

#PredictingTheDow

Daniel Mooney, Mohammed Faiz Shaikh, Apurva Tripathi

Theory

It has long been theorized that the news affects the stock market.

September 11th - when the markets opened on 9/17, NYSE went down 680 points (7.1%)

Lehman Brothers collapse led Dow closing 4.4% or 504 point down.

Quants are already doing this - Hathaway effect

Can we predict the stock market movement from the news?

Efficient Market Theory

it is impossible to "beat the **market**" because stock **market efficiency** causes existing share prices to always incorporate and reflect all relevant information

Problem

Using /r/worldnews (Reddit) to predict whether the stock market(measured by Dow Jones Index) will go up or down

We are given Top 25 News Items of the day, along with

'Open', 'High', 'Low', 'Close', 'Volume', 'Adj Close'

Predictor: 'Label' (1 if Open-Close >0 else 0)



reddit

WORLDNEWS

hot

new

rising

controversial

top

gilded

wiki

promoted

Other Subs:

Related

N. America

S. America

Europe

Asia

Middle East

Africa

Oceania

讀賣新聞

DER SPIEGEL

人民日報

USA

REUTERS

Ap

BBC

The Sydney Morning Herald

The Sydney Morning Herald

The Sydney Morning Herald

Are you a programmer, quant or just love math? Break into algorithmic forex trading with FXCM. Free advanced trading APIs. Sign up and test for free. Losses can exceed your deposit. (landing.fxcm.com)

promoted by FXCM_
comment share

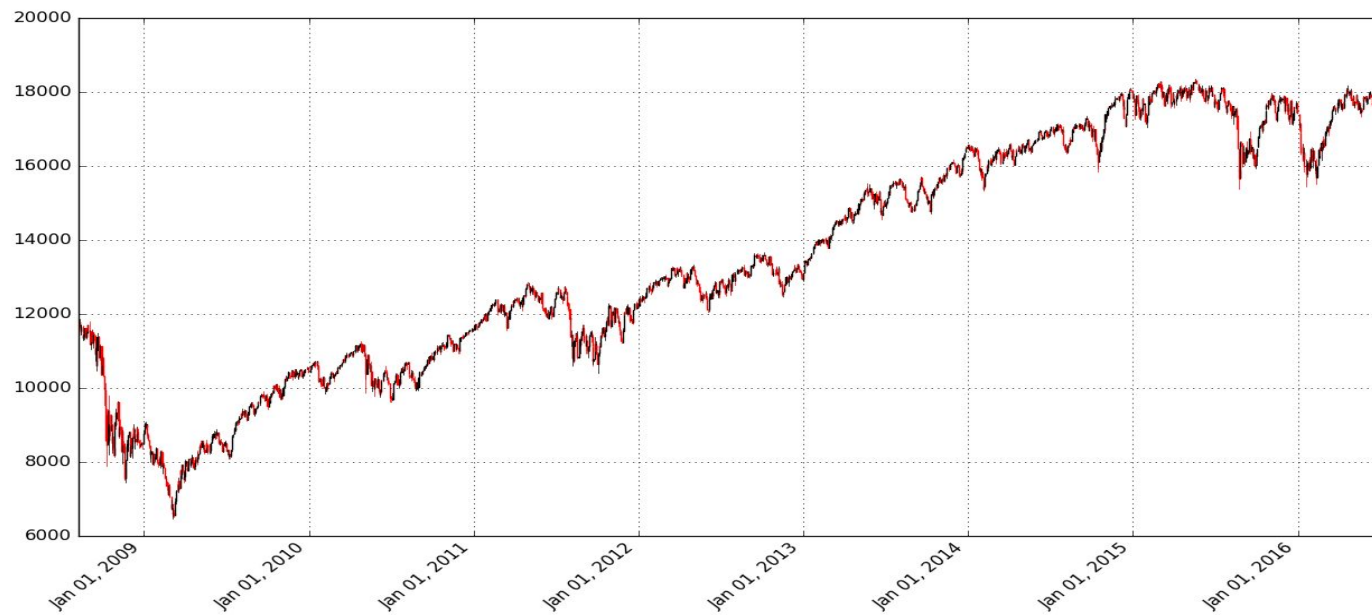
promoted post what's this?

- ↑ 6400 ↓
Elon Musk says the solar roof that will be sold under a combined Tesla-SolarCity will likely cost less than a normal roof to install. (sciencealert.com)
submitted 13 hours ago by maxwellhill
2802 comments share
- ↑ 3343 ↓
BBC Angela Merkel to stand for fourth term (bbc.co.uk)
submitted 13 hours ago by Crazyfrog3214
1366 comments share
- ↑ 381 ↓
BBC French ex-President Nicolas Sarkozy admits defeat in conservative presidential primary - BBC News (bbc.com)
submitted 4 hours ago by JustJoeB
82 comments share
- ↑ 883 ↓
REUTERS Obama tells Latin America and world: give Trump time, don't assume worst (reuters.com)
submitted 11 hours ago by lnewsjournal
312 comments share
- ↑ 279 ↓
theguardian Big turnout as French right votes for candidate to oppose Marine Le Pen (theguardian.com)
submitted 5 hours ago by stylishreddit
139 comments share
- ↑ 1358 ↓
INDEPENDENT Tony Blair 'returning to politics' because he thinks Jeremy Corbyn 'is a nutter' and Theresa May 'is a lightweight' (independent.co.uk)

