Applied Data Science Project 5

Jason LI Dejian WANG Yuxin ZHU Ke HAN

Summary

We combined core value of 8 published articles related to Earnings Release and improve it with machine learning strategy. Finally, we got two models:

long-before earning release
short-after earning release

The model is based on one assumption:
Before earning release, stocks with strong momentum will attract people to buy since people believe that the Earnings Report exceed expectations, the price will increase in a short period. And after earning release, some people will cash in and the price will drop in a short period. In summary, long before Earning release, short after it.

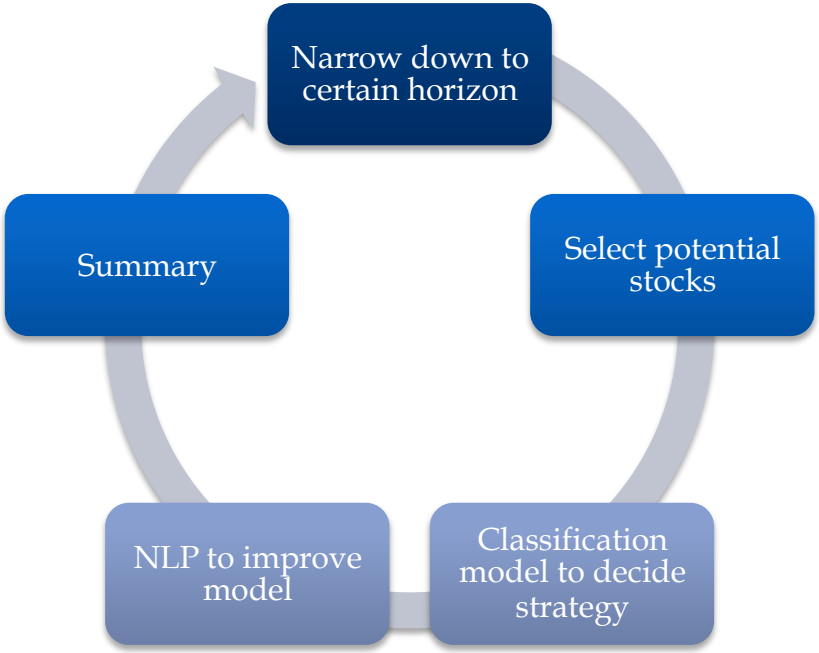
Our model could tell you, if you long the stock before earning announcement, whether you could a positive return. If you short the stock after earning announcement, whether you could get a positive return.

1. Initially, we use 1000 stocks data in 10 years, got 963 K observation and use forward stepwise to select the most important two factors which are momentum. This step is to find a pattern of stocks which are sensitive to Earning announcement.
2. Based on the pattern we found in step 1, we selected fewer observations and use machine learning model to do classification and got 0.88 accurate rate with Random Forest.

Introduction

Scope

- Using the financial data set
- Focused on Earning release date related days
- Using the US market most liquidity stock
- Across more than 10 years investment horizon



Data Description

- Using more than 1000 stocks
- Using more than 39k+ entries
- Including financial fundamental information
- This is the reason there are so many NA, and too dirty the data set

Data Science related

- Predictive modeling
- Classification models
- NLP to generated features
- Data visualization

	A	B	C	D	E	F	G	H	I	J	K
1	RETURN	DY	EBITG	EV2EBITDA	M2B	MOMENTU	PB	PE	PF	PS	days
2	0.35517	0	188.45	7.6547	0.4416	8.3629	0.7015	11.057	4.0495	0.2503	3
3	0.42648	0	188.45	7.596	0.4144	7.8535	0.6584	10.377	3.8005	0.2349	4
4	-0.06206	0	188.45	7.9142	0.5616	7.7587	0.8922	14.063	5.1503	0.3183	1

Step 0: Related Papers and Projects

Adaptive Asset Allocation

- Adaptive asset allocation is a process of constantly rotating into asset with
  - Adaptive momentum
  - Volatility
  - Correlation factors

Momentum

- Defined as the an asset with the greatest relative performance over the past 1 month to 1 year is more likely to exhibit stronger performance over the next few days or weeks
- Some reasons for Momentum:
  - perceptions of risk
  - human cognitive biases
  - the operation of markets

