Kaggle : Two Sigma

Using News to Predict Stock Movement

Anirban Sen
Charu Sreedharan
Tigmanshu Chaudhary

Introduction

- Live Kaggle competition hosted by Two Sigma
- Market data provided by Intrinio and News data provided by Thomson Reuters
- Training Data: (2007-2016) Test Data: (Jan 2017 - July 2018)
- Aim of the competition: To predict whether each stock will have increased or decreased value ten trading days into the future based on news information.

Data

- Market Data : 9 Million records
- News Data : 4 Million records
- Target value : ‘returnsOpenNextMktres10’ (market-residualized return 10 days into the future)

Understanding Market data

	volume	close	open	returnsClosePrevRaw1	returnsOpenPrevRaw1	returnsClosePrevMktres1	returnsOpenPr
count	4.072956e+06	4072956.00	4072956.00	4072956.00	4072956.00	4056976.00	4056968.00
mean	2.665312e+06	39.71	39.71	0.00	0.01	0.00	0.01
std	7.687606e+06	42.29	42.61	0.04	7.08	0.03	6.97
min	0.000000e+00	0.07	0.01	-0.98	-1.00	-1.24	-615.85
25%	4.657968e+05	17.25	17.25	-0.01	-0.01	-0.01	-0.01
50%	9.821000e+05	30.30	30.29	0.00	0.00	-0.00	-0.00
75%	2.403165e+06	49.86	49.85	0.01	0.01	0.01	0.01
max	1.226791e+09	1578.13	9998.99	45.59	9209.00	45.12	8989.21

- Discrepancies in the market data observed

Contd.

	time	assetCode	assetName	volume	close	open	returnsClosePrevRaw1	returnsOpenPrevRaw1	retu
627547	2008-09-29 22:00:00+00:00	BK.N	Bank of New York Mellon Corp	18718479.0	26.5	3288.1136	-0.271578	99.125262	-0.0
1127598	2010-01-04 22:00:00+00:00	TW.N	Towers Watson & Co	223136.0	50.0	9998.9900	-0.058470	185.988360	-0.0

- The data row with extreme openvalues. (Great than 1600)

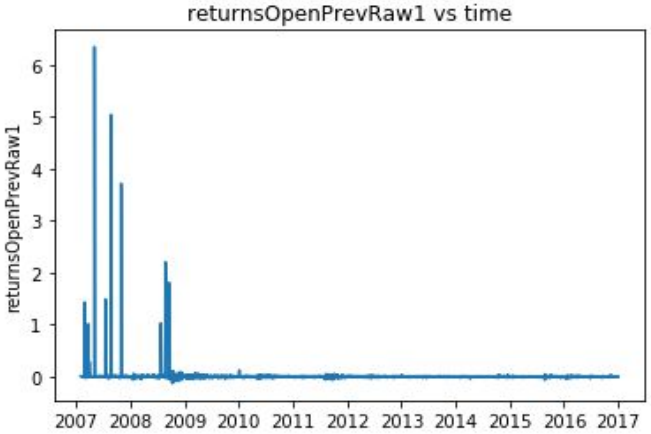
- The difference between raw return and market adjusted returns have some extreme values.

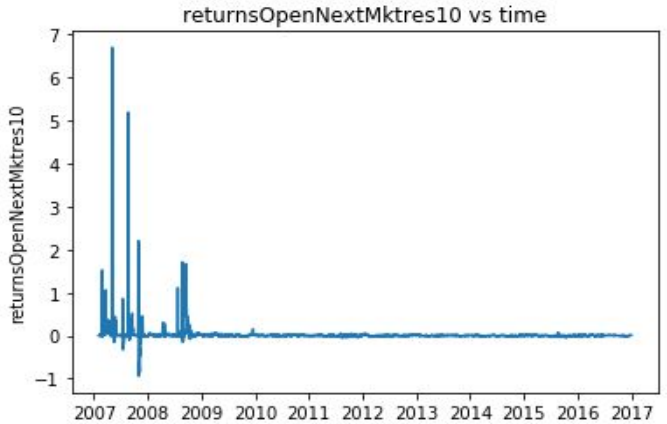
```
Check null data:

time                          0
assetCode                     0
assetName                     0
volume                        0
close                         0
open                          0
returnsClosePrevRaw1          0
returnsOpenPrevRaw1           0
returnsClosePrevMktres1   15980
returnsOpenPrevMktres1    15988
returnsClosePrevRaw10         0
returnsOpenPrevRaw10          0
returnsClosePrevMktres10  93010
returnsOpenPrevMktres10   93054
returnsOpenNextMktres10       0
universe                      0
dtype: int64
```

- Null values observed in market-adjusted columns.

