**Two Sigma: Using news to predict stock movements**

# Introduction

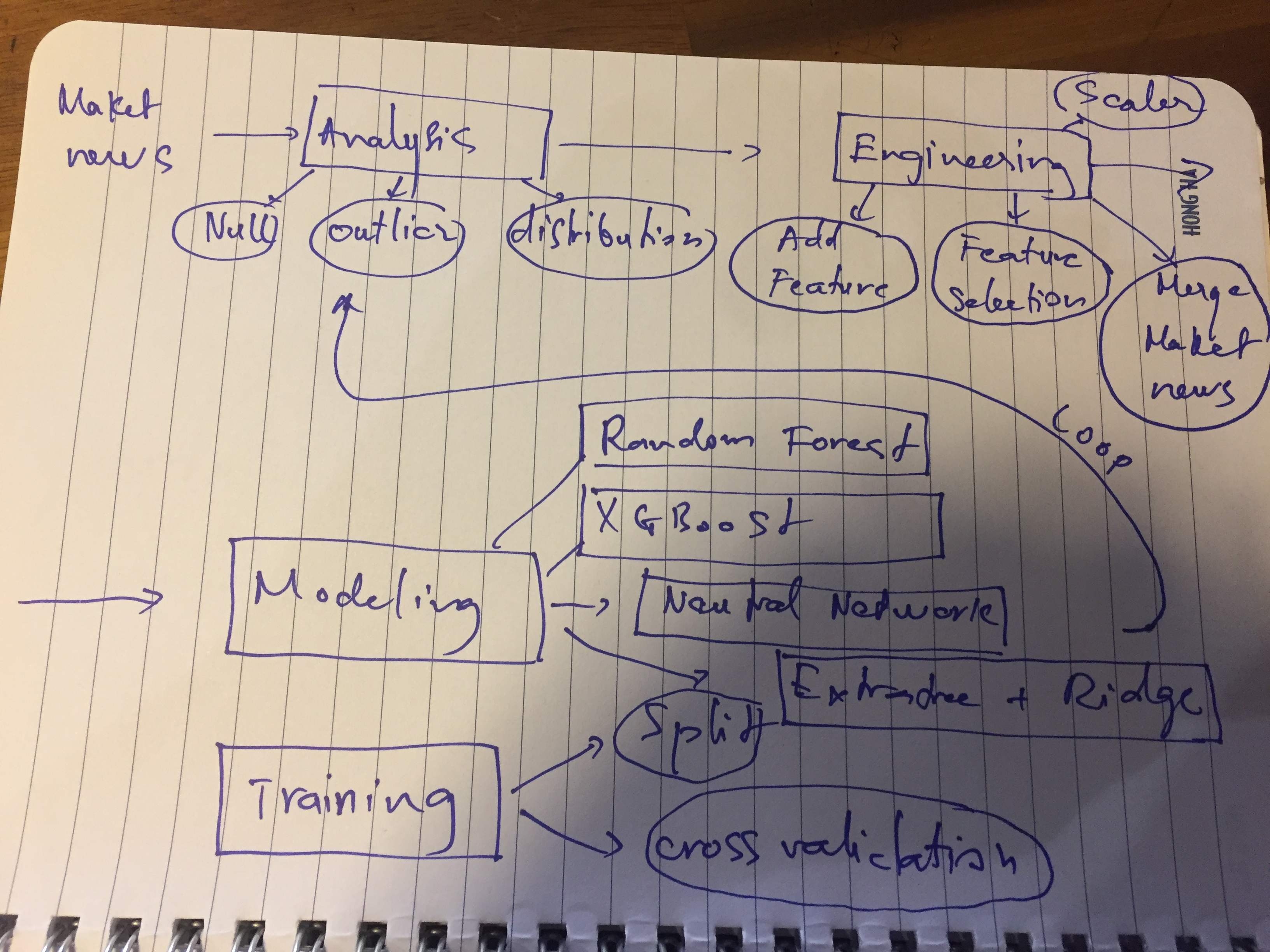
This project is a kaggle competition that hosted by Two Sigma, the purpose of this project is to find a solution for predicting stock movements using news and market data. The competition use sigma score as evaluation metric to rank the result of the competition.

# Process

We follow process like this:

Try stage of data analysis and data engineering fast to build simple model to make prediction and compute accuracy along with sigma score.

After measure the result we repeat the steps again, we pay more time on



# Data analysis

## Market data

About the market data description the details is on Kaggle. In general, each row of market data is a market history information of a stock (instrument) for a date include volume, open, close price previous change price of previous 10 days and 1 day of open and close, the same for next 1 and 10 days.

