Sexually-transmitted grades kills quality education
Top9	New Beijing Traffic Laws Take 800,000 Cars Off the Road in China
Top10	A Jamaican lottery scam draws in millions of US dollars every day, providing Jamaican gangs with high-calibre weapons
Top11	Terror bill: 42-day detention rejected
Top12	Ringo 'too busy' for autographs
Top13	EU warns youth: turn your MP3 players down!
Top14	For three decades the Free Aceh Movement fought for independence from Indonesia, settling finally for autonomy. Now its founder, long in exile,
Top15	North Korea restores U.N. monitoring of atom site-diplomats
Top16	Aids in Africa: The power of the pulpit's message
Top17	AskReddit: Out of all the significant world events that you have experienced in your lifetime, where would you rank the current financial crisis ?
Top18	A different spin on textbook controversy: rewriting history in the interest of peace
Top19	The Soleckshaw, a new solar powered rickshaw, has been unveiled this month in Delhi, and is being touted as a solution to traffic jams, pollution
Top20	Venezuela shuts down McDonald's
Top21	Children of the black dust
Top22	UK banks receive 37bn bail-out
Top23	Bank shares fall despite bail-out
Top24	Fighting the Financial Crisis: Stocks Surge As EU Nations Unveil Bailout Packages
Top25	EU to ban traditional light bulbs, despite dangers of CFLs to health and the environment

Past two days's news headlines from Reddit WorldNews Channel - 14/10/2008

Top1	Russian Lawyer Who Defended Journalists Is Poisoned In France
Top2	Daughter of Mossad chief sent to military prison for refusing to enlist. Stay classy, Israel!
Top3	US Surrenders Power to Appoint World Bank President
Top4	India gets a new view of US: Collection agents at call centers hear tales of woe from a land whose lifestyles they once idealized.
Top5	I was tortured and sentenced to death in Saudi Arabia. As a westerner, I was eventually released, but others are not so lucky
Top6	Redditors in Canada: don't forget to get out and vote today!
Top7	Friendly-Fire Cover Up Revealed by New Video
Top8	Farrakhan To Announce 'A New Beginning' Oct. 19th - Nation of Islam leader to make major statement at mosque dedication
Top9	How Did GOP Get $8 Million from Wachovia? (Was It Before or After They Were 'Bailed Out?'... If That Even Matters)
Top10	N. Korea defectors drop leaflets condemning leader
Top11	Protests in London Against the Financial Bailout Plan for Banks (videos)
Top12	Syria, Lebanon establish diplomatic ties
Top13	PDF: Amnesty International reports a sharp increase in beheadings in Saudi Arabia, with migrants and the poor bearing the brunt
Top14	UK "Shariah TV" Episode 5 asks "How can you be a Muslim in the Land of the Great Satan?"
Top15	Some of my best friends are gay: Pete Mullen apologizes and explains his suggestion that homosexuals have "sodomy warnings" tattooed on their bodies.
Top16	Unexploded Bombs in Germany: The Lethal Legacy of World War II
Top17	Global Economic Crisis Likely To Have Profound Consequences For US Politics, World Relations
Top18	Among the 64 Israelis arrested in the wake of Acre's 4 days of riots, the Arab man who started it all by driving his car during Yom Kippur.
Top19	My eight-year-old daughter - who has kidney problems - was dragged by her hair and one man tried to push her through the railings on the windows.
Top20	Stephen Harper has been a terrible Prime Minister and has run perhaps the poorest campaign, despite his millions, of all the major parties
Top21	A photo-timeline of the Darfur conflict, now in its fifth year, by Doctors Without Borders
Top22	Bush Blinks on North Korean Nukes. Again.
Top23	Who knows what disasters lurk within the world's scattered lands? The Hungarians know.
Top24	Science Based Approach to HIV/AIDS Returns to South Africa
Top25	UK: Church of England Signs Pact With Radical Sunni Sect Responsible for the Taliban, Equates Terrorism With the "Excesses of Western Foreign Policy"

Related works

Listening to Chaotic Whispers: A Deep Learning Framework for News-oriented Stock Trend Prediction[Reference 1]

Problem: how to get good features from news even if it is irrelevant or a rumor.

Solution: used Chaos theory and mimicking the way humans would have react to the situation

1. “Sequential Context Dependency”
- a single news would most probably be useless or give no information alone, we need to view news sequentially and weight them together.
1. “Diverse Influence”
- in when false news or rumors temporarily affect the price but will surely change back afterwards.
1. “Effective and Efficient Learning”.
- new doesn’t always provide obvious information, thus human normally gain knowledge first.

The authors proposed a new learning framework, particularly a Hybrid Attention Network (HAN) with self-paced learning mechanism, for stock trend prediction from online news. Models used in this paper are RNN, LSTM which provides training data with previous data and memory.

Related works

Using Burstiness to Improve Clustering of Topics in News Streams [Reference 2]
Qi He, Kuiyu Chang, Ee-Peng Lim (ICDM.2007.17)

Problem : a new clustering method for word vector.

Solution : Bursty Vector Space Models (B-VSM).

Find optimal K of clusters by entropy.

We want to make entrophy as small

as possible.

clustering/method		cluster entropy	class entropy	cluster purity
K-means	TFIDF	0.3233	0.3667	0.6970
	TFIDF-DF	0.2362	0.3373	0.7966
	SAB	0.3071	0.3676	0.7183
	BAB	0.2831	0.3468	0.7405
	SMB	0.2078	0.3023	0.8094
	BMB	0.1784	0.2587	0.8449
	BT	0.2456	0.3167	0.7791

Dataset Source

https://www.kaggle.com/aaron7sun/stocknews

Description:

There are two channels of data provided in this dataset:

(Range: 2008-08-08 to 2016-07-01)