**I MADE ALL MY MONEY IN
THE STOCK MARKET**

**MOSTLY TRADED ON
THE MEOW JONES**

Exploration Time...(candlestick issues)



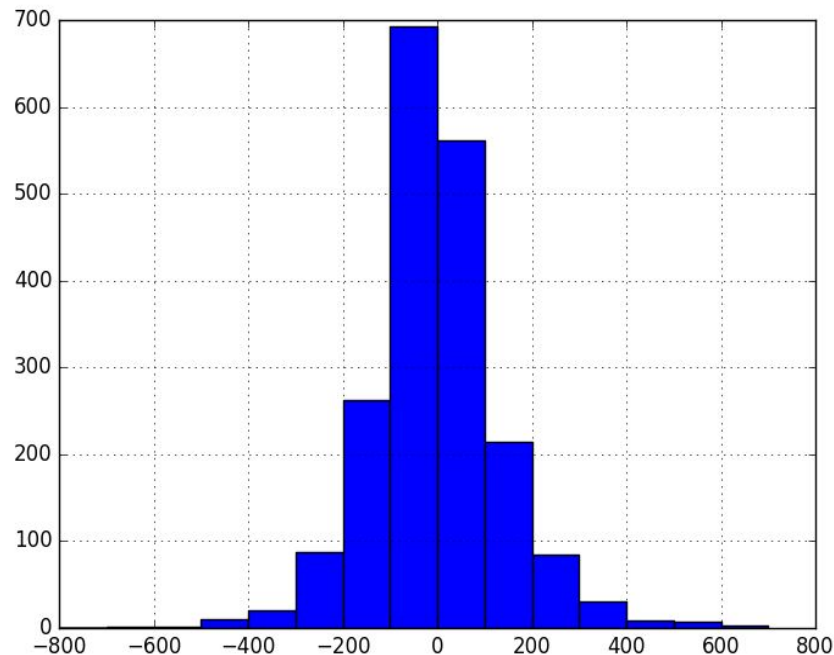
Exploration Time...(cont..)

