MSBD 6000I - group project

Tencent Stock Opening Price Prediction

Hui Kwat Kong, Jay 20123133
Cheng Man Ho 20545628
Agniva Debnath 20647050

Contents

- Introduction (Horace)
- Related Work (Horace)
- Data Set (Jay)
- Data Transformation
  - BERT (Horace)
  - Sentiment analysis (Agniva)
  - Other Transformation (Jay)
- Modeling (Jay)
- Conclusion(Agniva)

Introduction

Stock Trend Prediction (Up / Down) by T+1 open price v.s. T closing price

Target Stock: Tencent

“Efficient-market hypothesis” => stock change based on new information

Traditional technical indicators + market news

Analysis of market news:

1) Sentimental analysis
2) BERT

Related Work

Trend: unigram -> sentiment analysis -> attention

On the Importance of Text Analysis for Stock Price Prediction. (Lee et al. 2014)

Sentiment analysis on social media for stock movement prediction. (Nguyen et al. 2015)

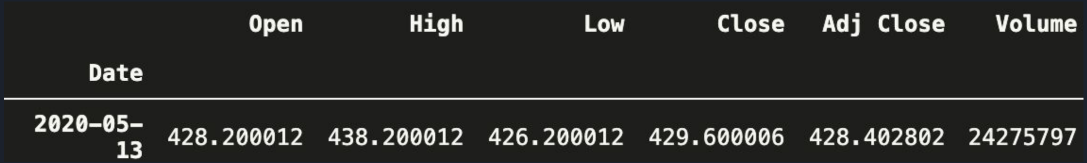
Attention-Based Event Relevance Model for Stock Price Movement Prediction. (Liu et al. 2017)

BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability. (Hiew et al. 2019)

Dataset

Two sources of data: (From Nov 2017 to Apr 2020, around 2.5 years)

- Yahoo Finance API to call the stock and HSI data:

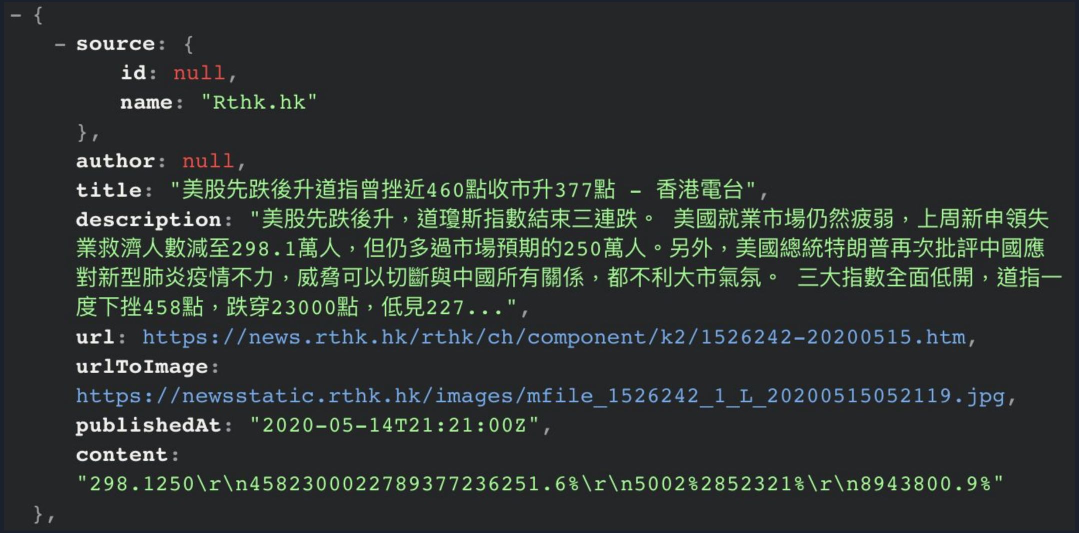


	Open	High	Low	Close	Adj Close	Volume
Date						
2020-05-13	428.200012	438.200012	426.200012	429.600006	428.402802	24275797

Dataset

Two sources of data: (From Nov 2017 to Apr 2020, around 2.5 years)

