#PredictingTheDow

Daniel Mooney, Mohammed Faiz Shaikh, Apurva Tripathi

Agenda

- Theories
- Problem
- Exploration
- Experimenting
- Refining
- Feature Selection/Importance
- Backtesting
- Conclusions and Future Work

Github Repo

https://github.com/ddm7018/-PredictingTheDow

Supporting 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

Correlating Financial Time Series with Micro-Blogging Activity

Can we predict the stock market movement from the news?

Efficient Market (Opposing) 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

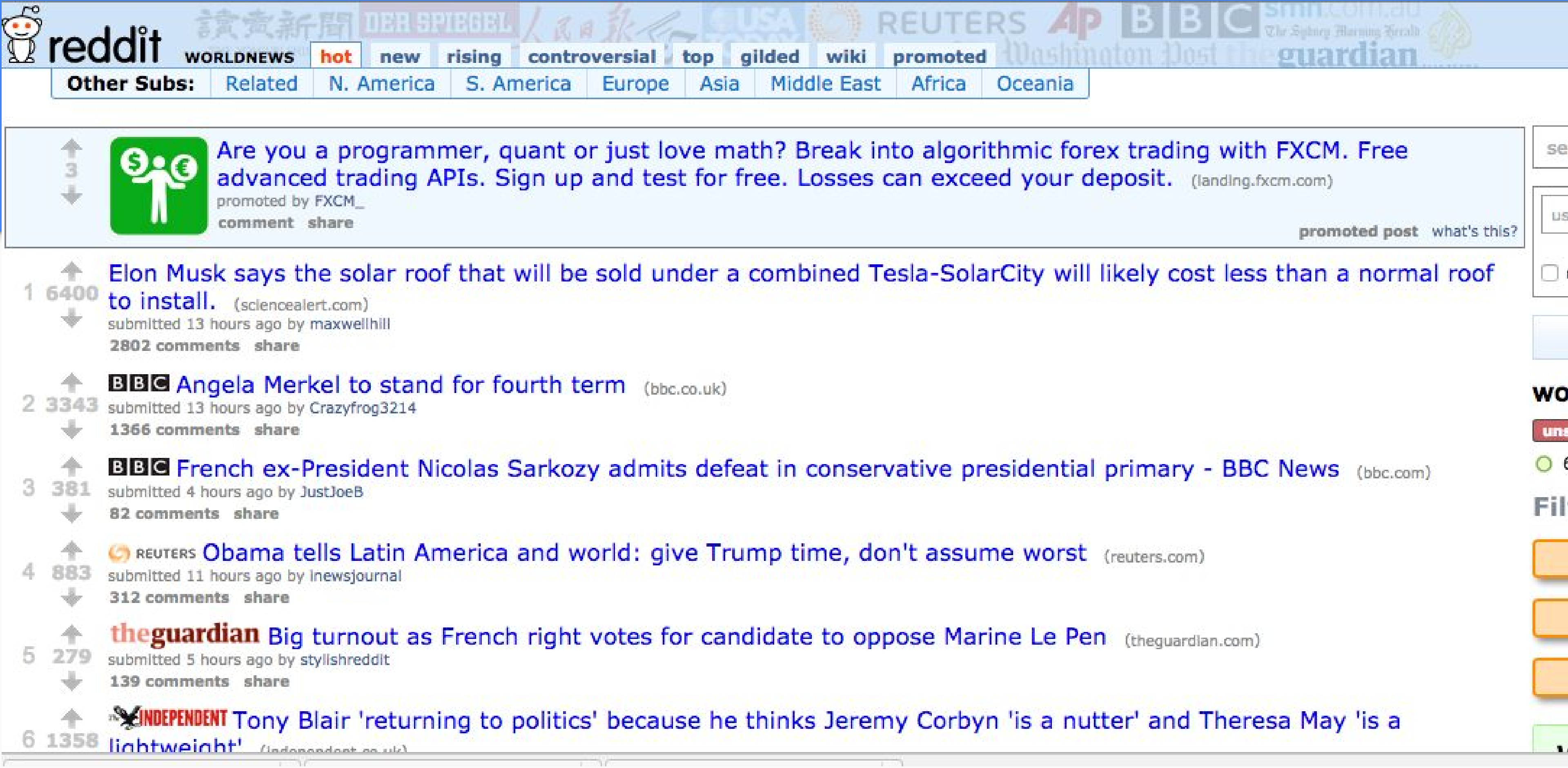
Dataset: https://www.kaggle.com/aaron7sun/stocknews

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)



3 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?

1 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

2 3343 BBC Angela Merkel to stand for fourth term (bbc.co.uk)
submitted 13 hours ago by Crazyfrog3214
1366 comments share

3 381 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

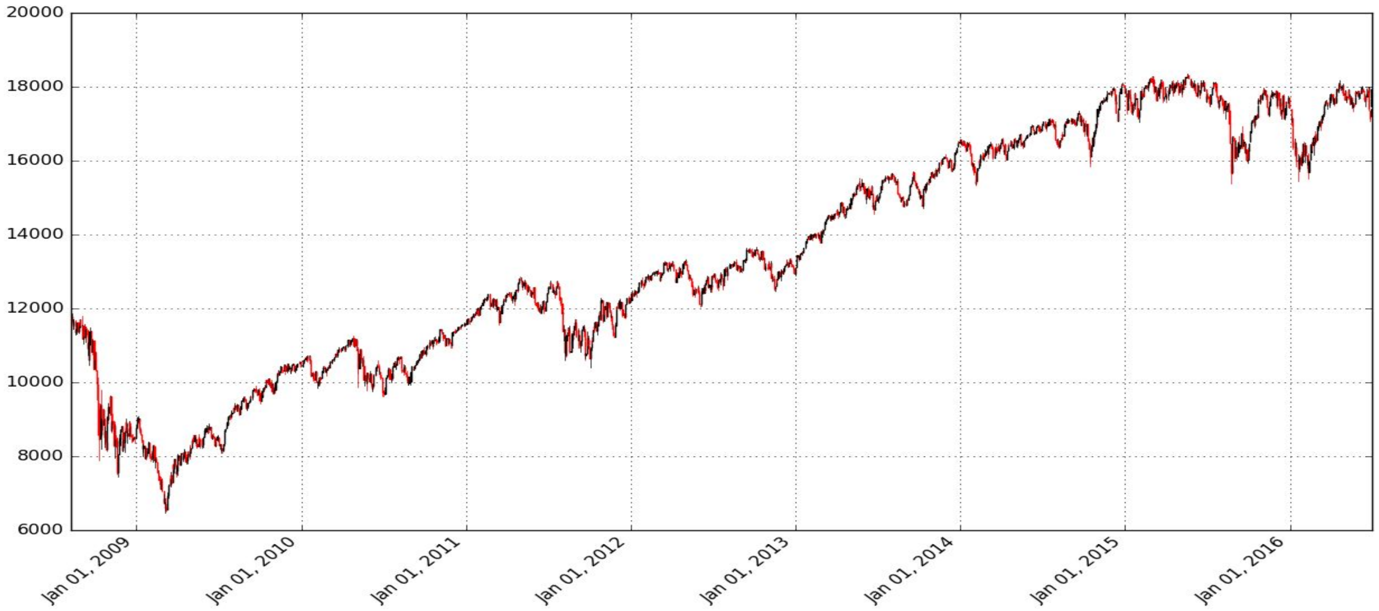
4 883 REUTERS Obama tells Latin America and world: give Trump time, don't assume worst (reuters.com)
submitted 11 hours ago by inewsjournal
312 comments share