Earning announcement related

- There are two main indicators:
  - SUE (Standardized unexpected earnings $SUE_{i,q} = \frac{X_{i,q} - E(X_{i,q})}{\sigma_{i,q}}$
  - EAR (Earning announcement return) $EAR_{i,q} = \prod_{j=t-1}^{t+1}(1+R_{i,j}) - \prod_{j=t-1}^{t+1}(1+FF_j)$

  Here, FF stands for the benchmark size &book t- market FF portfolio
- And the related trading strategy are:
  - Long the stocks with positive announcement
  - Short the portfolio with negative announcement

Step 1: Data collection and Description

Load data

- Data source: Using Bloomberg, Scraping techniques
- Size: including financial data 39034 entries (based on the volatility, choose the most volatility ones), news related txt files
- Size: 9 files with a total size of 150 MB

Horizon

- Based on the Earning released projects on the previous page, we focused on the earning released date -6 to 6 horizon
- In our data set, the earning release is covering 2006 to 2016 earning release (can refer to earning calendar)

Data processing

- Calculated and processed the dividend yield ratio, enterprise value to EBITDA ratio, market to book ratio, Momentum, price to sales ratio, days(based on the data to earning release) etc.
- For example here,
  - Momentum is based on the price (CLOSE(J) / CLOSE(J – N) * 100)
  - Fama French Book to market ratio

$$\text{Book to Market} = \frac{\text{Book Value of Firm}}{\text{Market Value of Firm}}$$

*Full code can be found in the R markdown file in github
*http://www.nasdaq.com/earnings/earnings-calendar.aspx

Step 2: Select the stocks based on their Momentum

Select Stocks

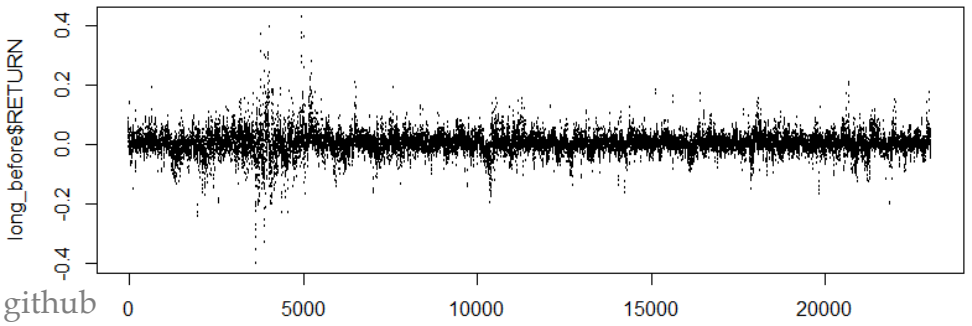
- After comparing some subset of the stocks, we decided to pick up a most sensitive, also known as influential
- We run a stepAIC to figure out the importance of the variables
- In the regression, we see that Momentum and Price to sales ratio is the most important one
- Choose stocks based on those two criteria, limited our sample size, so we focused on the Momentum
- We choose most momentum stocks!

	Before Earning Release	After Earning Release
All stocks	1.58%	1.86%
Momentum	3.09%	3.05%

*Full code can be found in the R markdown file (Main) in github

Major strategies

- Long VS Short
- Before Earning Release VS After
- Our assumptions: Before earning release, stocks with strong momentum will attract people to buy since people believe that the Earnings Report exceed expectations, the price will increase in a short period. And after earning release, some people will cash in and the price will drop in a short period.
- Summarized to 2 models:
  - long-before earning release
  - short-after earning release