Total: 1987 Train: 1589 Test: 398

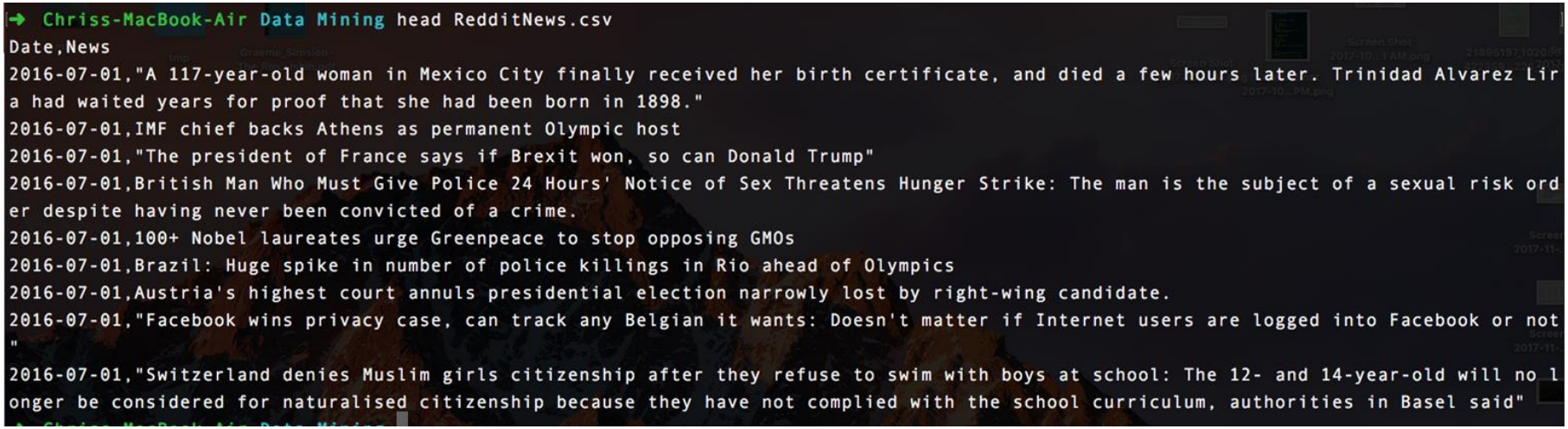
1. News data: crawled historical news headlines from Reddit WorldNews Channel (https://www.reddit.com/r/worldnews?hl). They are ranked by reddit users' votes, and only the top 25 headlines are considered for a single date.
2. Stock data: Dow Jones Industrial Average (DJIA) is used to "prove the concept".

Dataset DJIA_TABLE.csv

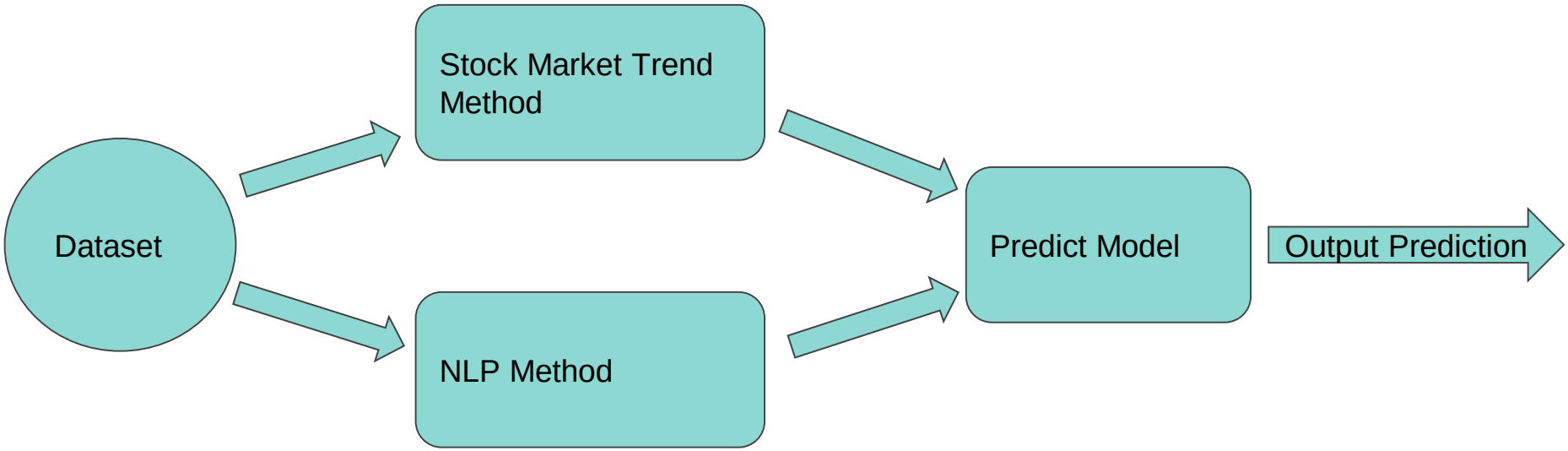
Preview | Column Metadata

Date	Open	High	Low	Close	Volume	Adj Close
2016-07-01	17924.240234	18002.380859	17916.910156	17949.369141	82160000	17949.369141
2016-06-30	17712.759766	17930.609375	17711.800781	17929.990234	133030000	17929.990234
2016-06-29	17456.019531	17704.509766	17456.019531	17694.679688	106380000	17694.679688
2016-06-28	17190.509766	17409.720703	17190.509766	17409.720703	112190000	17409.720703
2016-06-27	17355.210938	17355.210938	17063.080078	17140.240234	138740000	17140.240234
2016-06-24	17946.630859	17946.630859	17356.339844	17400.75	239000000	17400.75
2016-06-23	17844.109375	18011.070312	17844.109375	18011.070312	98070000	18011.070312
2016-06-22	17832.669922	17920.160156	17770.359375	17780.830078	89440000	17780.830078
2016-06-21	17827.330078	17877.839844	17799.800781	17829.730469	85130000	17829.730469
2016-06-20	17736.869141	17946.359375	17736.869141	17804.869141	99380000	17804.869141
2016-06-17	17733.439453	17733.439453	17602.779297	17675.160156	248680000	17675.160156

Dataset RedditNews.csv

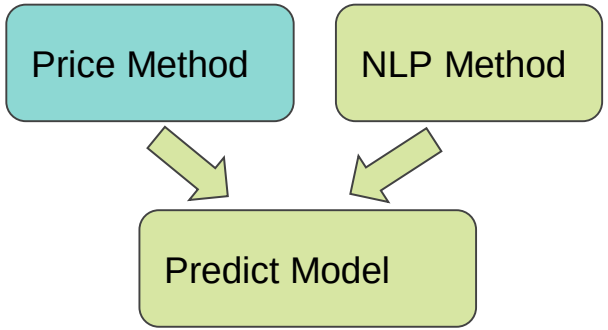


Methods



Each results are the average after 3 times tested.

