Predict Stock Market Trend

Using News Headlines

Pei Guo

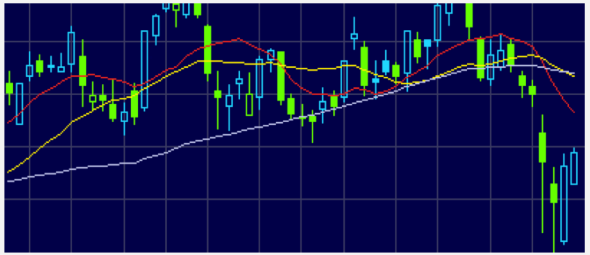
DATA

Daily News Headlines

8 years from 2008-2016

Dow Jones

Industrial Average(DJIA)
Same day close value



Targets	DJIA close value comparing with the previous day
0	Decrease
1	Rise or Stay the same

Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8
2008-08-08	0	b"Georgia 'downs two Russian warplanes' as cou...	b'BREAKING: Musharraf to be impeached.'	b'Russia Today: Columns of troops roll into So...	b'Russian tanks are moving towards the capital...	b"Afghan children raped with 'impunity,' U.N. ...	b'150 Russian tanks have entered South Ossetia...	b"Breaking: Georgia invades South Ossetia, Rus...	b"The 'enemy combatent' trials are nothing but...
2008-08-11	1	b'Why wont America and Nato help us? If they w...	b'Bush puts foot down on Georgian conflict'	b"Jewish Georgian minister: Thanks to Israeli ...	b'Georgian army flees in disarray as Russians ...	b"Olympic opening ceremony fireworks 'faked'"	b'What were the Mossad with fraudulent New Zea...	b'Russia angered by Israeli military sale to G...	b'An American citizen living in S.Ossetia blam...

Agenda

01 Data Preparation

02 Model Training

03 Deployment

04 Summary

05 Future Work

Data Cleaning

01
Data
Prep

```
'b"Georgia \'downs two Russian warplanes\' as countries move to brink of war"'
```

Punctuations
Stop words
Lower letters

```
georgia two russian warplane country move brink war
```

Word Embedding

01
Data
Prep

```
georgia two russian warplane country move brink war
```

Google News
Word2vec - Doc2vec

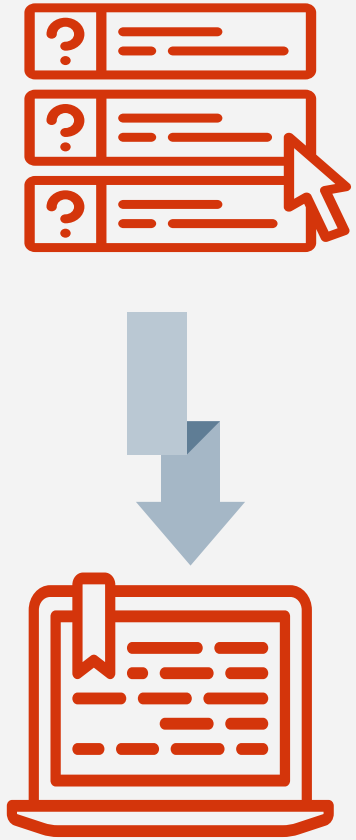
Bag of words
TF-IDF (character)
TF-IDF (word)

Vectors

```
-0.00466015,  0.04605767,  0.03547058,  0.10155483, -0.03422928,
-0.01389613,  0.0094356 , -0.14502455,  0.03542512,  0.0929451 ,
-0.02405647, -0.13372633, -0.0746136 ,  0.03855896, -0.08115924,
 0.10458709, -0.0641892 ,  0.0719087 ,  0.0019091 , -0.12026239,
 0.0486027 ,  0.02366476,  0.06977414, -0.02491679,  0.00629754,
 0.01696101, -0.09939197,  0.02952102,  0.05048709, -0.03987644,
 0.04234939, -0.02253801, -0.08623367, -0.02431879, -0.04120318,
-0.01384873, -0.00259539,  0.0541687 ,  0.06273022,  0.04681931,
 0.0330398 , -0.03522548,  0.1864641 ,  0.01297338,  0.03296788,
-0.02038486, -0.00516242, -0.0396328 , -0.10085952,  0.04830088,
```

```
[0., 0., 0., ..., 0., 0., 0.],
[0., 0., 0., ..., 0., 0., 0.],
[0., 0., 0., ..., 0., 0., 0.],
...,
[0., 0., 0., ..., 0., 0., 0.],
[0., 0., 0., ..., 0., 0., 0.],
[0., 0., 0., ..., 0., 0., 0.]]
```

news headlines vectors



Machine Learning Models	
Random Forest	Logistic Regression
XGBoost	Naïve Bayes
SVM	KNN
AdaBoost	EXTRA trees

02
Models

02
Models

Best Baseline Model

Character based
TF-IDF embedding
&
Naïve Bayes model

Overall F1 score: 0.56

Labels	Precision	Recall	F1 - score
0	0.5	0.35	0.41
1	0.61	0.74	0.67

Successfully predict 74% of market increase or no change

Of all the predictions of increase or no change, 61% are correct

Can we improve the score?