Step 2(continue): Exploratory analysis

AIC step result

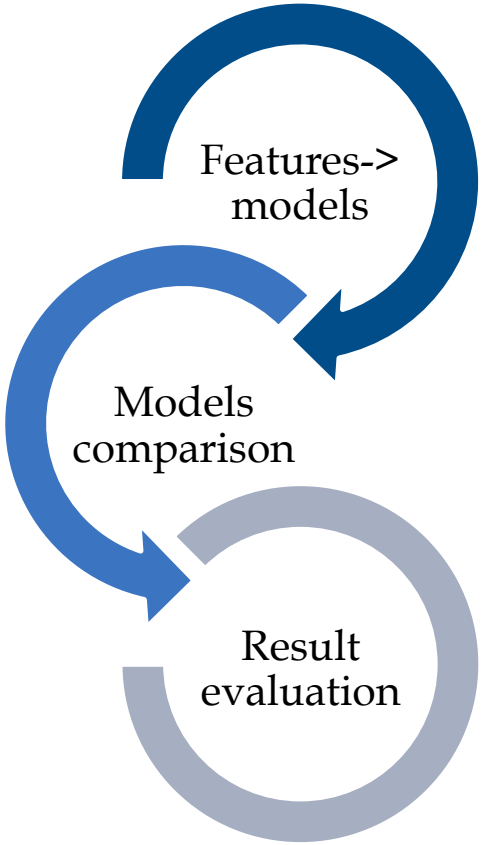
```
summary(stepSA)

##
## Call:
## lm(formula = RETURN ~ DY + EV2EBITDA + M2B + MOMENTUM + PB +
##     PF + PS + days, data = shortAfter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71892 -0.03096  0.00200  0.03229  0.49368
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.316e-03  1.069e-03   3.103  0.00192 **
## DY           2.270e-04  6.400e-05   3.547  0.00039 ***
## EV2EBITDA   -5.917e-05  2.748e-05  -2.153  0.03132 *
## M2B          3.388e-04  2.129e-04   1.591  0.11156
## MOMENTUM    -1.854e-02  1.087e-03 -17.054  < 2e-16 ***
## PB          -4.553e-04  2.595e-04  -1.754  0.07936 .
## PF           4.705e-06  2.580e-06   1.824  0.06822 .
## PS          -1.863e-03  2.178e-04  -8.557  < 2e-16 ***
## days        -5.523e-04  2.398e-04  -2.303  0.02128 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06209 on 22978 degrees of freedom
## Multiple R-squared:  0.02419,    Adjusted R-squared:  0.02385
## F-statistic:  71.2 on 8 and 22978 DF,  p-value: < 2.2e-16
```

Statistical Description of the data

Name	Mean	Median
Dividend yield	3.07	2.069
EBITG	40.35	0.24
EV/EBITDA	11.77	9.19
M/B	5.64	2.80
Momentum	-0.032	-0.104
P/E ratio	22.59	13.96
P/S ratio	2.11	1.64
Return	0.0033	0.0020
Days	3.5	4.0
Surprise	10.73	3.70
P/F ratio	28.19	14.35

Step 3: Classification model

Final model



- After feature selection and running Gradient Boosting Machine, BP Neural network, Random Forest, Logistic regression, Supported Vector Machine,
- Model comparison:
  - SVM: $$J_\lambda(\alpha) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{L}(K^{(i)T}\alpha, y^{(i)}) + \frac{\lambda}{2}\alpha^T K\alpha$$

    tuning on bandwidth and using kernel RBF
  - Random Forest: tuning on mtry
  - Logistic regression: $\beta = (X'WX)^{-1}X'WY$ tuning on bandwidth
  - GBM: Gradient Boosting Machine
  - NNET: feed-forward neural networks with a single hidden layer, and for multinomial log-linear models
- We summarized the prediction accuracy on test data set and runtime analysis

- The best result we can get is 82%!