The market data is from 2007 to 2016

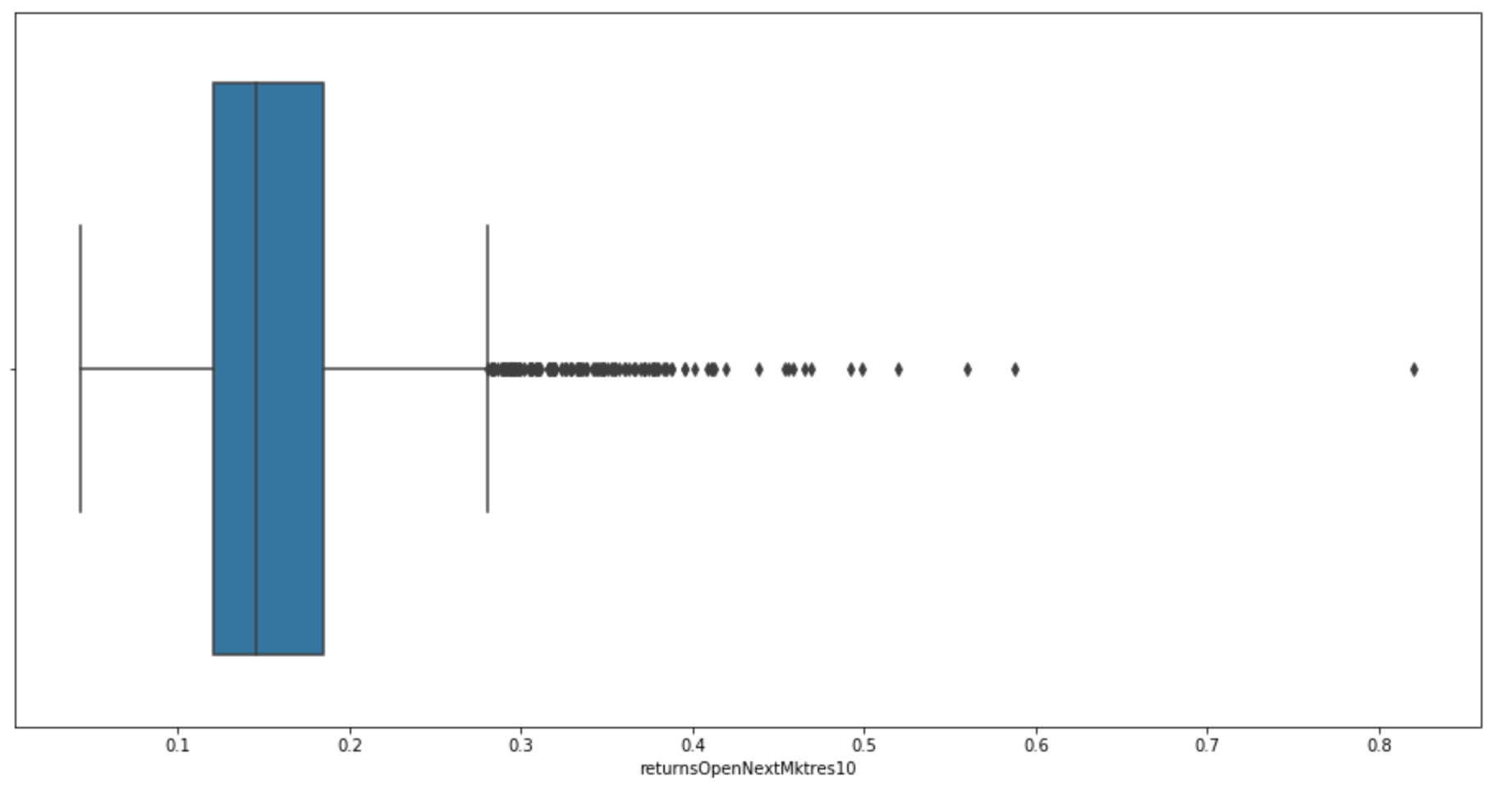
*Price of 5 random assets*



*returnOpenNextMktres10 of 5 random assets*



*Outlier of returnOpenNextMktres10*

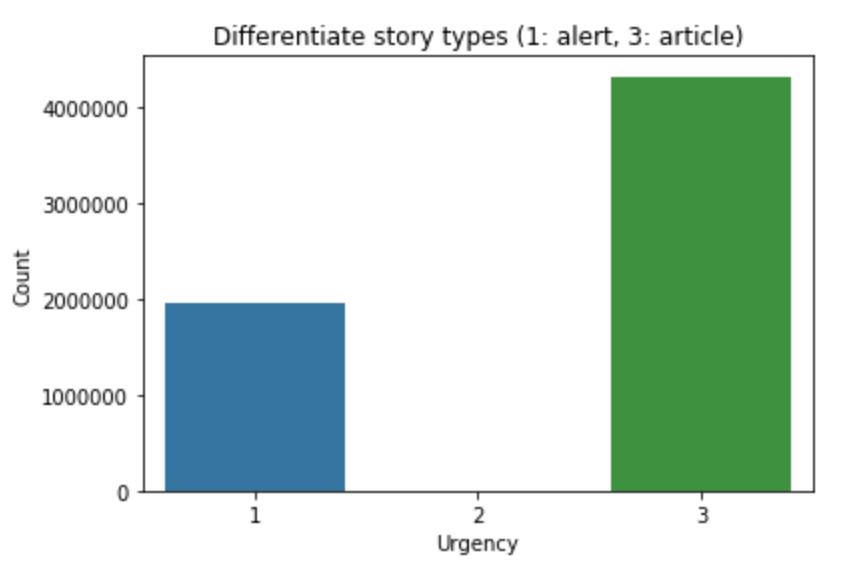


## News data

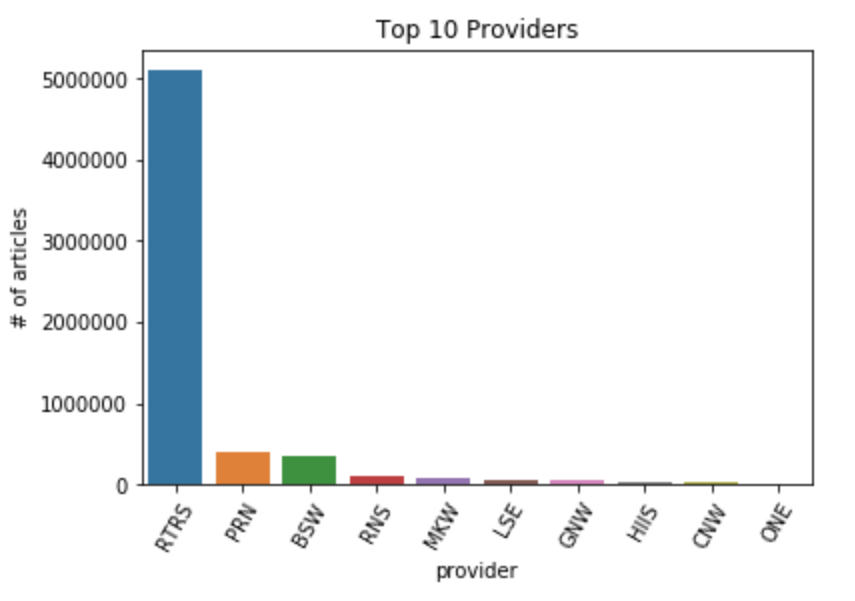
The details description about news is data on Kaggle. For each date, the news is provided by providers (Thomson for example). And that news related to some stocks, there are some important information like sentiment score and relevance score provided by two sigma.

Another field is firstMentionSentence, the good news for stock if it is mentioned at headline or beginning of the story body

New base on urgency



Providers:



## Evaluation

For each day t, we calculate:

xt = sum of all stock(y\_hat\*r\*u) (1)

Inside ( and ) is a expression used to calculate for one particular stock

After that the score is calculated by mean divided by standard deviation of all daily xt

Score = mean(xt)/ standard deviation(xt) (2)

In (1), y\_hat is confident value that we have to predict

r as the return of next 10 days that provided by dataset

u is 0 or 1 indicate that the stock is calculated score or not

Data of market is very noisy so using regression to predict is not good so we choose classification for stock prediction.

From returnOpenNextMktres10 (10 day market-residualized return) we label data

if returnOpenNextMktres10 larger than 0 label 1 else label 0

The problem becomes the binary classification that to classify up (1) or down (0) base on the probability of prediction output in range [0,1] we map it to [-1,1] (y\_hat = y\_prob\*2 - 1). The purpose of mapping from [0,1] to [-1,1] is completion requirement of prediction range

So there are 4 cases: we predict right predict right direction it mean (y\_hat>0 and r>0) or (y\_hat < 0 and r<0) the xt will be larger than 0 (we have revenue that day t) otherwise we have a lost at that day t