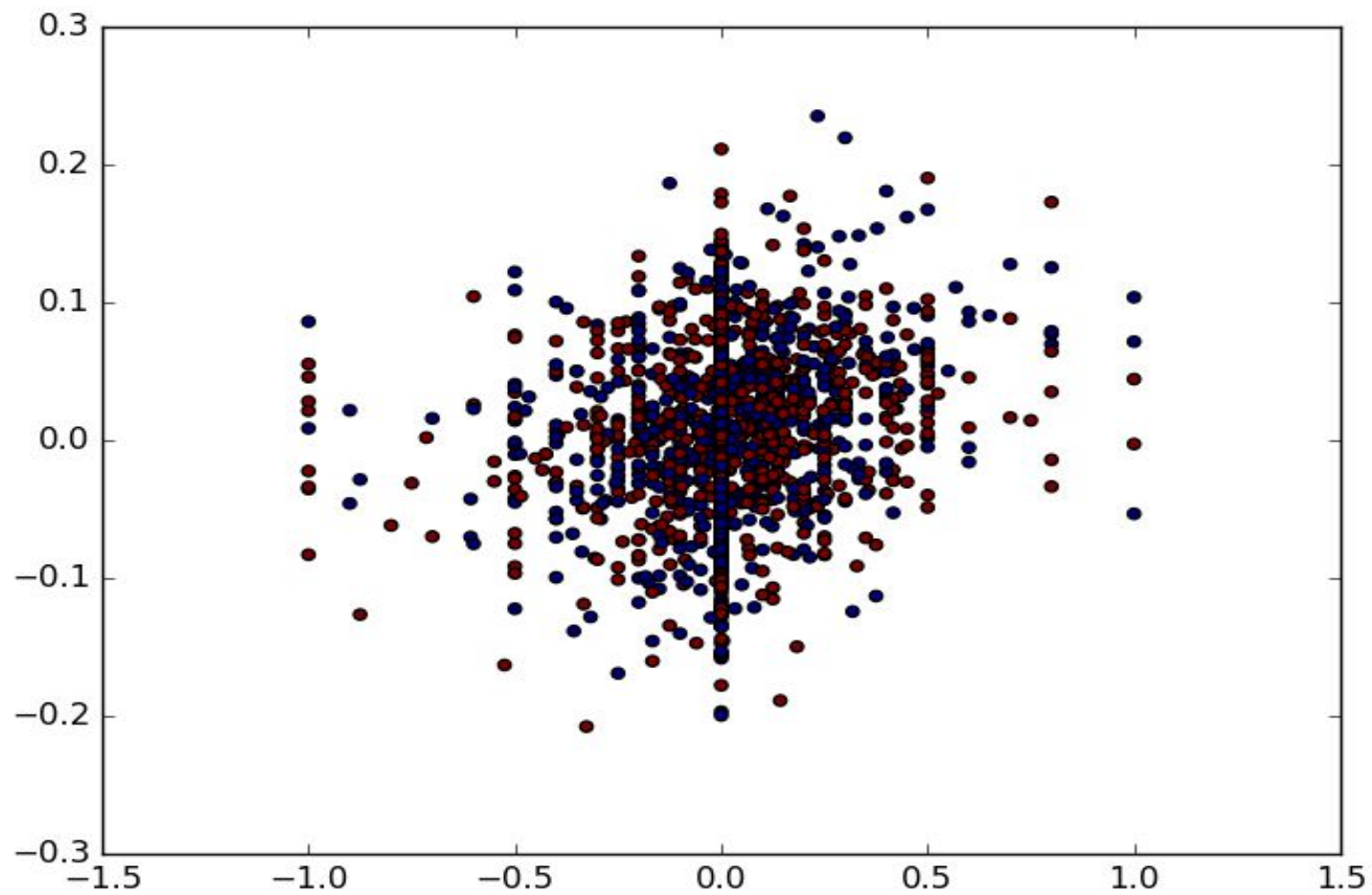
06/08/2008 -> 07/01/2016

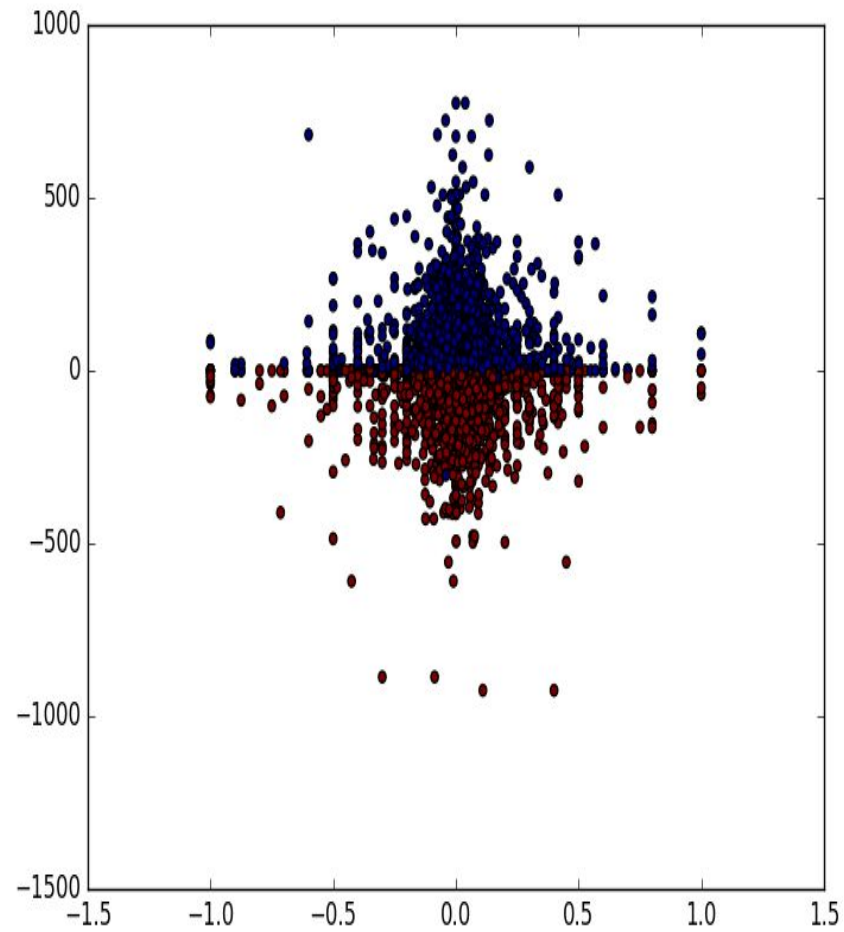
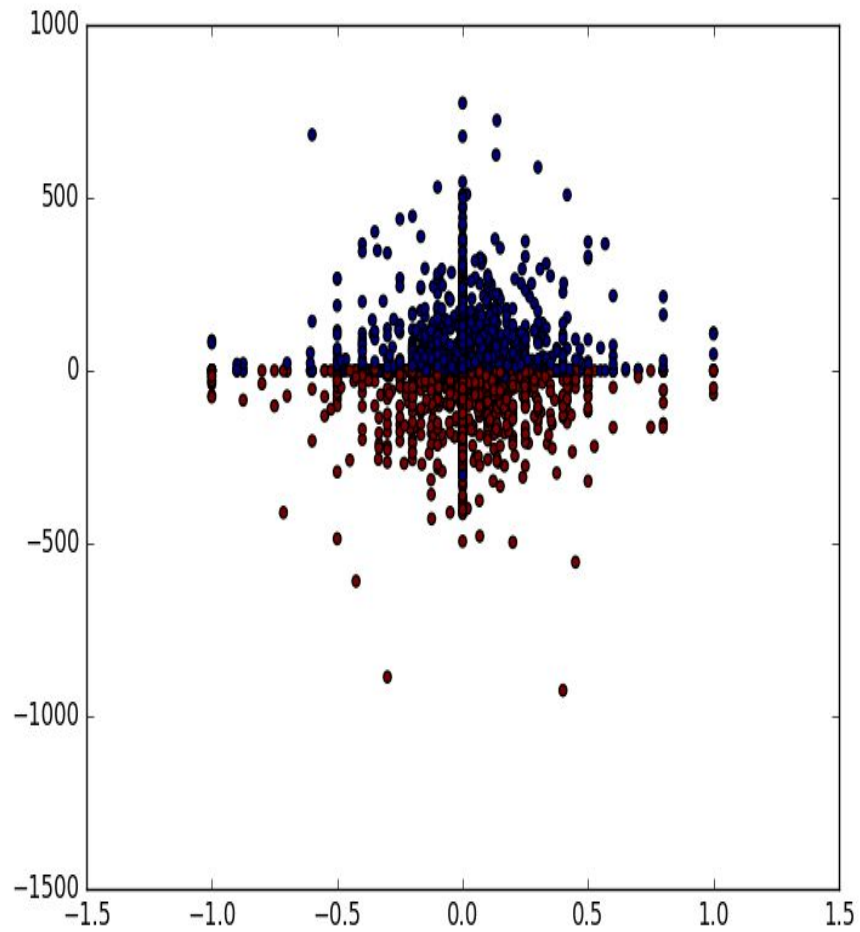
Stock Market went up 53% during that time (the goal to beat)