Topic modelling

Topic model F1: 0.53

HDP: Hierarchical Dirichlet Process	determine topic number: 20
LDA: Latent Dirichlet Allocation	generate topic distribution numbers

Baseline F1: 0.56

02 Models

```
 (2,
  '0.000*"korea" + 0.000*"israel" + 0.000*"china" + 0.000*"libya" + 0.000*"north" + 0.000*"go
vernment" + 0.000*"new" + 0.000*"year" + 0.000*"world" + 0.000*"killed"'),
 (3,
  '0.000*"russia" + 0.000*"china" + 0.000*"government" + 0.000*"new" + 0.000*"iran" + 0.000
*"world" + 0.000*"people" + 0.000*"america" + 0.000*"year" + 0.000*"russian"'),
 (4,
  '0.001*"israel" + 0.001*"new" + 0.001*"world" + 0.001*"year" + 0.001*"government" + 0.001
*"china" + 0.001*"police" + 0.001*"russia" + 0.001*"people" + 0.001*"war"'),
```

Can we improve the score?

Models on the cloud!

- Linear Learners
- XGBoost

02 Models

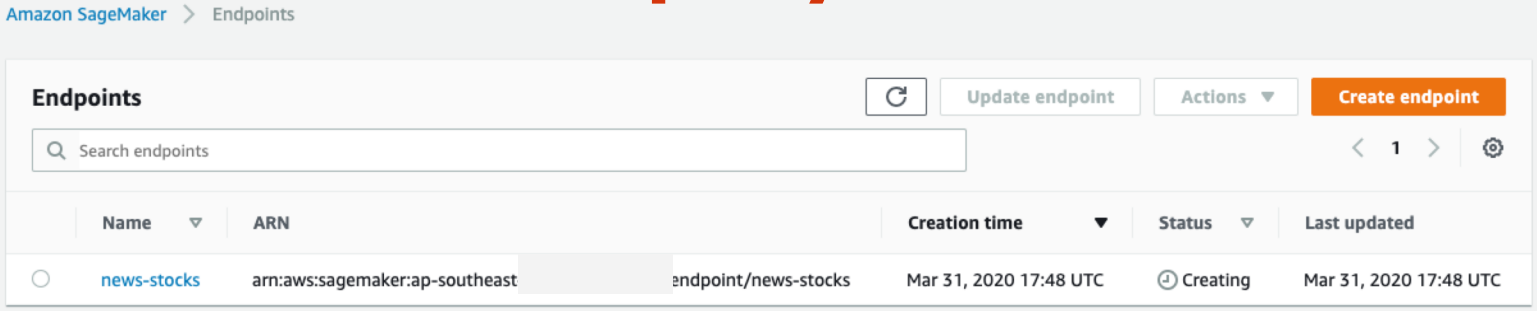
AWS F1: 0.52

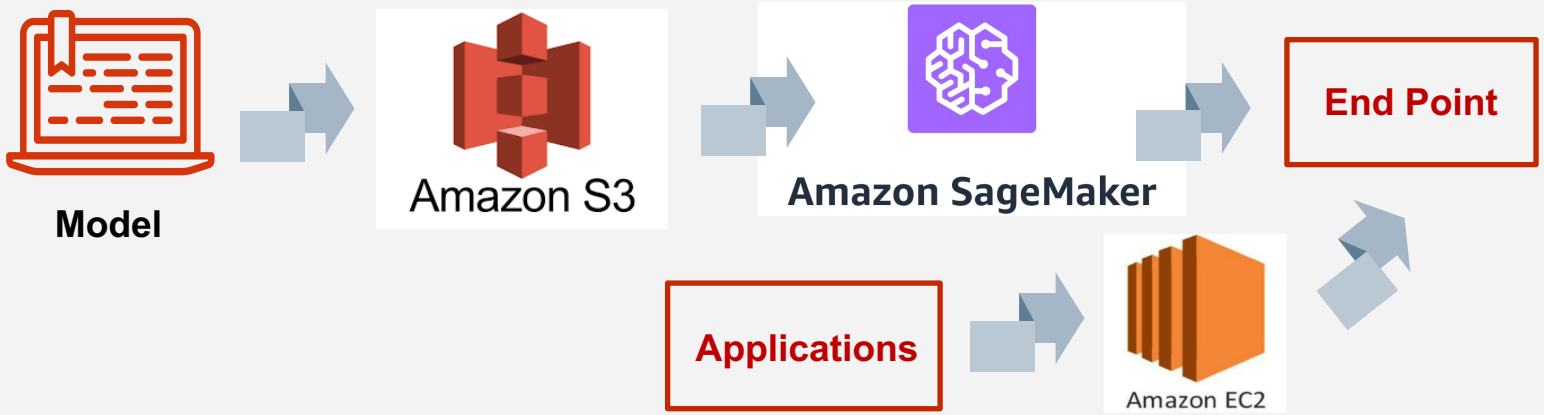
Baseline F1: 0.56

03 Deploy

How to make the model useful?

Deployment!





Predict stock market trend using news headlines ?

- End to end machine learning project
- The best F1 score 0.56
  Successfully predict 74% of market increase or no change
- Stock prices trend is a complex problem

04
Summary

05 Future Work

- Collect more data
- Try deep learning models
- Make the ML pipeline more automatic

Thank you!

Pei Guo

guopei123@gmail.com

linkedin.com/in/pei-guo

SCAN ME