- Subscribe NewsAPI (every day Tencent news as it is very popular)



```
- {
    - source: {
          id: null,
          name: "Rthk.hk"
      },
      author: null,
      title: "美股先跌後升道指曾挫近460點收市升377點 - 香港電台",
      description: "美股先跌後升，道瓊斯指數結束三連跌。 美國就業市場仍然疲弱，上周新申領失
      業救濟人數減至298.1萬人，但仍多過市場預期的250萬人。另外，美國總統特朗普再次批評中國應
      對新型肺炎疫情不力，威脅可以切斷與中國所有關係，都不利大市氣氛。 三大指數全面低開，道指一
      度下挫458點，跌穿23000點，低見227...",
      url: https://news.rthk.hk/rthk/ch/component/k2/1526242-20200515.htm,
      urlToImage:
      https://newsstatic.rthk.hk/images/mfile_1526242_1_L_20200515052119.jpg,
      publishedAt: "2020-05-14T21:21:00Z",
      content:
      "298.1250\r\n4582300022789377236251.6%\r\n5002%2852321%\r\n8943800.9%"
  },
```

BERT

Google (2018) state-of-art NLP model

Attention with Transformer

BERT-Base Uncased option

Pre-trained by English on Wikipedia and BooksCorpus

“CLS” + [Title] + “SEP” + [Description] + “SEP” of the news as input

“Standard Tokenizer” => lowercase + greedy match for unseen word

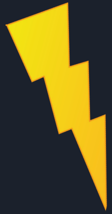
“TENCENT” => “TEN” + “##CENT” => “COMPANY”

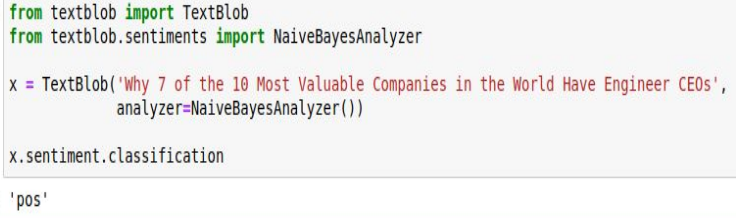
Pooled-output of dimension 768 then PCA (as 10 components)

Sentiment analysis

- TextBlob Sentiment Analyzer VS NLTK sentiment VADER

(Valence Aware Dictionary and sEntiment Reasoner)



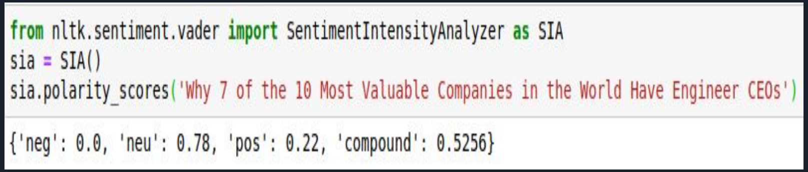


```
from textblob import TextBlob
from textblob.sentiments import NaiveBayesAnalyzer

x = TextBlob('Why 7 of the 10 Most Valuable Companies in the World Have Engineer CEOs',
             analyzer=NaiveBayesAnalyzer())

x.sentiment.classification
```

```
'pos'
```



```
from nltk.sentiment.vader import SentimentIntensityAnalyzer as SIA
sia = SIA()
sia.polarity_scores('Why 7 of the 10 Most Valuable Companies in the World Have Engineer CEOs')
```

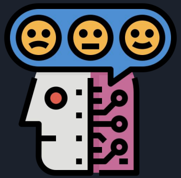
```
{'neg': 0.0, 'neu': 0.78, 'pos': 0.22, 'compound': 0.5256}
```

- Analyser : Naive Bayes Analyzer
- Corpus: Movie Reviews
- Slow

- Corpus: Social Media Text
- Fast

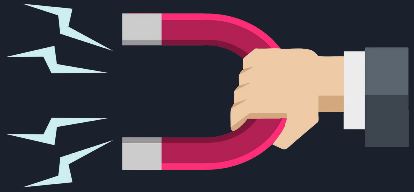
Scope of Improvement: Domain specific Corpus, Combine Twitter data

Sentiment Analysis

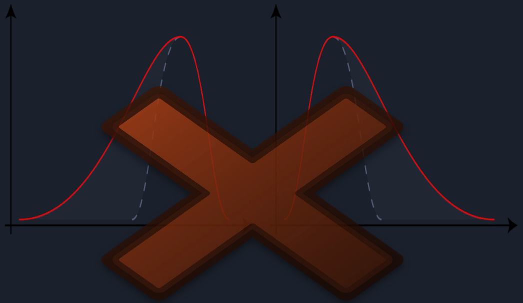


Method :