Can we beat it?!

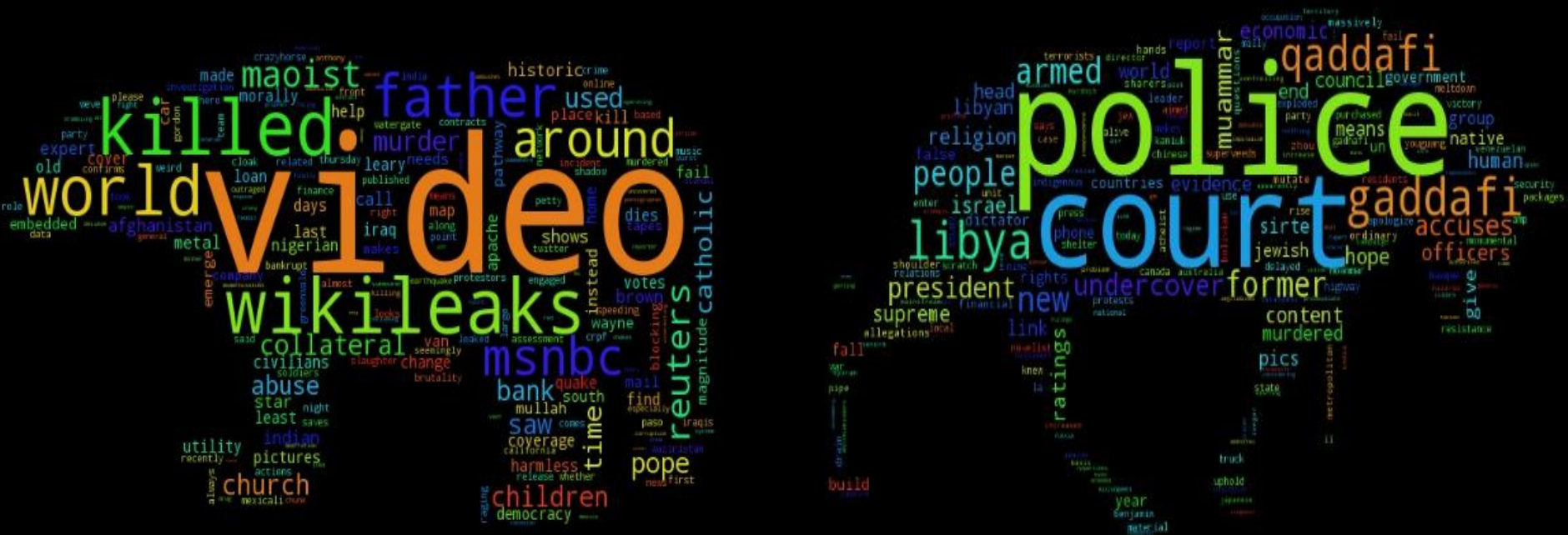
Distribution







Word Clouds



We ran into problems...