Methods Stock -> Trend



Test and Find The Highest Accuracy Algorithm

1. Classification
2. Regression
3. Time Series Forecasting

Methods Stock -> Trend (Classification)

Features **Label**

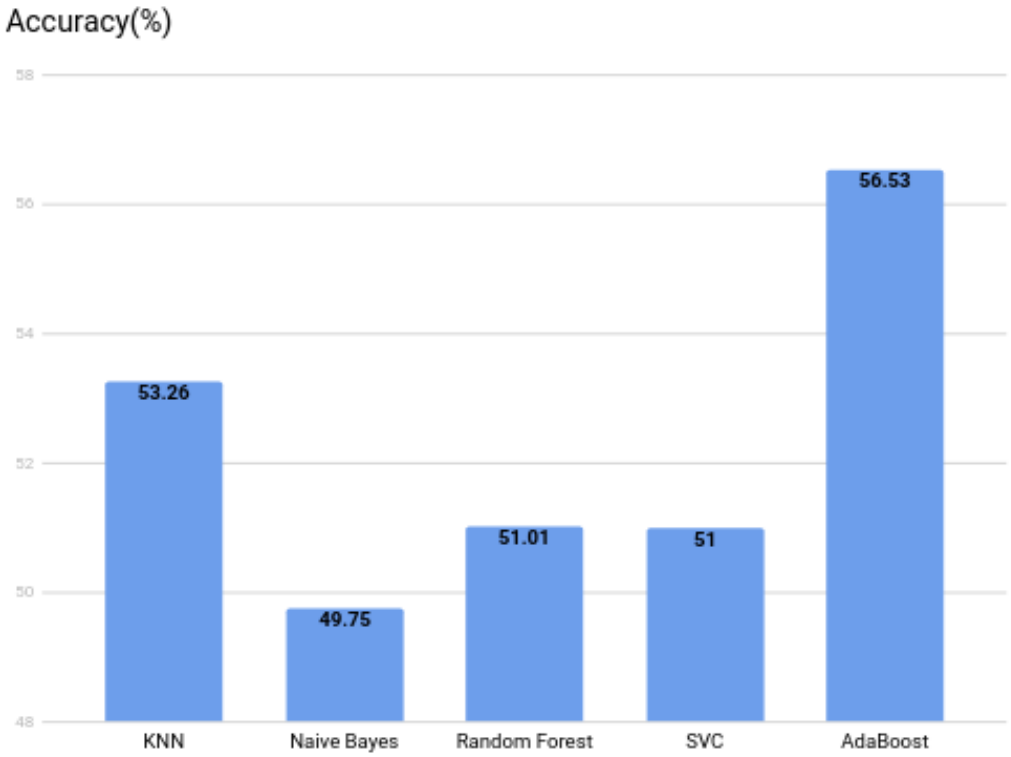
High1	Low1	Volume1	label1	day_of_week1	anomaly1	High2	Low2	Volume2	label2	day_of_week2	anomaly2	real_label
11759.959961	11388.040039	212830000.0	1.0	4.0	0.0	11867.110352	11675.530273	183190000.0	1.0	0.0	0.0	0.0
11867.110352	11675.530273	183190000.0	1.0	0.0	0.0	11782.349609	11601.519531	173590000.0	0.0	1.0	0.0	0.0
11782.349609	11601.519531	173590000.0	0.0	1.0	0.0	11633.780273	11453.339844	182550000.0	0.0	2.0	0.0	1.0

Features: 2 previous days' data

Label: label of the 3rd day

High:highest price of that day
Low: lowest price of that day
Volume: total number of a security that were traded on that day
label: "1" when DJIA Adj Close value rose or stayed as the same;
"0" when DJIA Adj Close value decreased.
day of week: 1(Monday)-5(Friday)
anomaly: "1" when DJIA Adj Close percent changes > 0.02 or <-0.03;
"0" when DJIA Adj Close percent changes -0.03~0.02.

Methods Stock -> Trend (Classification)



Methods Stock -> Trend (Regression)

2. Regression

Features

	High1	Low1	Volume1	label1	day_of_week1	anomaly1	percent_change1
1	11759.959961	11388.040039	212830000.0	1.0	4.0	0.0	0.026496
2	11867.110352	11675.530273	183190000.0	1.0	0.0	0.0	0.004093

High2	Low2	Volume2	label2	day_of_week2	anomaly2	percent_change2
11867.110352	11675.530273	183190000.0	1.0	0.0	0.0	0.004093
11782.349609	11601.519531	173590000.0	0.0	1.0	0.0	-0.011872