1. [Title] + [Description] + [Content]
   - Fetch polarity scores for Each Component
   - Aggregate for Each Day



2. Normalise
   - Removes biases for days with more number of news articles



Other Transformations

1. Aggregate the different transformed news from BERT and Sentiment analysis into daily format (mean, sum)

2. Generate the technical indicators (such as: Relative Strength Index (RSI))

3. Create Holiday, T-1, T-2, T-3, T-4 inputs by shifting our dataset

4. Create day, weekday, month to check seasonal effects

Modeling Experiments

Set-up:

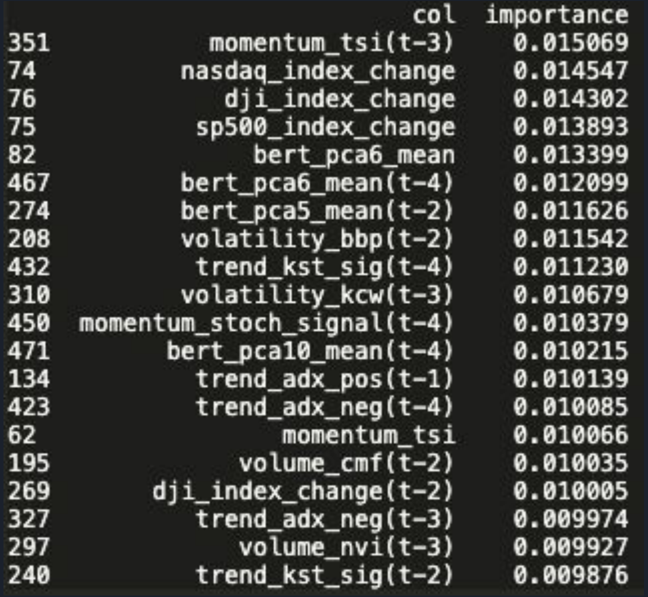
1. Scoring criteria: roc auc score (due to unbalanced size)
2. Selecting hype-parameter by Randomized search
3. Cross validation with stratified by the target

Modeling Experiments

Metrics		SVM	Voting (SVM+XG Boost)	LSTM	XGBoost (With text information)
ROC AUC	Training	1	0.799	0.743	0.858
	Testing	0.577	0.723	0.633	0.738
Accuracy	Training	1	0.702	0.671	0.781
	Testing	0.612	0.669	0.633	0.736

Modeling Experiments

Importance features in XGBoost:

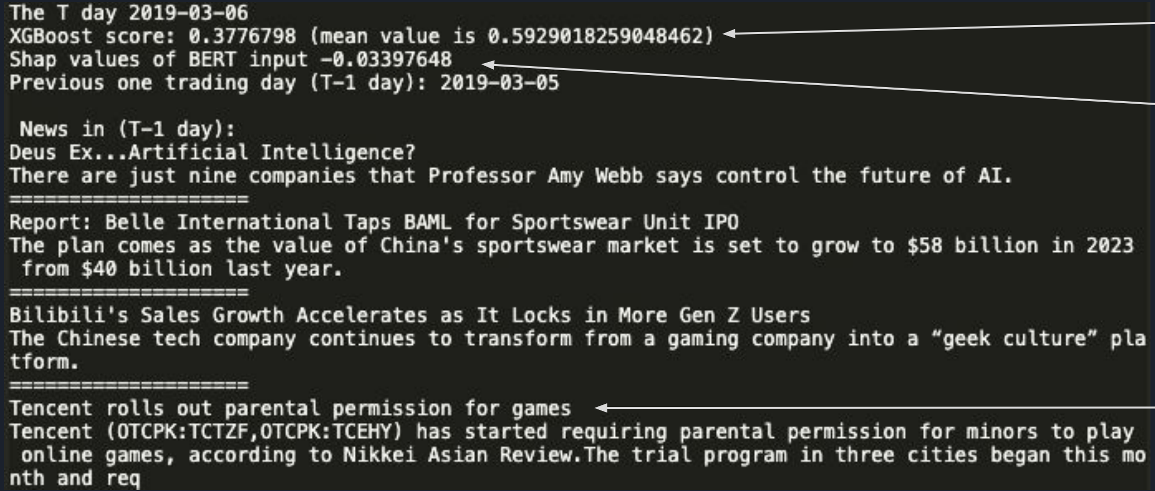


```
                         col  importance
351           momentum_tsi(t-3)    0.015069
74          nasdaq_index_change    0.014547
76             dji_index_change    0.014302
75           sp500_index_change    0.013893
82               bert_pca6_mean    0.013399
467         bert_pca6_mean(t-4)    0.012099
274         bert_pca5_mean(t-2)    0.011626
208         volatility_bbp(t-2)    0.011542
432          trend_kst_sig(t-4)    0.011230
310         volatility_kcw(t-3)    0.010679
450  momentum_stoch_signal(t-4)    0.010379
471        bert_pca10_mean(t-4)    0.010215
134          trend_adx_pos(t-1)    0.010139
423          trend_adx_neg(t-4)    0.010085
62                 momentum_tsi    0.010066
195             volume_cmf(t-2)    0.010035
269       dji_index_change(t-2)    0.010005
327          trend_adx_neg(t-3)    0.009974
297             volume_nvi(t-3)    0.009927
240          trend_kst_sig(t-2)    0.009876
```