In our first or second checkpoint, we reported accuracies of 80% and better

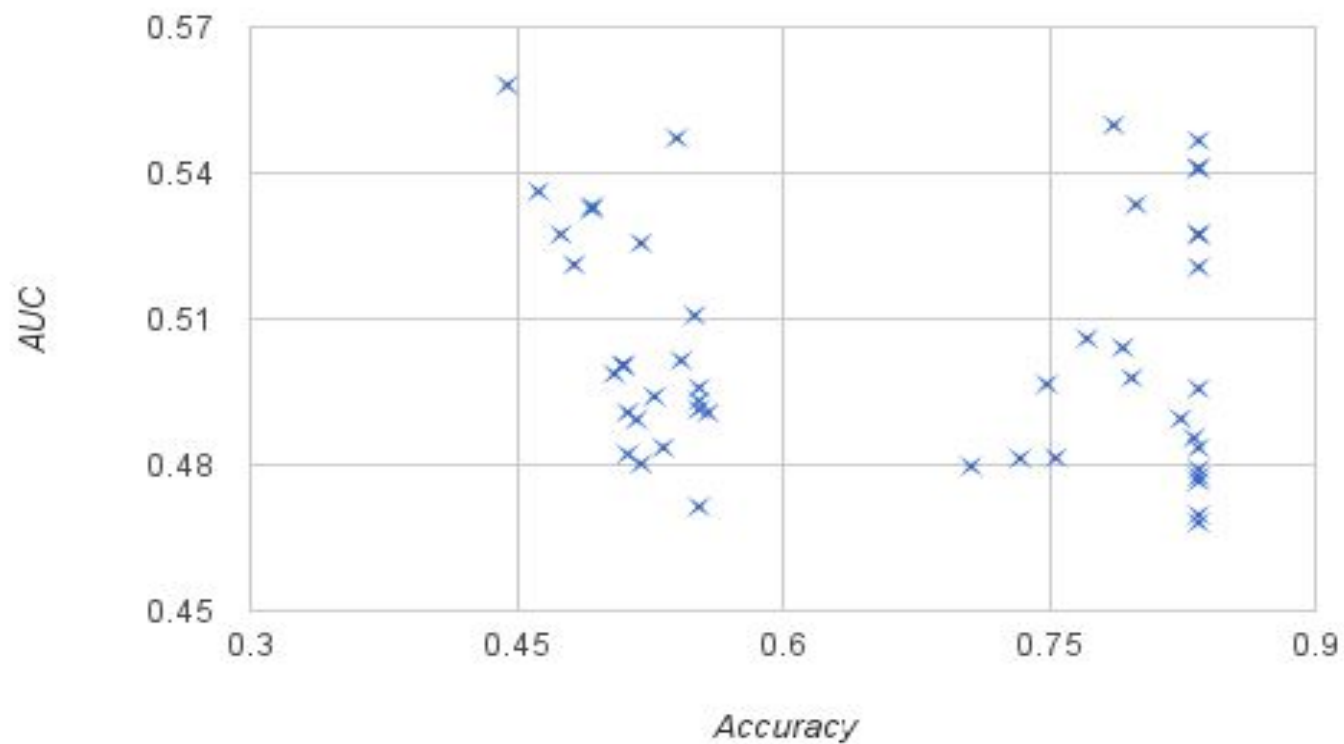
But..we split our training and testing data incorrectly..and some of our training data found its way into our test data as well.

The data was quite messy and CountVectors and Td-idfVectors required stemword removal and stemming

95% of words found when the stock market went down were also found in news articles when the stock market went up

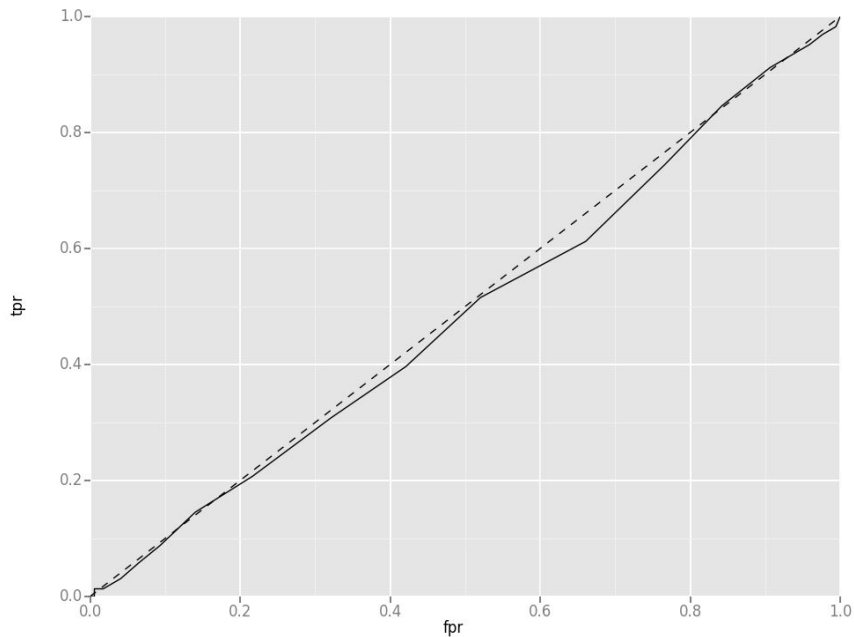
Minor foreign event did not seem to have any effect of the stock market