*Full code can be found in the R file attached

Step 3(continue): Models Summary

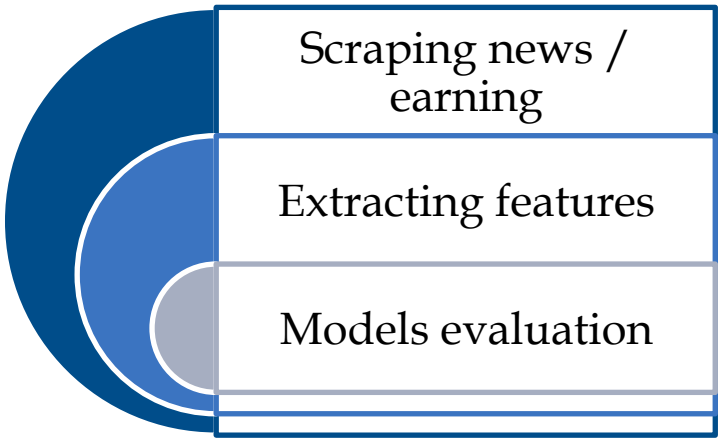
Short After Model Summary

Method	Precision	Recall	F1	Accuracy	Time
GBM	0.71	0.88	0.78	0.70	0.14666295 secs
SVM	0.76	0.79	0.78	0.72	33.51036286 secs
NNET	0.68	0.84	0.75	0.66	84.84540105 secs
RF	0.84	0.88	0.86	0.88	4.17855310 secs
Logistic	0.62	1.00	0.76	0.62	0.01478505 secs

Long Before Model Summary

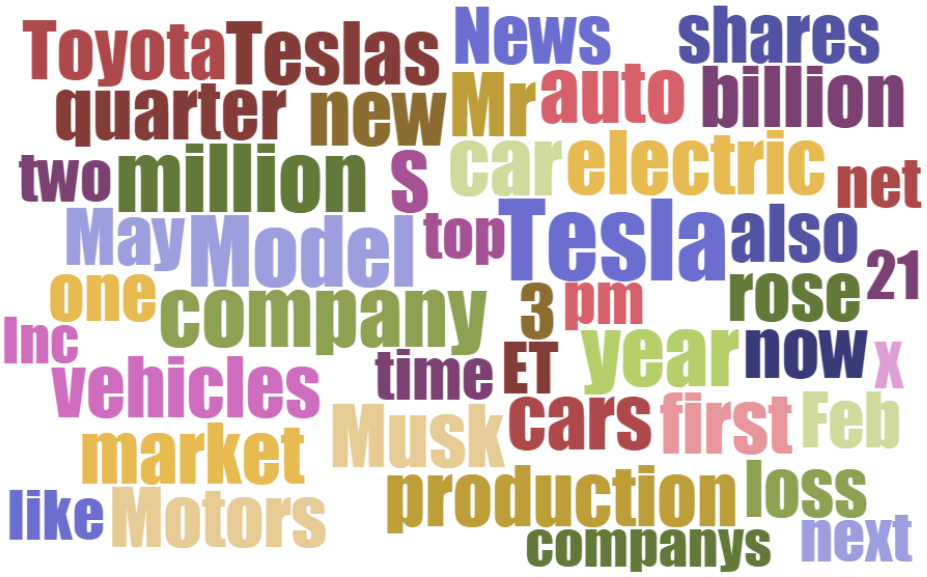
Method	Precision	Recall	F1	Accuracy	Time
GBM	0.55	0.59	0.57	0.61	0.3874459 secs
SVM	0.61	0.72	0.66	0.68	107.2432630 secs
NNET	0.56	0.65	0.60	0.63	72.2320290 secs
RF	0.67	0.70	0.69	0.72	4.8966820 secs
Logistic	0.47	0.50	0.48	0.54	0.0203681 secs

Step 4: Including news features from NLP



- Building up the advanced model from the features, we would have more accrete prediction
- Compared the mixed portfolio and specific stock, we see that specific is better, improved by **5-8%??**
- How to utilized this, the investment horizon changed on this model, we can more flexibility change our portfolio and thus gain better result

- **News and analytic articles**
- We would like to analyze if these article from mainstream media could influence stock market and investor's trading strategies
- **EPS vs Forecast**
- How much the actual EPS bits consensus expectation is also a key factor to be considered for investment strategy
- Those features would expand our info



Step 4(Continue): For mixed portfolio

- Mixed portfolio, commercial words would stand out
- We see that the model would outperform our basics by **5%??**

Step 5: Strategy and Back test

- After building up the portfolio based on our analysis, we have the following lines!