Shap individual checking

“Shap”, a python package, aims to explain the output of machine learning models by a game theory approach. TreeExplainer in shap can explain feature importance in individual prediction.

What does the input news look like for most negative shap values? (it means it should have very negative news for Tencent and its opening price should go down)



```
The T day 2019-03-06
XGBoost score: 0.3776798 (mean value is 0.5929018259048462)
Shap values of BERT input -0.03397648
Previous one trading day (T-1 day): 2019-03-05

 News in (T-1 day):
Deus Ex...Artificial Intelligence?
There are just nine companies that Professor Amy Webb says control the future of AI.
====================
Report: Belle International Taps BAML for Sportswear Unit IPO
The plan comes as the value of China's sportswear market is set to grow to $58 billion in 2023
 from $40 billion last year.
====================
Bilibili's Sales Growth Accelerates as It Locks in More Gen Z Users
The Chinese tech company continues to transform from a gaming company into a “geek culture” pla
tform.
====================
Tencent rolls out parental permission for games
Tencent (OTCPK:TCTZF,OTCPK:TCEHY) has started requiring parental permission for minors to play
 online games, according to Nikkei Asian Review.The trial program in three cities began this mo
nth and req
```

Very low likelihood

Negative shap values in text columns

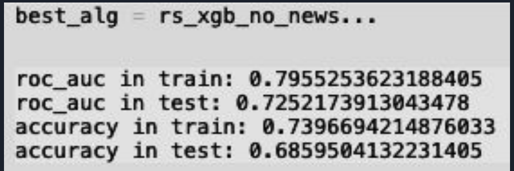
Setting parental permission is very harmful for gaming business in Tencent

Conclusion

- XGboost has the best modelling result with 73.6% accuracy, it is because boosting is a powerful algorithm and it can sufficiently handle the over-fitting issue by early stopping

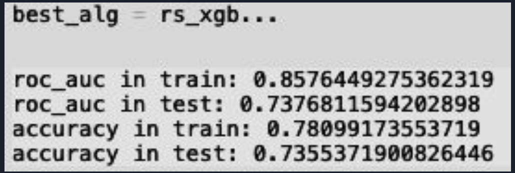
- XGboost performance with and without news
  - Without News

    With News



```
best_alg = rs_xgb_no_news...

roc_auc in train: 0.7955253623188405
roc_auc in test: 0.7252173913043478
accuracy in train: 0.7396694214876033
accuracy in test: 0.6859504132231405
```



```
best_alg = rs_xgb...

roc_auc in train: 0.8576449275362319
roc_auc in test: 0.7376811594202898
accuracy in train: 0.78099173553719
accuracy in test: 0.7355371900826446
```

Limitations

1. Limited news data
   - We have data from Nov 2017 to Apr 2020, it is better to extract as many years
   - Desirable to have Chinese news and English news. Engine Capable to combine them together

2. Sentiment analysis and BERT encoding is just an unsupervised approach
   - Domain specific Pretraining and Corpus

3. Unprecedented circumstances
   - Covid-19

Thank You

Thank You for being here today.

We really appreciate that you took the time to be here and listen to our presentation.

Special thanks to Prof. Zhang Chen for his teaching guidance and support.