AUC vs Accuracy

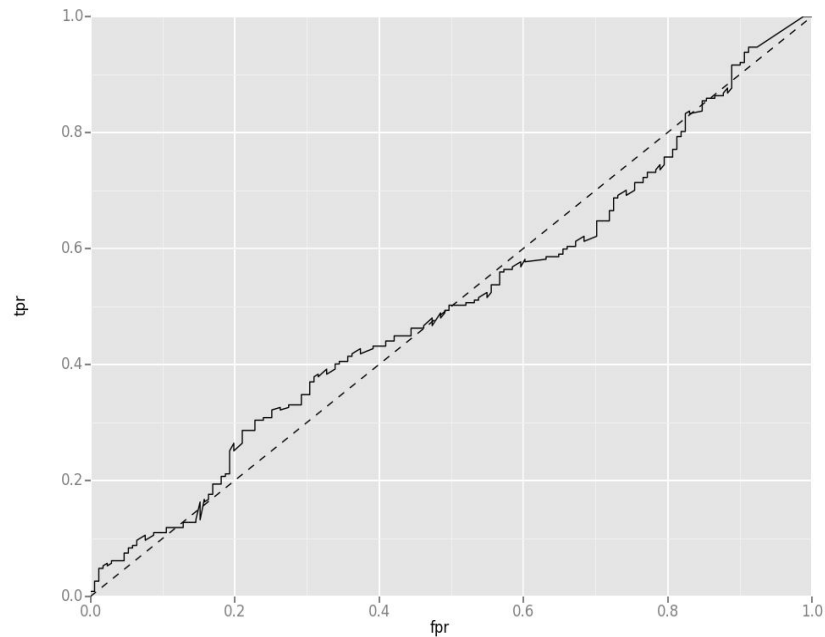


Our Best AUC Curves...

Roc Curve of KNeighborsClassifier



Roc Curve of SVC



Cross Validation and Model Refining

Cross Validating brought most algorithms to have an accuracy of 53%

We concentrated on refining KNN. When we optimizing AUC, accuracy was neglected and vice-versa. We were not able to solidly get AUC above $>.5$ and accuracy above 53%

AUC $< .6$ is quite poor!