Label

real_label
-0.011872
-0.009406

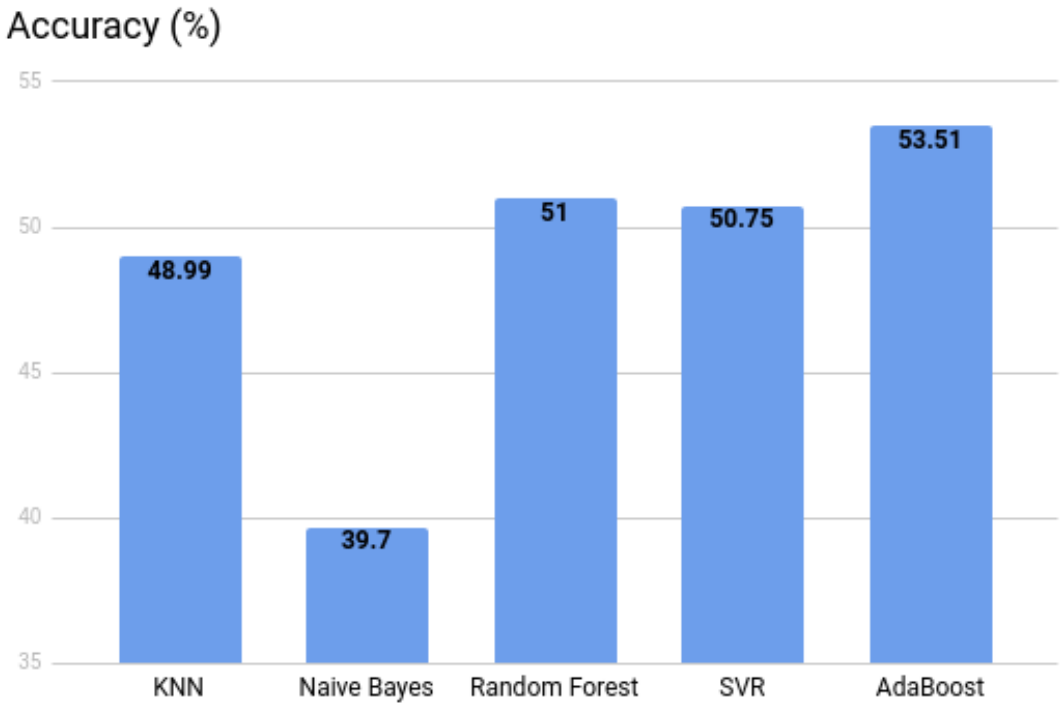
Features: 2 previous days' data of the 3rd day

Label: percent change

High:highest price of that day
Low: lowest price of that day
Volume: total number of a security that were traded on that day
label: "1" when DJIA Adj Close value rose or stayed as the same;
"0" when DJIA Adj Close value decreased.
day of week: 1(Monday)-5(Friday)
anomaly: "1" when DJIA Adj Close percent changes > 0.02 or <-0.03;
"0" when DJIA Adj Close percent changes -0.03~0.02.

percent_change: (close(j)-close(j-1)) / close(j-1)

Methods Stock -> Trend (Regression)



Methods Stock -> Trend (Forecasting)

Time Series Forecasting - Prophet

input :

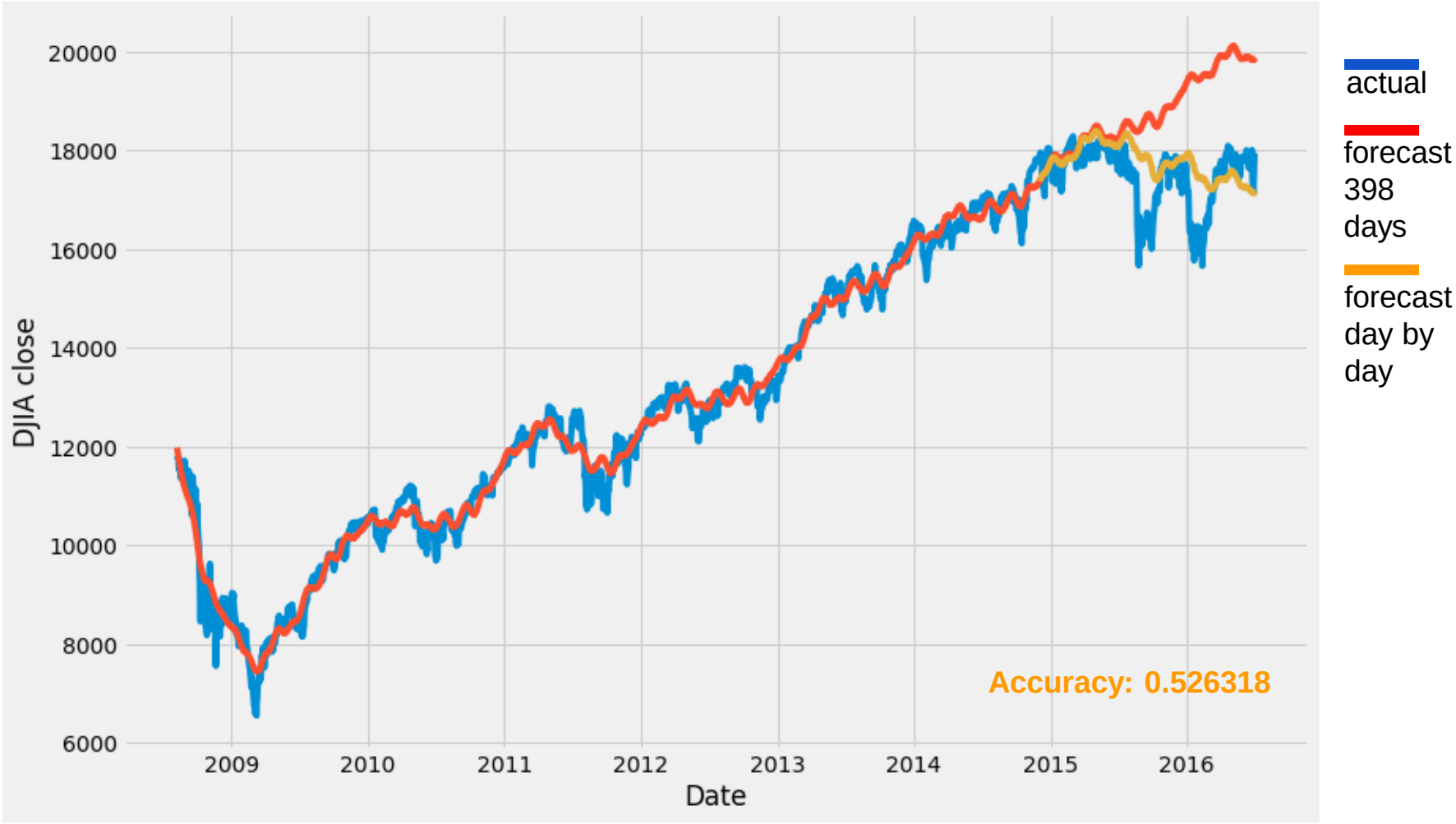
DJIA close (float)

output :

DJIA close (float)

	ds	y
0	2008-08-08	11734.320312
1	2008-08-11	11782.349609
2	2008-08-12	11642.469727
3	2008-08-13	11532.959961
4	2008-08-14	11615.929688

ds: date
y: DJIA close

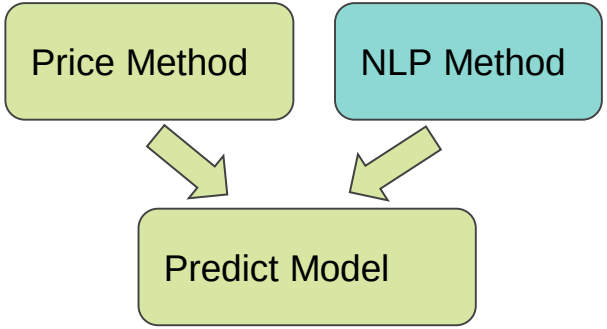

20000
18000
16000
14000
12000
10000
8000
6000
DJIA close
2009
2010
2011
2012
2013
2014
2015
2016
Date
Accuracy: 0.526318
actual
forecast 398 days
forecast day by day

Methods Stock -> Trend

The highest accuracy I get is from AdaBoostClassifier, which is 0.565327.