*Full code can be found in the R file attached

Summary and Improvement

Summary

- Stock market is known as a chaotic system. We get more information on news by limiting our scope to earning release day, and are able to build the prediction model of more than 85% accuracy, and better with NLP
- Models and features: GBM, SVM, NNET, RF, Logistic, Random forest gives us the best among all of them
- Further, it is observed that data visualization from different companies shows that data set from a certain company is much more distinguishable than mixing data (portfolio and index)
- Compared with our goal, we just made it. Check whether the strategy still works with recent 10 years data/ Improve the model with statistical technique / Predict whether to long or short before & after Earnings Release

Discussion

- We use their core value: long before Earnings Release, Short after Earnings Release and use machine learning model to improve it

Future work

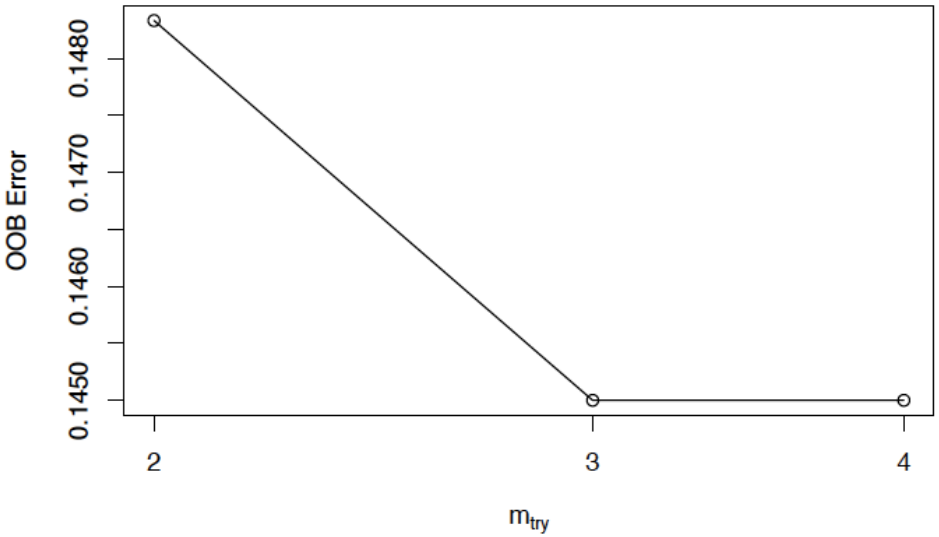
- NLP analysis: The sentiment result from toolkit is too general to apply for financial news and our word vector list is quite limited
- Model Improvement: high dimension can be handled better

Thank You !

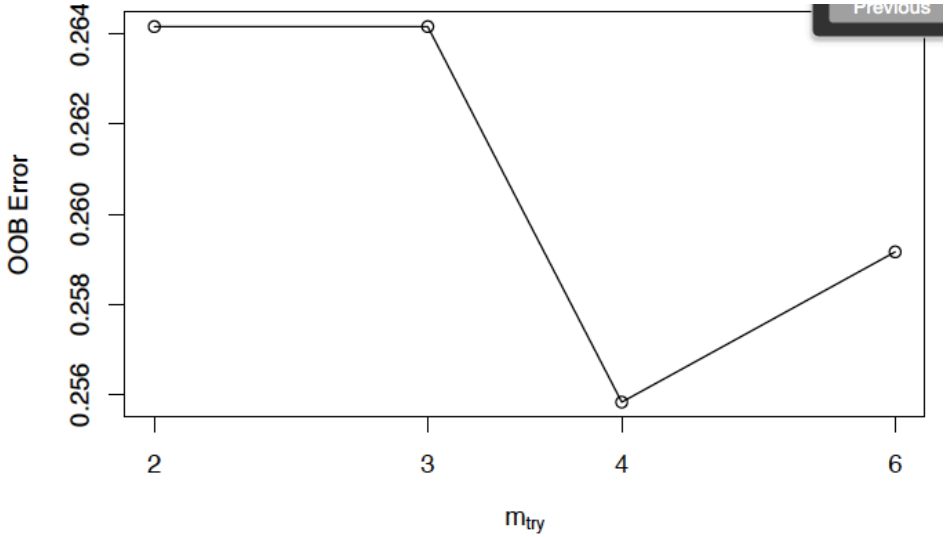
Contact Information: Jason LI
Dejian WANG
Yuxin ZHU
Ke HAN

Attachment: graphics related to training process

Short After (RF tuning)



Long Before (RF tuning)



Word Cloud for the mixed data set