Refining KNN

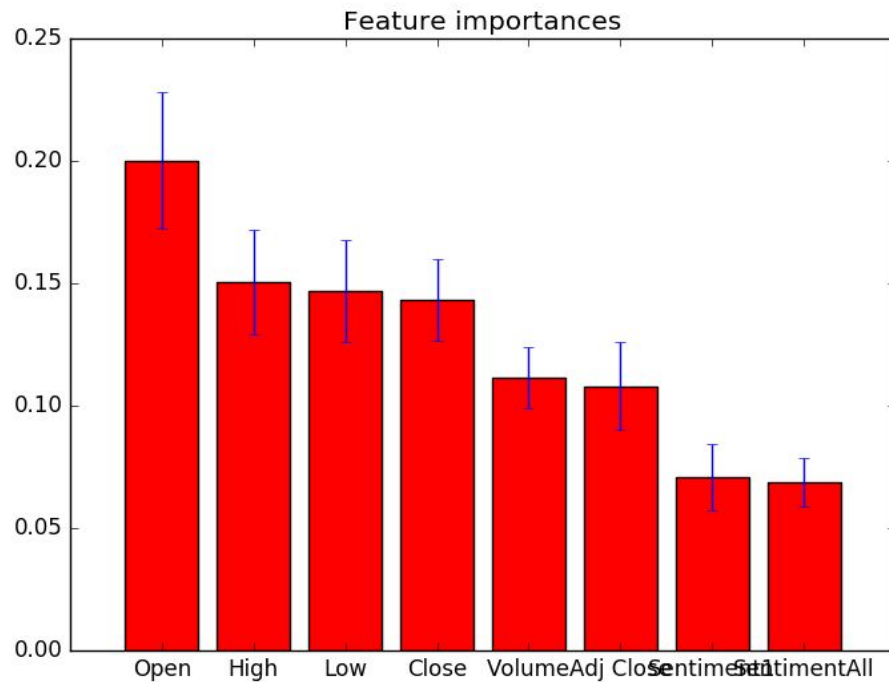
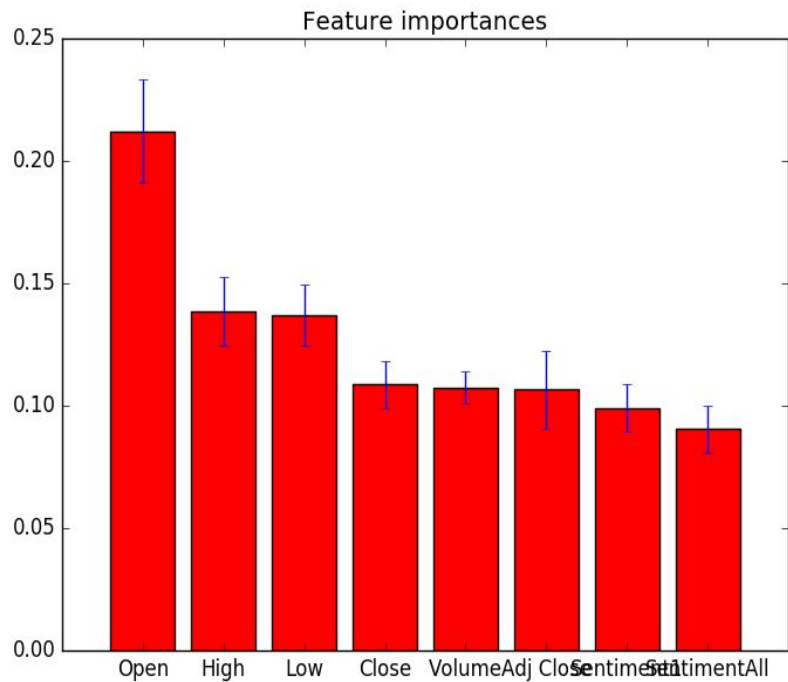
```
[4]+ Stopped          python -i refineKNN.py
Daniels-MacBook-Air:Project danielmooney$ python -i refineKNN.py
KNN using CountVector with neigh 238 had the best AUC  of 0.527938221256an accuracy of 0.499953608247
>>>

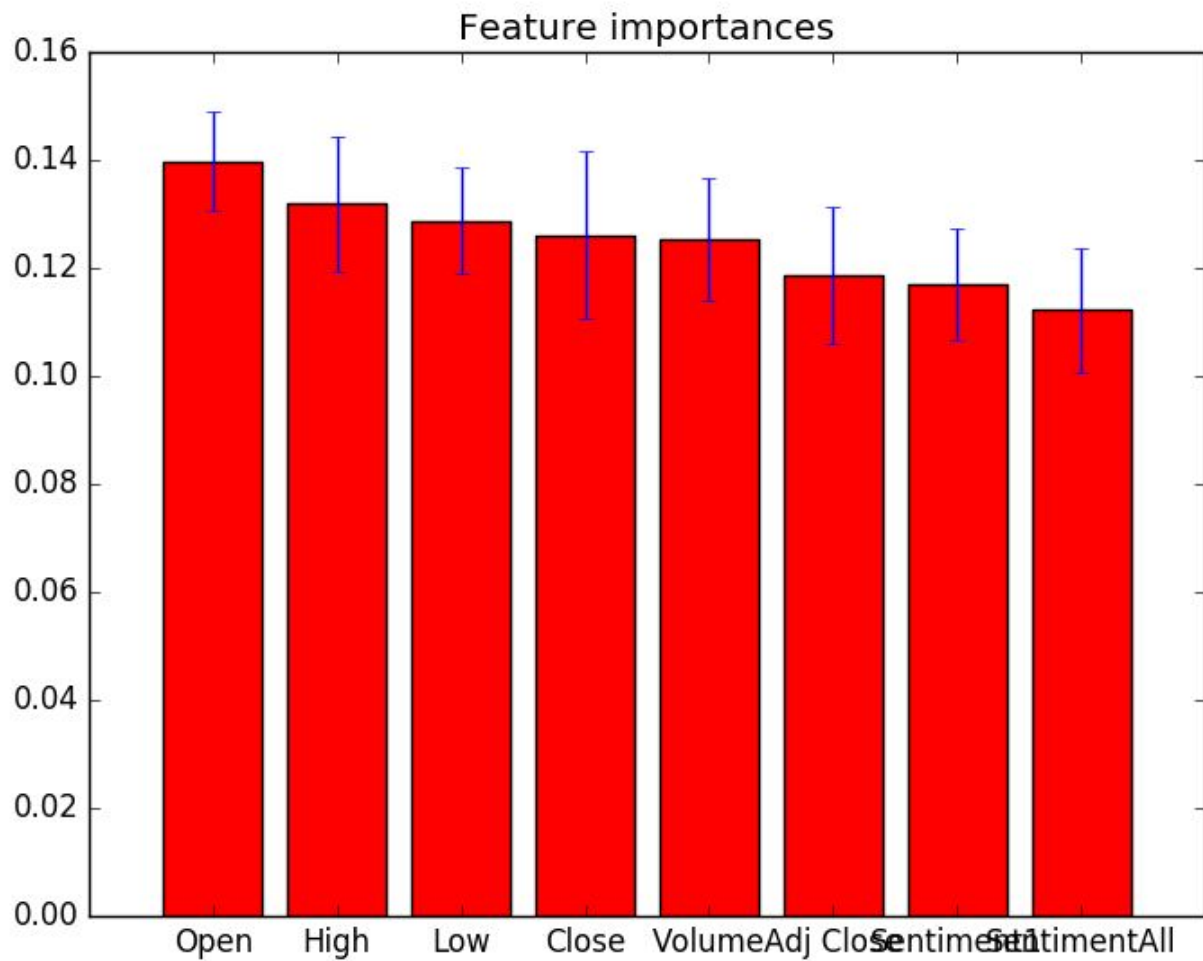
[5]+ Stopped          python -i refineKNN.py
Daniels-MacBook-Air:Project danielmooney$ python -i refineKNN.py
KNN using Tfidf with neigh 4 had the best AUC  of 0.517641765621an accuracy of 0.473829896907
>>>

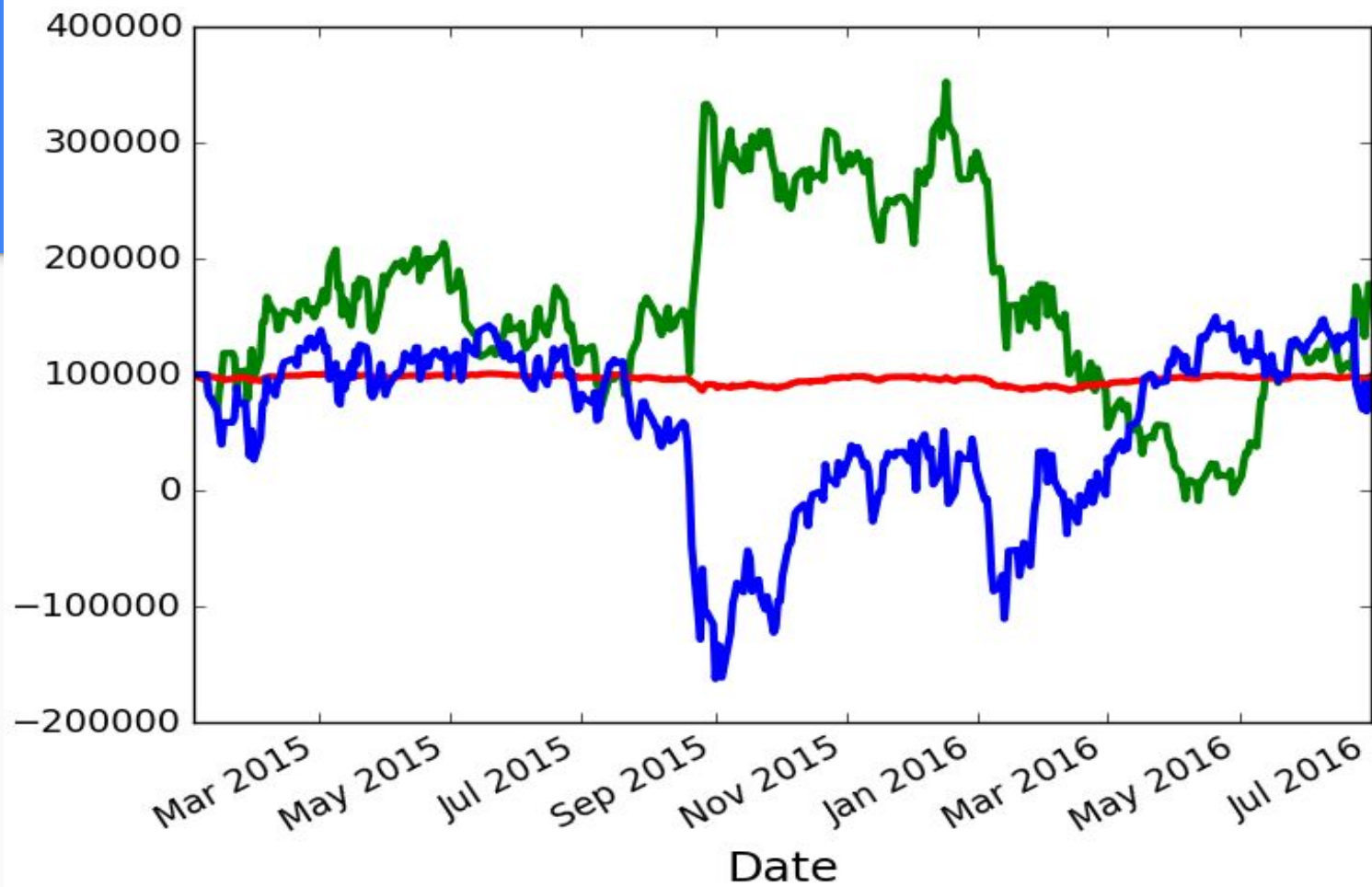
>>>
[9]+ Stopped          python -i refineKNN.py
Daniels-MacBook-Air:Project danielmooney$ python -i refineKNN.py
KNN using CountVector with neigh 299 had the best AUC  of 0.516656727079an accuracy of 0.532170103093
>>> x
299
>>>

[10]+ Stopped         python -i refineKNN.py
Daniels-MacBook-Air:Project danielmooney$ python -i refineKNN.py
KNN using TfidfVectorizer with neigh 280 had the best AUC  of 0.528955796047an accuracy of 0.525855670103
... █
```

Feature Selection







Future Work

Try different text data or more specific data (research with surprisingly good accuracy in this domain) **Correlating Financial Time Series with Micro-Blogging Activity**

More, different vectorization

Incorporate numeric data with text data

More backtesting

Regression Analysis

**YOU GOT 80%
RECOGNITION RATE**



WHY NOT 100%

quickmeme.com

**I'M NOT SAYING ITS THE
DATA**



**BUT ITS THE
DATA**

HISTORY.COM

memegenerator.net

Conclusions

Given the text data that we have, we can't accurately predict whether the stock market will go up or down

Dataset was created for Deep Learning course, maybe a deep learning approach would be beneficial

There's a lot a foreign news that doesn't impact a the stock market. What would happen if we tried a different text source?

What if we analyzed Apple Reddit News vs Apple Stock?