5 279 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

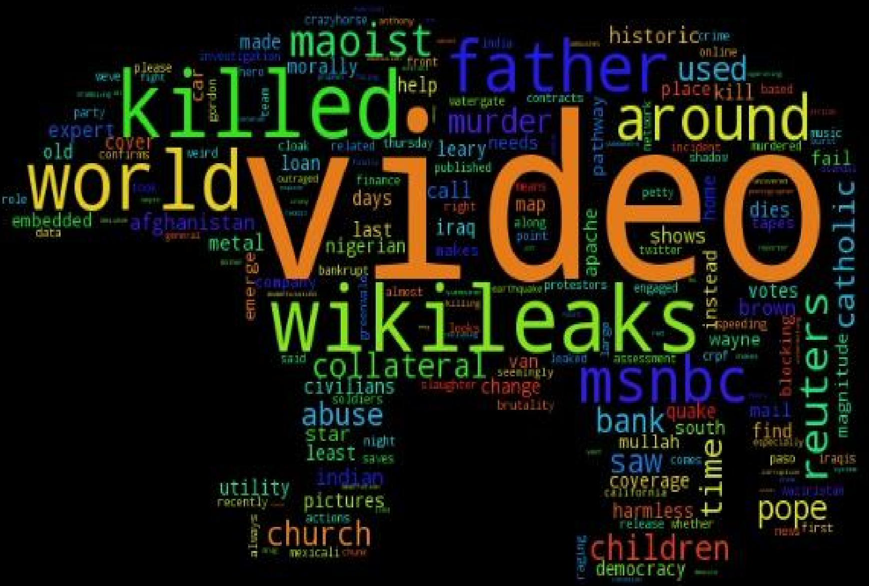
6 1358 INDEPENDENT Tony Blair 'returning to politics' because he thinks Jeremy Corbyn 'is a nutter' and Theresa May 'is a lightweight' (independent.co.uk)

I MADE ALL MY MONEY IN THE STOCK MARKET
MOSTLY TRADED ON THE MEOW JONES

Exploration Time...(candlestick issues)



Word Clouds



video
wikileaks
killed
world
father
around
maoist
murder
msnbc
reuters
catholic
collateral
abuse
church
children
bank
saw
pope
time
used
historic
civilians
afghanistan
nigerian
iraq
pictures
utility
coverage
harmless
democracy
mullah
south
quake
star
least
indian
embedded
expert
cover
old
metal
apache
pathway
instead
votes
wayne
brown
dies
fail
map
days
last
call
leary
needs
help
morally
made
loan
shows
kill
place
change
van
find
mail
magnitude
blocking
emerge
company
bankrupt
almost

police
court
qaddafi
gaddafi
libya
muammar
armed
people
president
undercover
former
new
supreme
religion
libyan
head
israel
dictator
economic
council
government
accuses
officers
hope
sirte
jewish
human
native
group
means
end
content
murdered
pics
rights
link
ratings
evidence
countries
world
report
false
phone
shelter
give
fall
build
year
truck
uphold
allegations

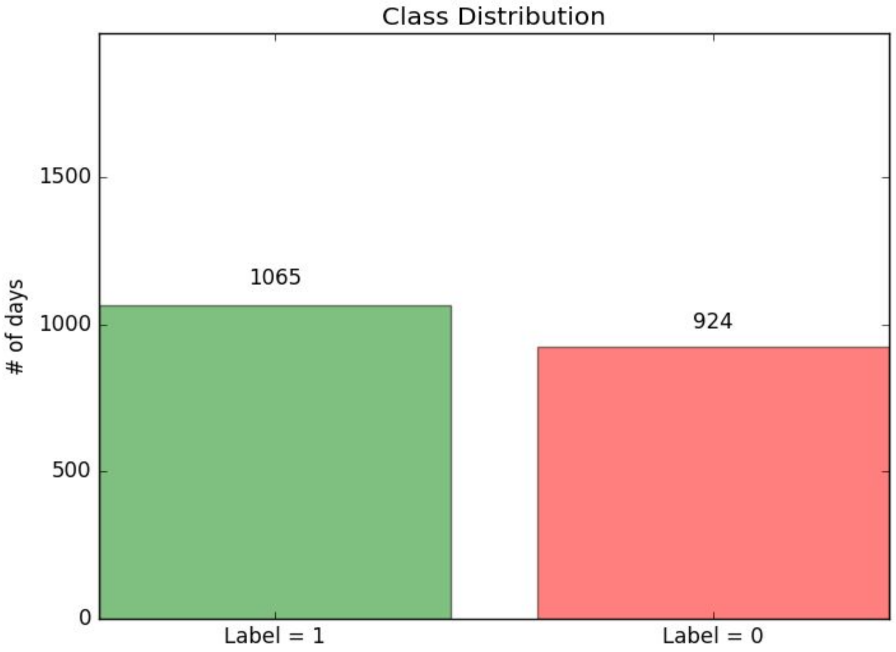
Exploration Time...(cont..)