- Possible Solution: Replace them with their respective raw data.

Contd.





- The returnsOpenPrevRaw1 & returnsOpenNextMktres10 parameters behave homogeneously after 2009, but is not so before 2009.

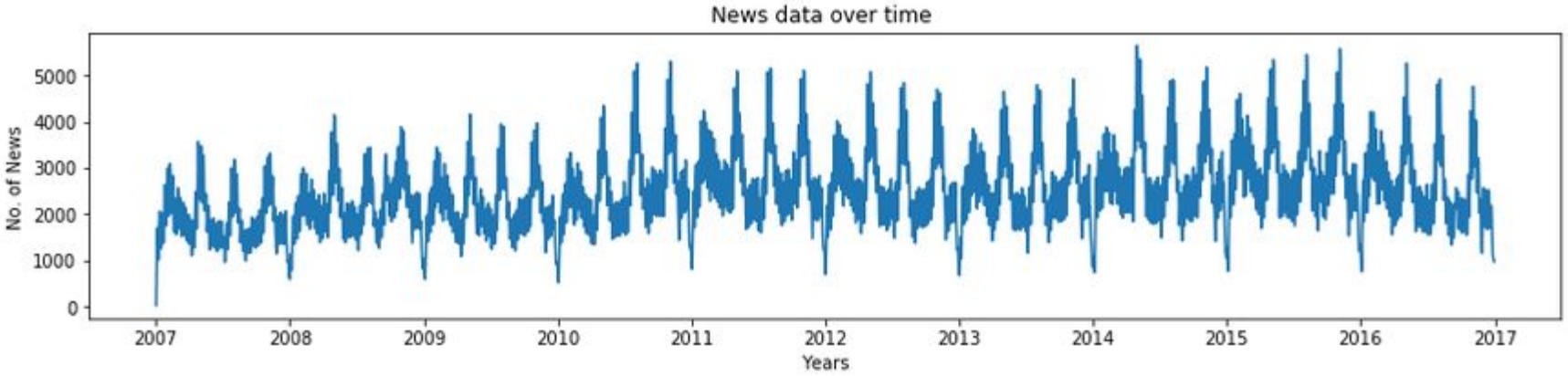
- Possible Solution: Delete data prior to 2009

Understanding News data

	urgency	takeSequence	bodySize	companyCount	sentenceCount	wordCount	firstMentionSentence	relevance	sen
count	9328750.00	9328750.00	9328750.00	9328750.00	9328750.00	9328750.00	9328750.00	9328750.00	932
mean	2.32	2.12	3768.92	5.03	22.51	580.43	4.82	0.74	0.0
std	0.95	2.94	7475.65	8.79	36.02	958.06	12.18	0.38	0.8
min	1.00	1.00	0.00	1.00	1.00	1.00	0.00	0.00	-1.0
25%	1.00	1.00	0.00	1.00	1.00	20.00	1.00	0.35	-1.0
50%	3.00	1.00	1571.00	1.00	10.00	259.00	1.00	1.00	0.0
75%	3.00	2.00	4504.00	5.00	30.00	765.00	2.00	1.00	1.0
max	3.00	97.00	122770.00	43.00	1205.00	20263.00	989.00	1.00	1.0

- Multiple asset codes for single news data observed

Contd.



- 4 peaks in no. of news article observed each year

- A fall in no. of news articles during the end of the year

Data Engineering

- Drop Dataframe columns which are not relevant.
- Sanitizing Data and encoding Categorical values.
- Normalize the News dataframe.
- Group news items having same ‘date’ and ‘assestCode’.
- Merging the News and Market Datasets and cleaning the resultant Data frame

Issues Faced

- Discrepancies in data set
- Issues while merging the 2 data sets
- Frequent kernel crashes during cross validations as data set is very large

Models Applied on Validation Datasets

- Linear Discriminant Analysis: **0.538**
- Quadratic Discriminant Analysis: **0.512**
- AdaBoost(n_estimators=300): **0.541**
- Logistic Regression (c=1, penalty=l2) : **0.53**
- XGBoost(n_jobs=4,n_estimators=2000,max_depth=8,eta=0.1) : **0.60**

XGB

- Very popular model when it comes to Kaggle competitions and is able to handle both classification and regression problems

- eXtreme Gradient Boosting like other boosting algorithms tries to successively improve weak learners.

- Stopping Criteria for XGB

- How does the model choose splits

Result

- Final submission score of **0.57**

- Currently the best score on the competition leaderboard is **0.78**

Future Improvements

- Implement cross validation on the models on the whole dataset.
- Closer study of stocks and variation in their prices due to splits and acquisition.
- Other models eg LGBM