	Classifier	Avg_Accuracy	Computation_Time
1	KNeighborsClassifier	0.532663	0.016393
2	MultinomialNB	0.497487	0.004931
3	RandomForestClassifier	0.510050	0.137040
4	SVC	0.510050	5.112153
5	AdaBoostClassifier	0.565327	0.752363

Methods News -> Trend Data Input

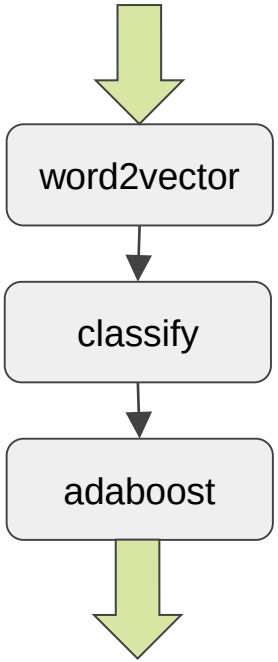


input : 2 * 25 news titles (string)

output : 0~1 (float)

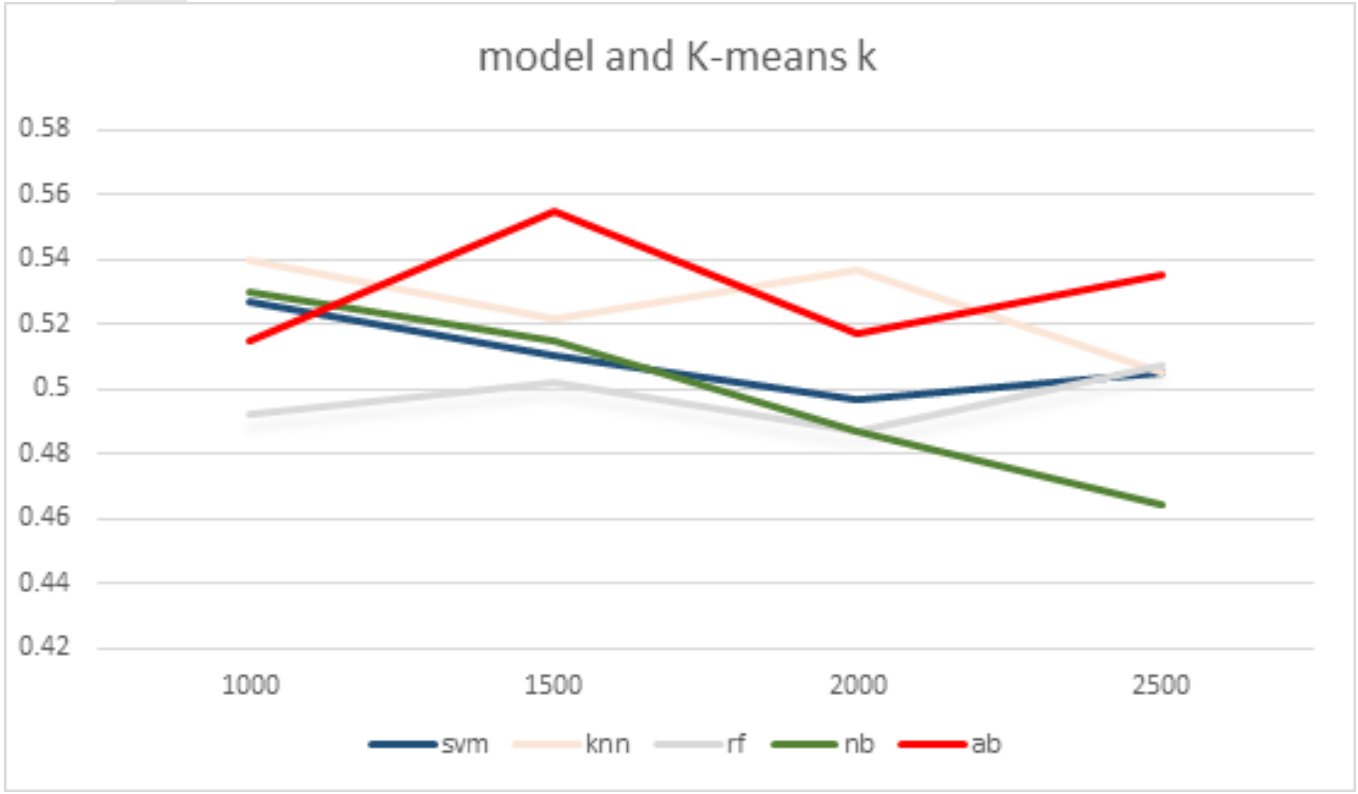
```
array([ ' jamaica proposes marijuana dispensers for tourists at airports following legalisation the
kiosks and desks would give people a license to purchase up to 2 ounces of the drug to use during th
eir stay stephen hawking says pollution and stupidity still biggest threats to mankind we have certa
inly not become less greedy or less stupid in our treatment of the environment over the past decade
boris johnson says he will not run for tory party leadership six gay men in ivory coast were abused
and forced to flee their homes after they were pictured signing a condolence book for victims of the
 recent attack on a gay nightclub in florida switzerland denies citizenship to muslim immigrant girl
s who refused to swim with boys report palestinian terrorist stabs israeli teen girl to death in her
 bedroom puerto rico will default on $1 billion of debt on friday republic of ireland fans to be awa
rded medal for sportsmanship by paris mayor afghan suicide bomber kills up to 40 - bbc news us airst
rikes kill at least 250 isis fighters in convoy outside fallujah official says',
       ' explosion at airport in istanbul yemeni former president terrorism is the offspring of wahh
abism of al saud regime uk must accept freedom of movement to access eu market devastated scientists
 too late to captive breed mammal lost to climate change - australian conservationists spent 5 month
s obtaining permissions planning for a captive breeding program but when they arrived on the rodents
 tiny island they they were too late british labor party leader jeremy corbyn loses a no-confidence
vote but refuses to resign a muslim shop in the uk was just firebombed while people were inside mexi
can authorities sexually torture women in prison uk shares and pound continue to recover iceland his
torian johannesson wins presidential election 99-million-yr-old bird wings found encased in amber -
finding things trapped in amber is far from rare but when researchers in burma found a pair of tiny
bird-like wings frozen inside they knew they had something special'], dtype=object)
```