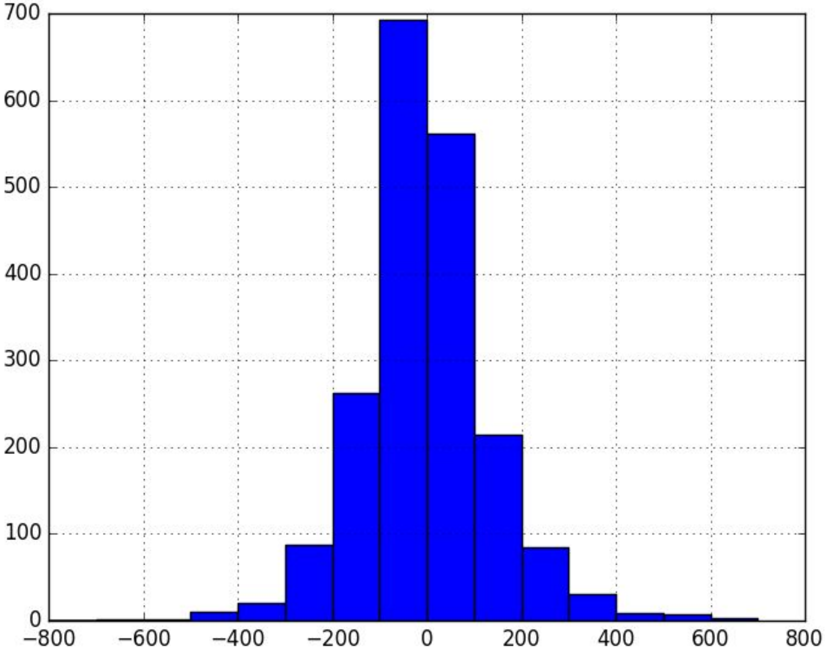
06/08/2008 -> 07/01/2016

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

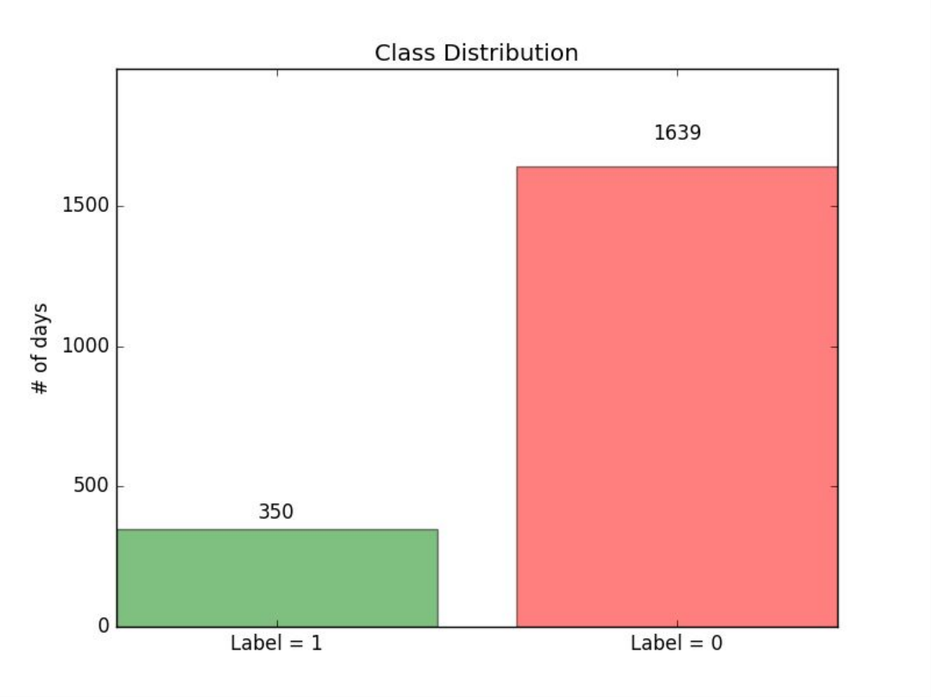
Can we beat it?!