Assume we predict right direction and value r is big but y\_hat is very small so the x(t) will be small so the result is not good

# Data engineering

## Market data

Handle missing data

Add Technical Analysis data:

We are strongly believe that TA is very helpful for trader to make trade decision so adding these TA features make model more predictive capacity

There are some TA feature that we use: Moving average, Log Return, 10 day percentage change, momentum and volatility

## News data

# Modeling

Data of stock is very noisy so we will use model can catch the weak signal of market using simple model random forest, logistic regression, linear classifier like Ridge.

To know how marker or news effect to prediction we build model that just use market data and news data. The result of using market data is about 0.54 (random forest) and news is about 0.5 (random forest). It mean that news seem to be not effect to our prediction confidence value. The result is disappointed because news is not important as market data.

After marge market data and news data we build model using: Random Forest, XGBoot and LGB (LGB is mentioned a lot on Kaggle so we have to try)

There are 2 measurements we use to evaluate: the first is accuracy that is up or down and the second is two sigma

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | train | test | Sigma score |
|  |  |  |  |  |
| Random Forest |  | 0.56 | 0.52 |  |
| XGBoost |  | 0.57 | 0.54 |  |
| LGB |  | 0.6 | 0.58 |  |

# Experiment

After Ensemble Extra Trees (ET) and Ridge we got a good result on training and acceptable result on test, it make us surprise because combine some week model can make a better model in this case the evaluation metric is sigma score. It means that although the prediction of up or down was not good but the confidence value is better than using just one simple model.

# Improvement

Transform news data to make it better for our prediction

Make more market data features

Using cross validation