Methods News -> Trend



1. google news words vectors
    a. 300-dimensional vectors for 3 million words (100 billion words trained)
2. Kmeans clustering 1500 groups
    a. distance : Euclidean (same as cosine after normalization)
    b. k = 1500 by experiment {500, 1000, 1500, 2000, 2500}
3. calculate daily news group counts
    a. eg : {0, 0, 2, 1, 0...}
4. Machine learning model
    a. adaboost is better for our case
5. output positive trend possibility (float 0~1)

Methods News -> Trend



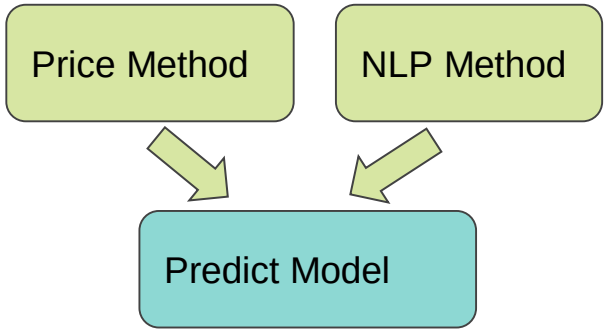
random forest
perform as random

naive bayes
worst prediction
(too many 0 in vector)

SVM classify good
regression bad
(too less data)

adaboost is the best
advantage to sparse

Combining Model



NLP : 0~1 (Float)

Stock: 0~1 (Float)

Output: 0, 1 (int)

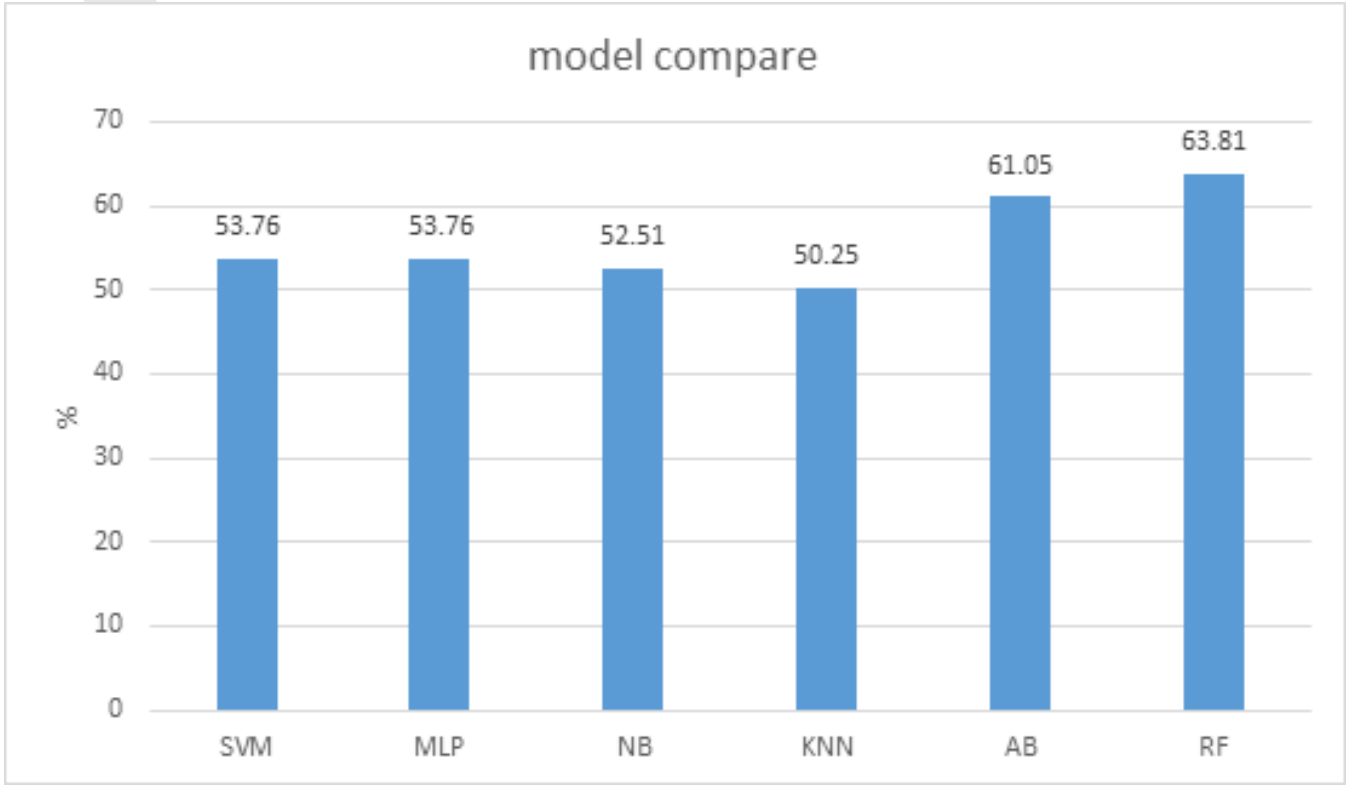
(0: Stock value drop, 1: Stock value rise)

```
X = pd.DataFrame(columns=['nlp', 'stock'])
X['nlp'] = df_nlp
X['stock'] = df_stock['stock_pred']
Y = df_stock['true'].tolist()

print(X.head())
print(Y[0:5])
```

```
        nlp     stock
0  0.499311  0.499963
1  0.505331  0.501456
2  0.489887  0.502461
3  0.489887  0.501412
4  0.505331  0.500476
[0, 0, 1, 1, 0]
```