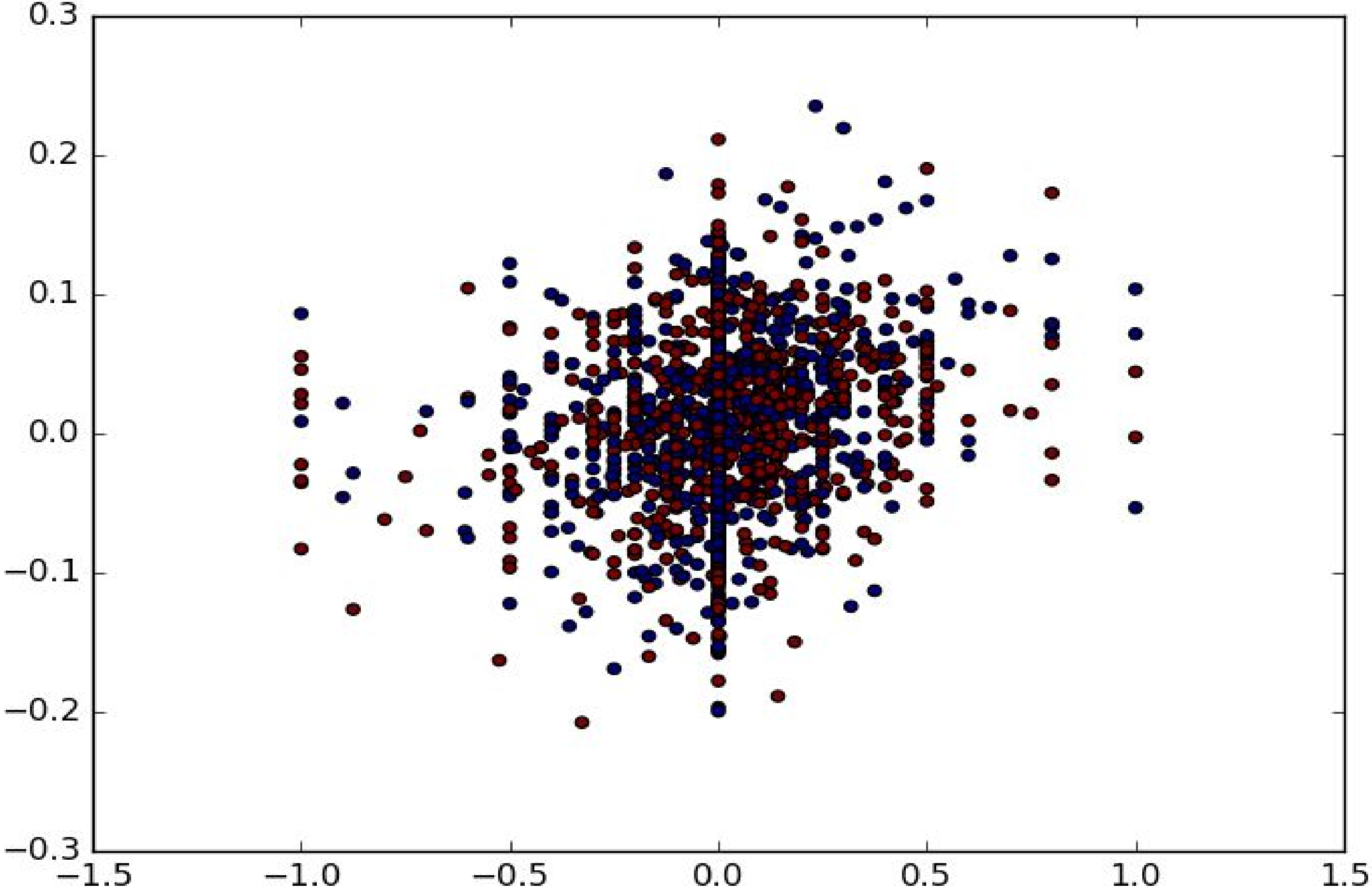
Distribution

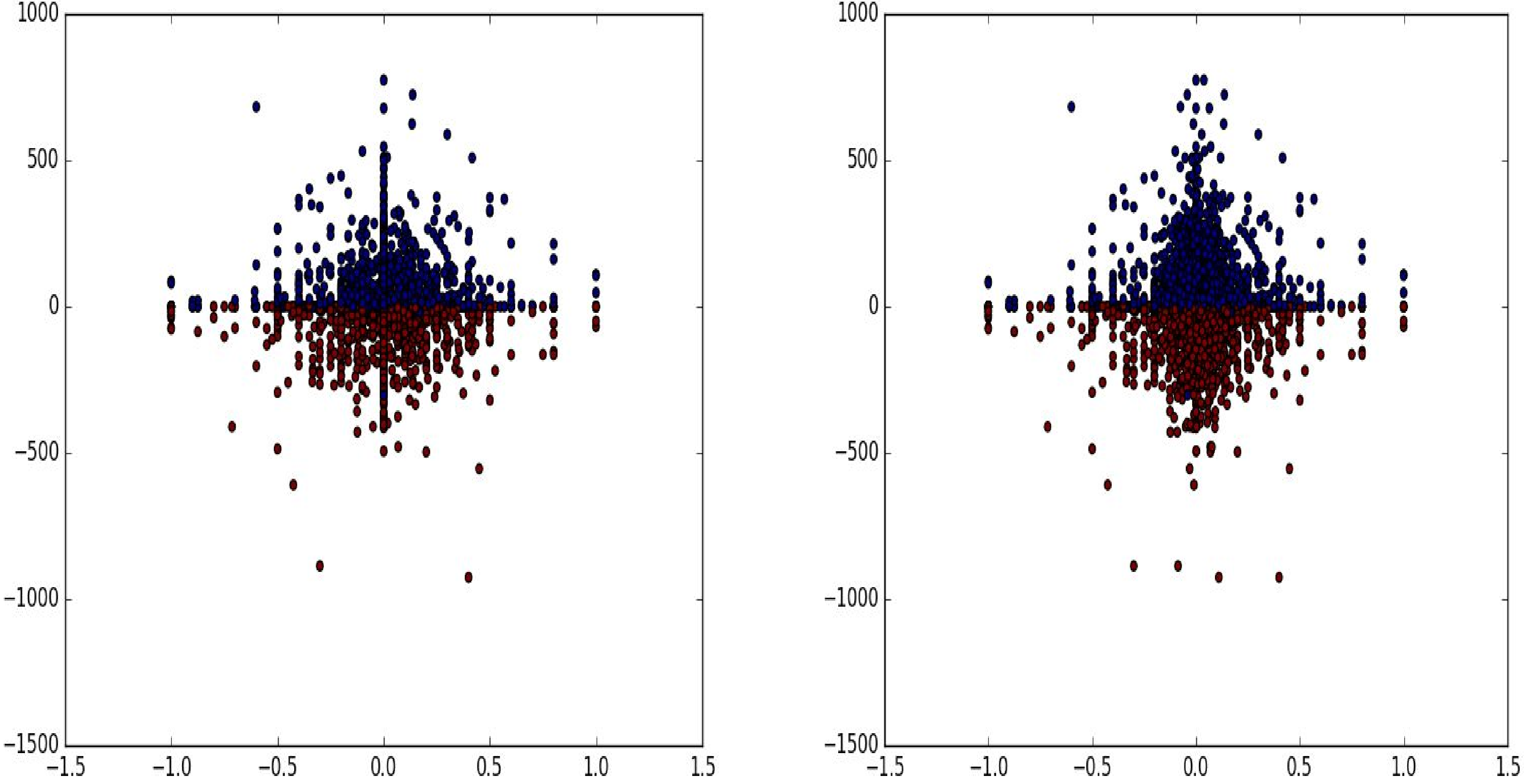




Distribution (cont.) - 100 Label




0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


1000
500
0
−500
−1000
−1500
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
1000
500
0
−500
−1000
−1500
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Workflow

Classifiers

- KNN
- AdaBoost
- DecisionTree
- RandomForest
- LogisticRegression
- SVC
- ExtraTrees
- BernoulliNB

Vectors

- Count Vector
- Count Vector of n = 2
- Tdidf Vectorizer
- Tdidf Vectorizer n =2

New Method - Ada Boosting

- short for "Adaptive Boosting", is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire who won the Gödel Prize in 2003
- popular boosting technique which helps you combine multiple “weak classifiers” into a single “strong classifier”. A weak classifier is simply a classifier that performs poorly, but performs better than random guessing

AdaBoosting in Scikit-learn

```
from sklearn.ensemble    import AdaBoostClassifier

Import pandas

Clf = AdaBoostClassifier(RandomForrestClassfier()),

def runModel(trainLines,testLines,train,test, label,key):
            clf.fit(trainLines, train[label])
            predictions = clf.predict(testLines)
            matrix = pandas.crosstab(test[label], predictions, rownames=["Actual"], colnames=["Predicted"])
```

Other New Terms Used

- ExtraTreesClassifer - Extremely Randomized Trees
- Truncated SVD - Singular Value Decomposition

We ran into problems...