Results



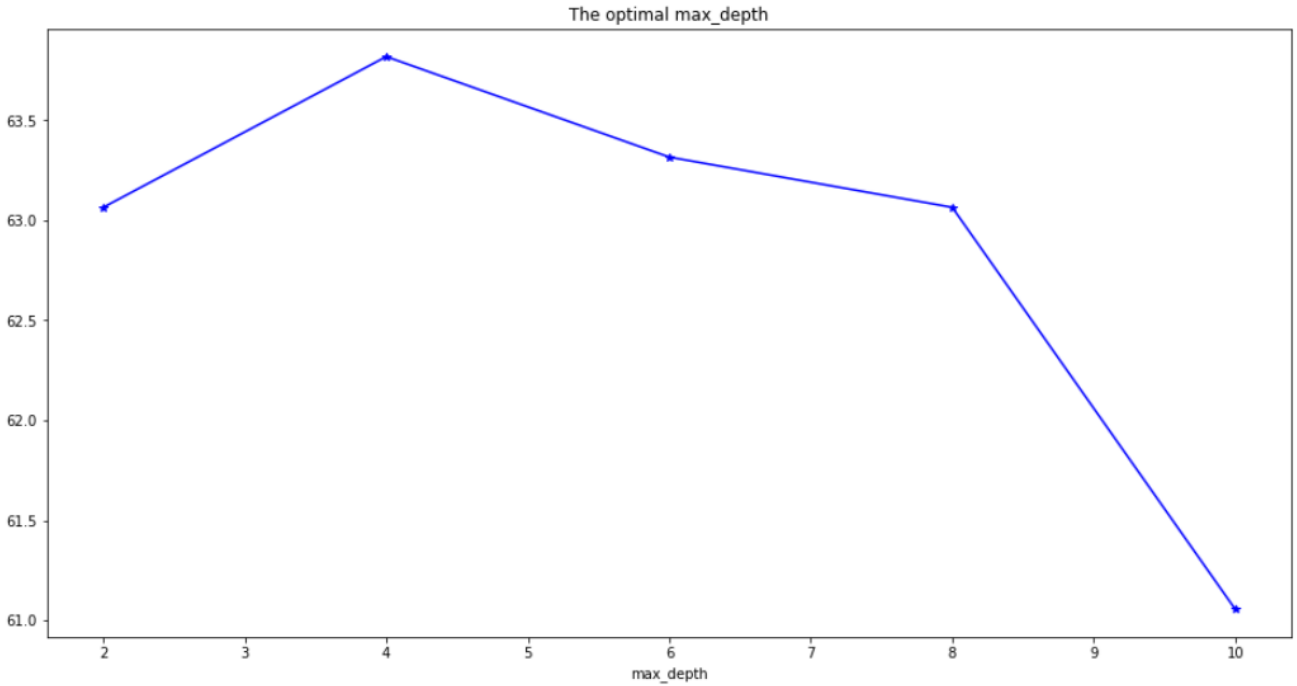
SVM、MLP
all guess 1

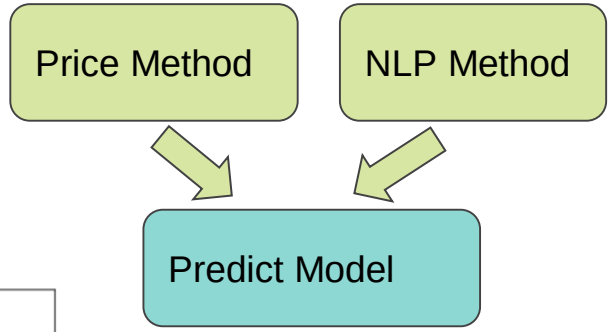
NB
most of predict is 1

KNN
worst predict
(overlapping)

Adaboost &
Random forest
balance prediction
predict good

parameters optimal





Optimal random forest

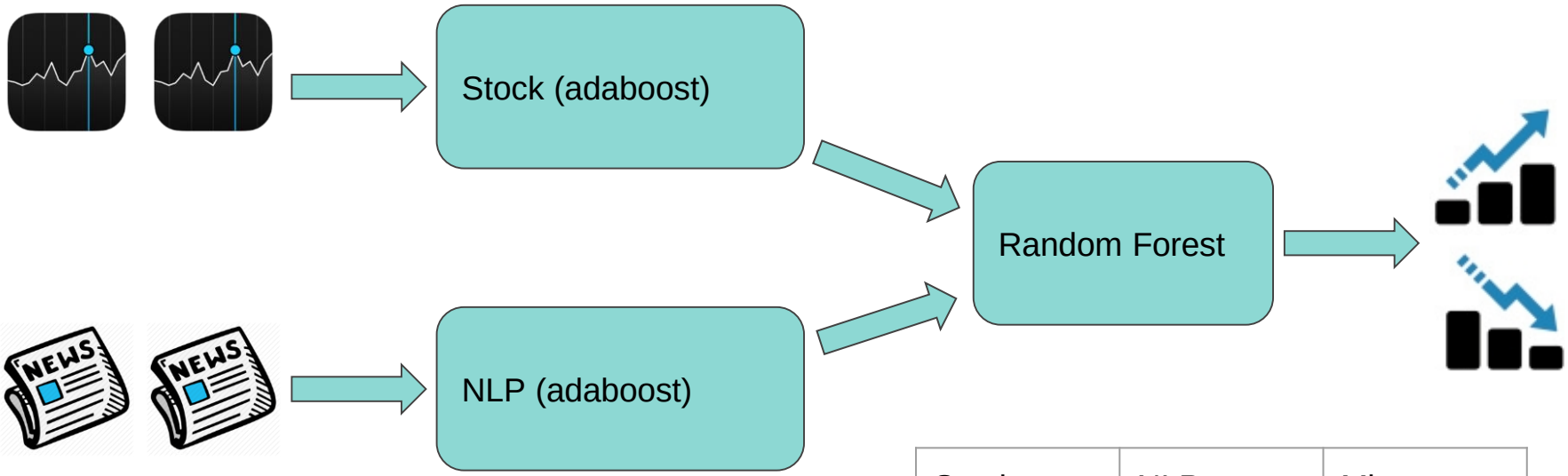
estimators = 100

max_depth = 4

max_depth > 4 accuracy drop

Because model overfitting!