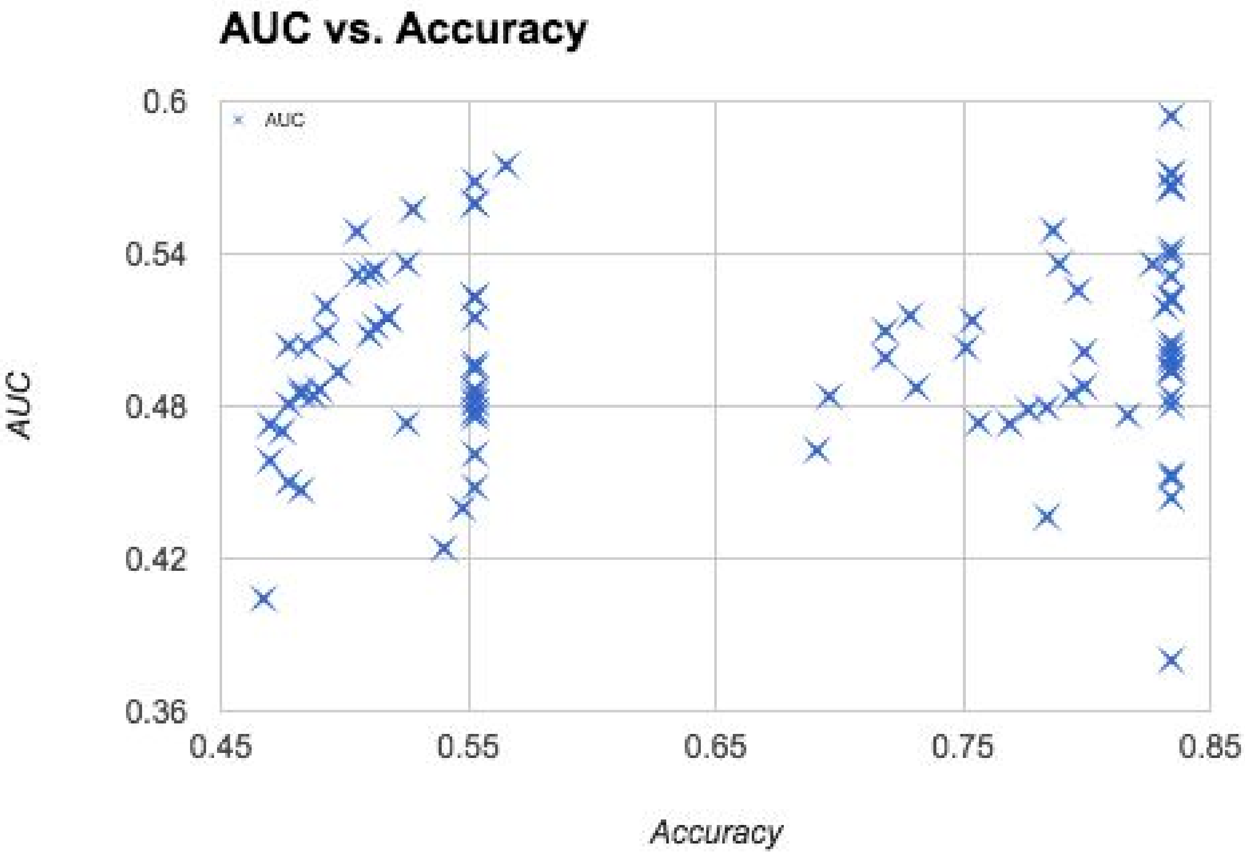
In our first or second checkpoint, we reported accuracies of 80% and better but..we split our training and testing data incorrectly..and some of our training data found its way into our test data as well.

Calculated AUC wrong!

The data was quite messy and Count Vectors and Td-idfVectors required stopword removal and stemming

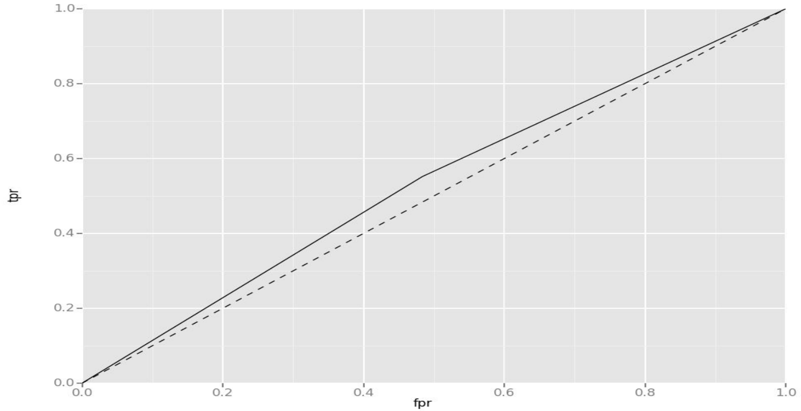
95% of words found when the stock market went down were also found in news articles when the stock market went up

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

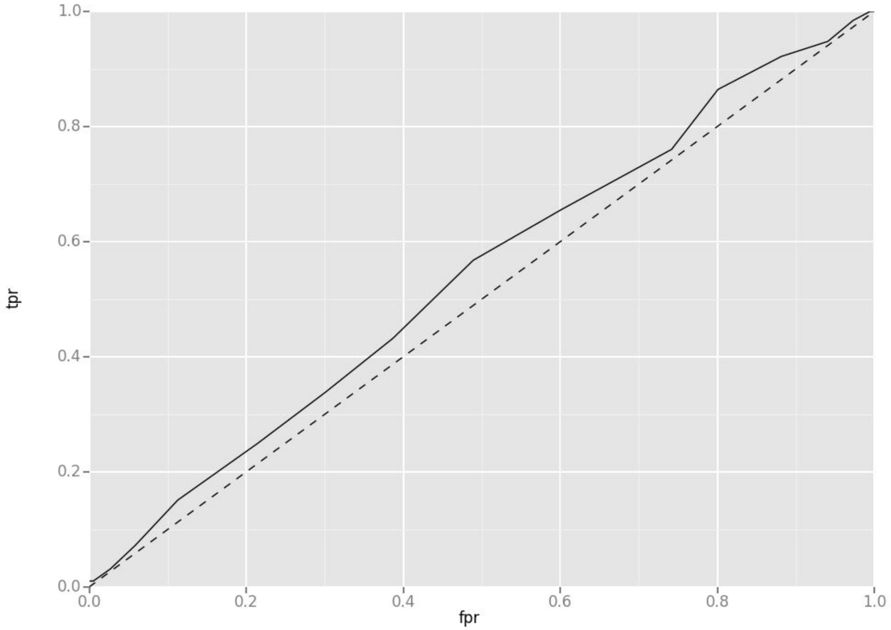

AUC vs. Accuracy
AUC
0.6
0.54
0.48
0.42
0.36
0.45
0.55
0.65
0.75
0.85
AUC
Accuracy

Our Best AUC Curves...

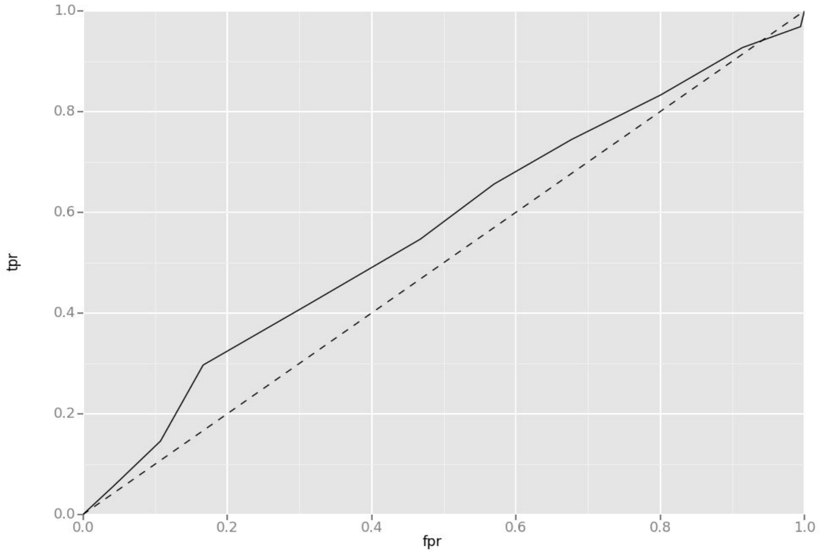
Roc Curve of DecisionTreeClassifier Accuracy(0.5344) with AUC of 0.5341



Roc Curve of KNeighborsClassifier Accuracy(0.5291) with AUC of 0.5396



Roc Curve of AdaBoostClassifier Accuracy(0.5397) with AUC of 0.5632



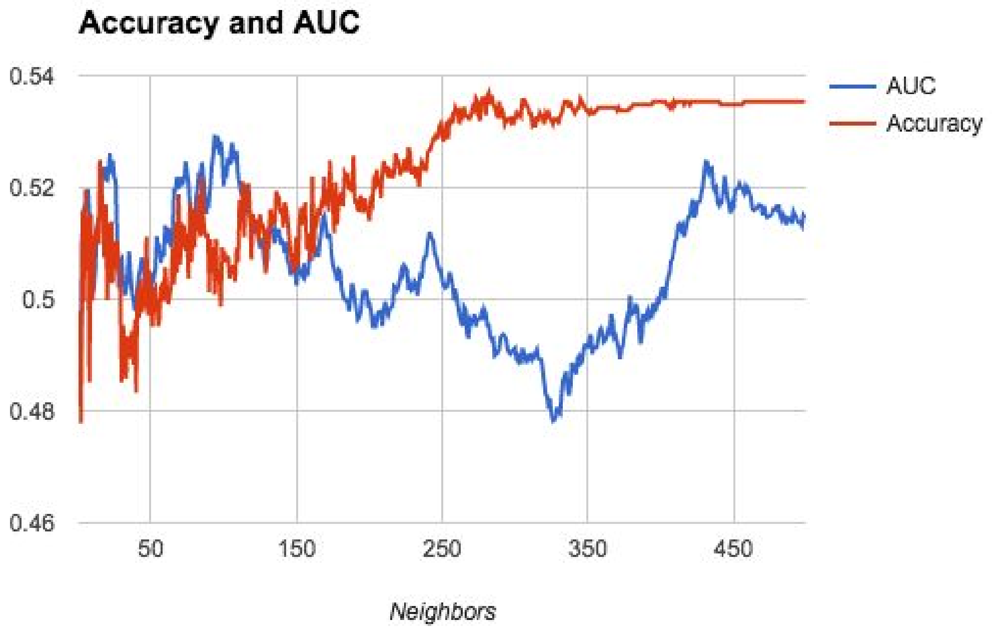
Cross Validation and Model Refining

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

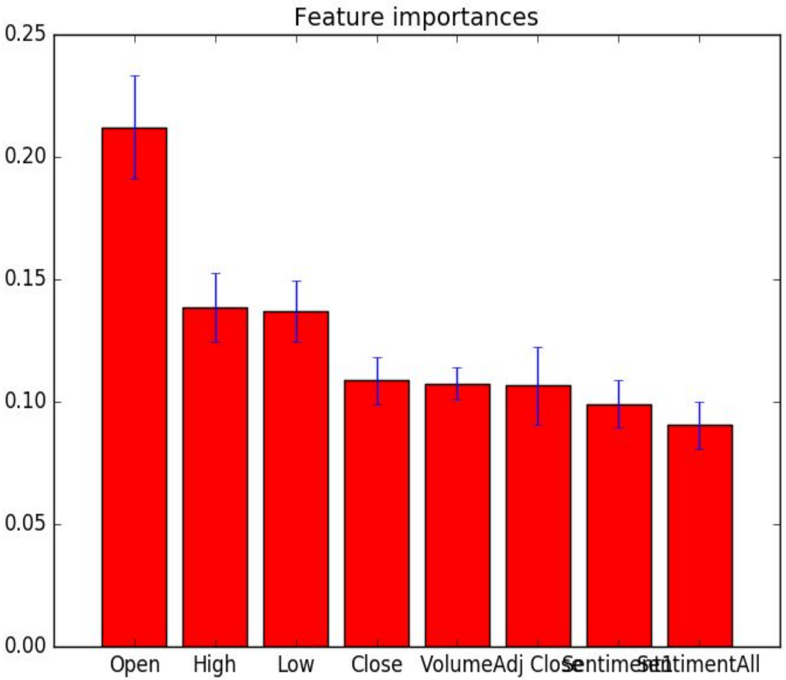
We concentrated on refining KNN. When we optimizing AUC, accuracy was neglected and vice-versa. We were not able to solidly get AUC above >.5 and accuracy above 53%

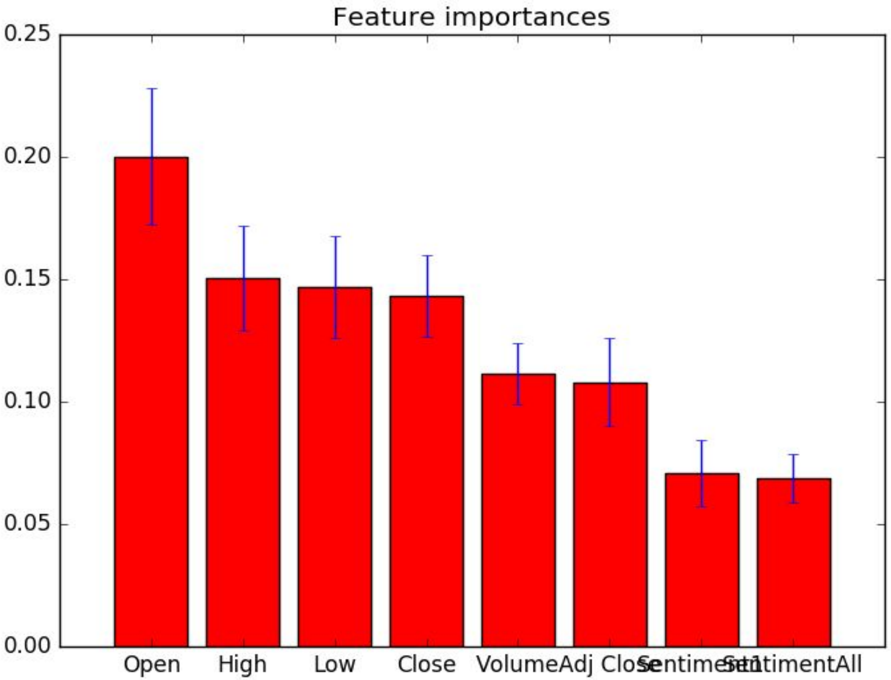
AUC < .6 is quite poor!