Experimental Results



Stock	NLP	Mixture
56%	55%	63%

Experimental Results

accuracy = 63.81%
(TP+TN) / (TP+FP+TN+FN)

precision = 77.10%
TP / (TP+FP)

recall = 63.46%
TP / (TP+FN)

F1 = 69.62%
(2*P*R) / (P+R)

Total=398	**Relevant**	**Non-Relevant**
Retrieved	TP=165	FP=49
Non-Retrieved	FN=95	TN=89

Total data : 1987; Train: first 1589; Test: last 398

The result is the average number after 5 times tested.

Issues faced in this project

Stock Market Trend Prediction:

- The DJIA index go higher, year by year
  - Normalization (percent change)
- Sudden drops affected by unpredictable events (like terrorism, war...)
  - Anomally indicator

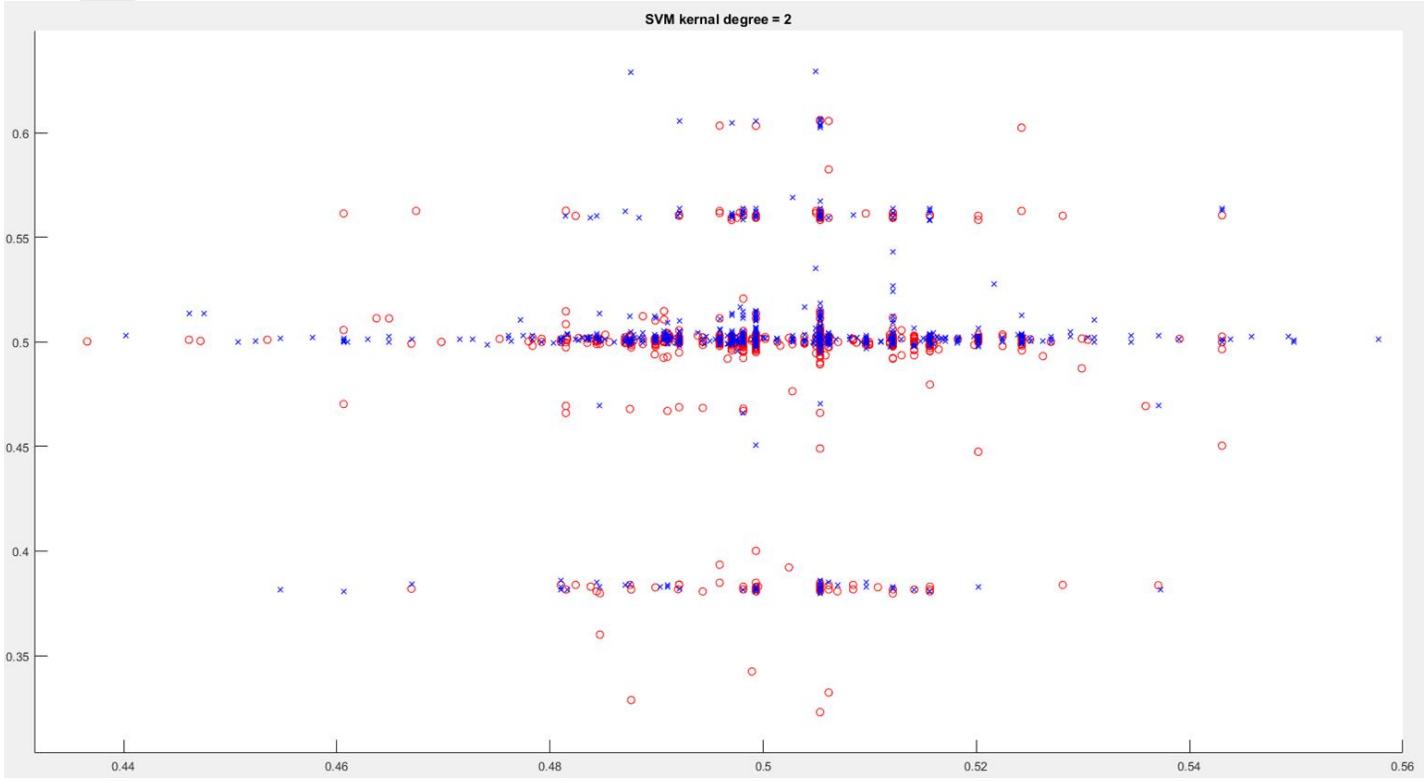
NLP prediction

- too less training data for words to vectors training
  - download pre-trained model from google
- Sklearn kmeans did not provide cosine distance
  - data normalization

Final prediction

- If prediction accuracy is greater than 70%, we consider it might be overfitting in luck

Other efforts - SVM regression



資料分佈重疊
SVM難切割
最終全預測1

自己手刻
SVM regression
(Gaussion)

kernel function
K = X'X

degree = 1,2
皆失敗

Conclusions and future works

Conclusions:

- Data normalization can improve the accuracy of the prediction.
- The combination of price and news prediction can increase the precision.

Future Works:

- Can find out all DJIA’s 30 companies data and increase the amount of data
- Deep Learning (RNN, LSTM ...) may be a better modal

Job description

陳子軒 25% price trend, combine model, ppt

楊光傑 25% price trend, paper survey, ppt

王威斌 25% NLP, result analysis, ppt

薛世恩 25% NLP, paper survey, ppt

References

1. Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. Listening to Chaotic Whispers: A Deep Learning Framework for News-oriented Stock Trend Prediction. WSDM 2018 .https://arxiv.org/pdf/1712.02136.pdf

1. Qi He, Kuiyu Chang, and Ee-Peng Lim. Using Burstiness to Improve Clustering of Topics in News Streams. ICDM.2007. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4470279&tag=1

Q&A