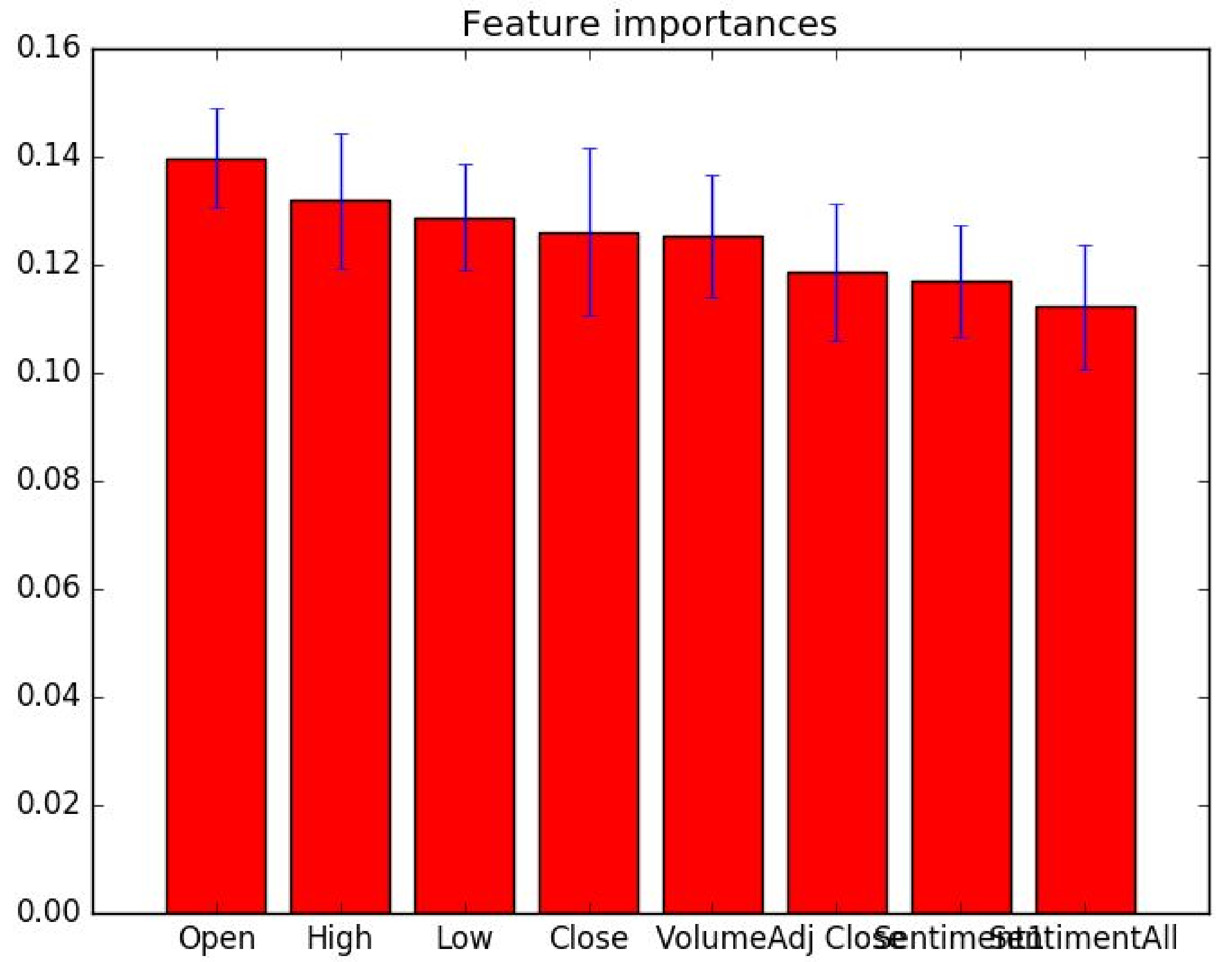
Refining KNN (k-fold k =5)

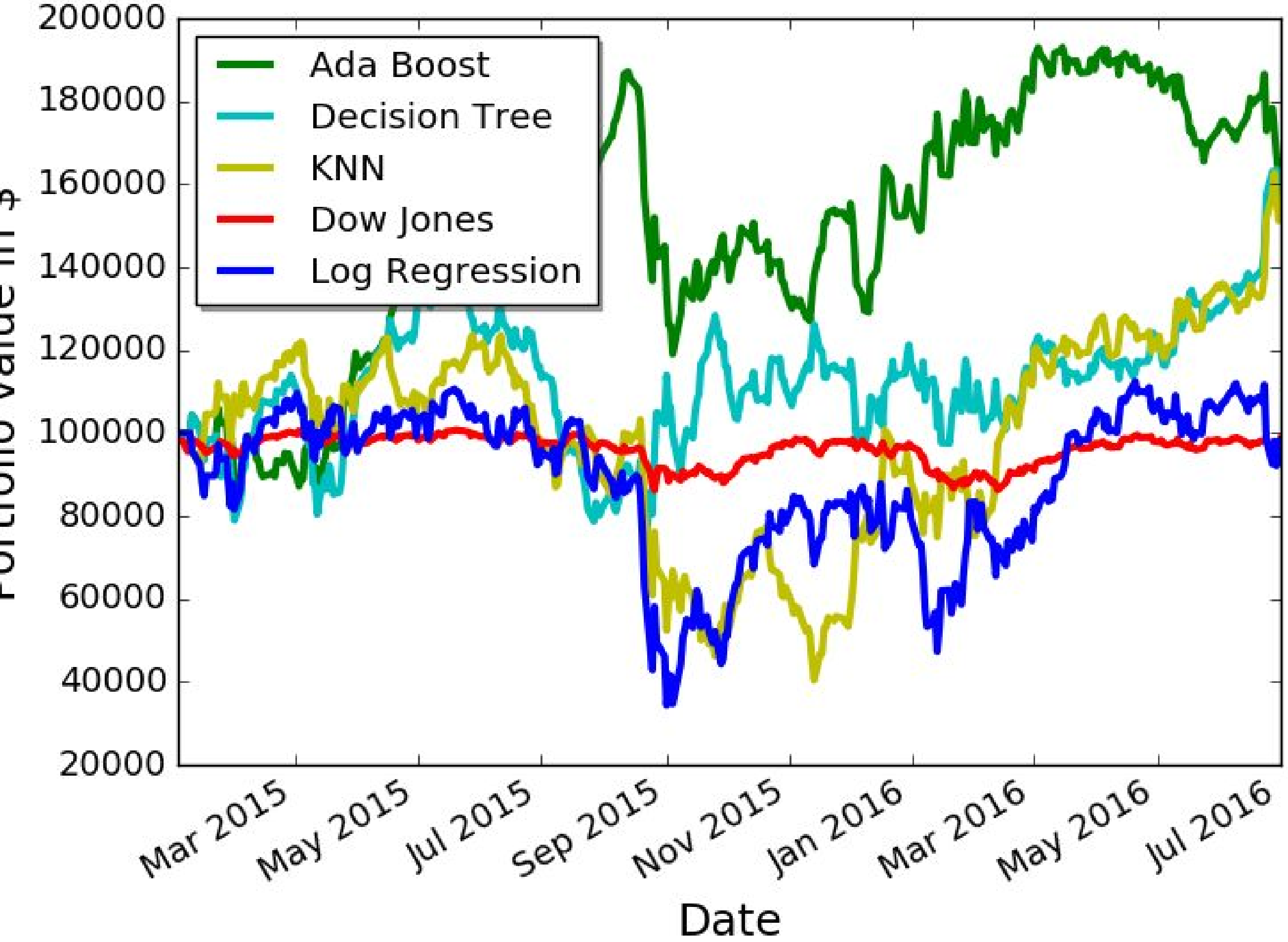


Feature Importance - ExtraTreesClassifier






Feature importances
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Open
High
Low
Close
Volume
Adj Close
Sentiment1
SentimentAll



AdaBoost - 62%

DT - 57.4%

KNN - 51.8%

LR   -1.8 %

Future Work

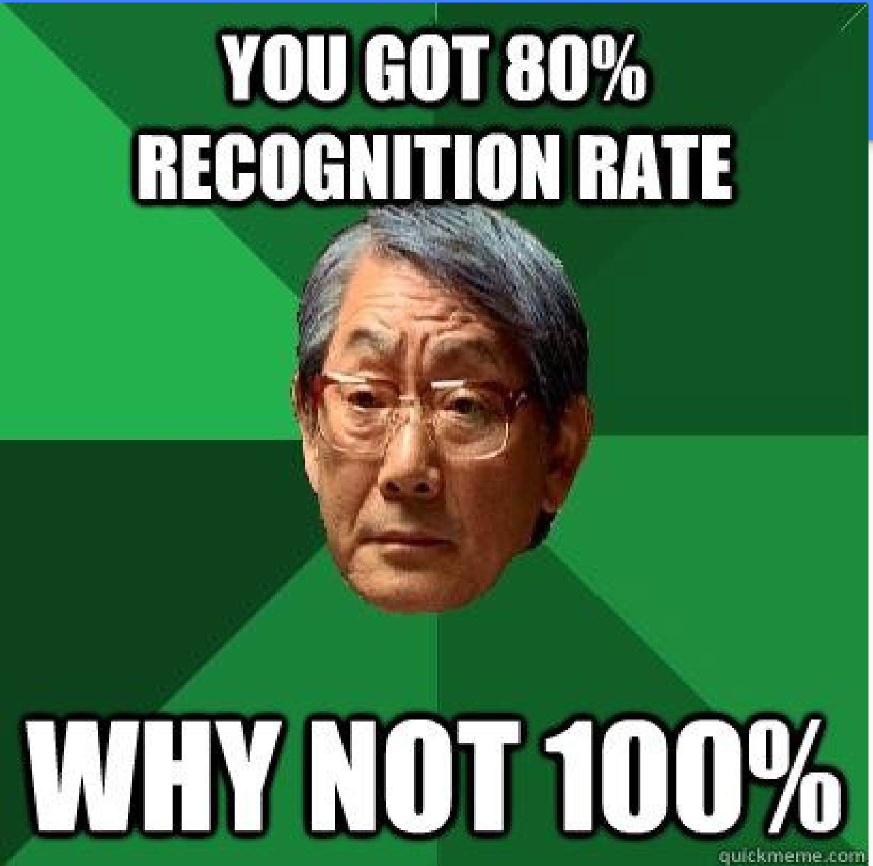
Try different text data or more specific data (research with surprisingly good accuracy in this domain) Correlating Financial Time Series with Micro-Blogging Activity

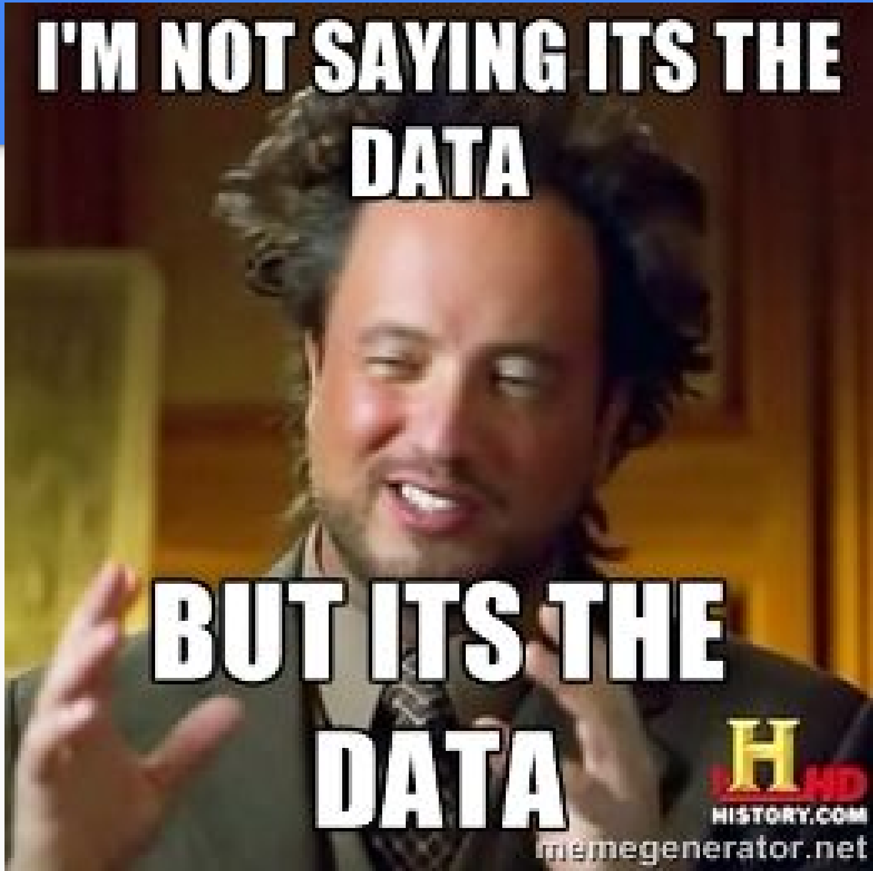
More, different vectorization

Incorporate numeric data with text data

More backtesting (maybe with Tomorrow LAbel and 100 Label)

Stanford Named Entity Recognizer


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?