

Applied Data Science Project 5

Jason LI Dejian WANG Yuxin ZHU Ke HAN

Introduction

Scope

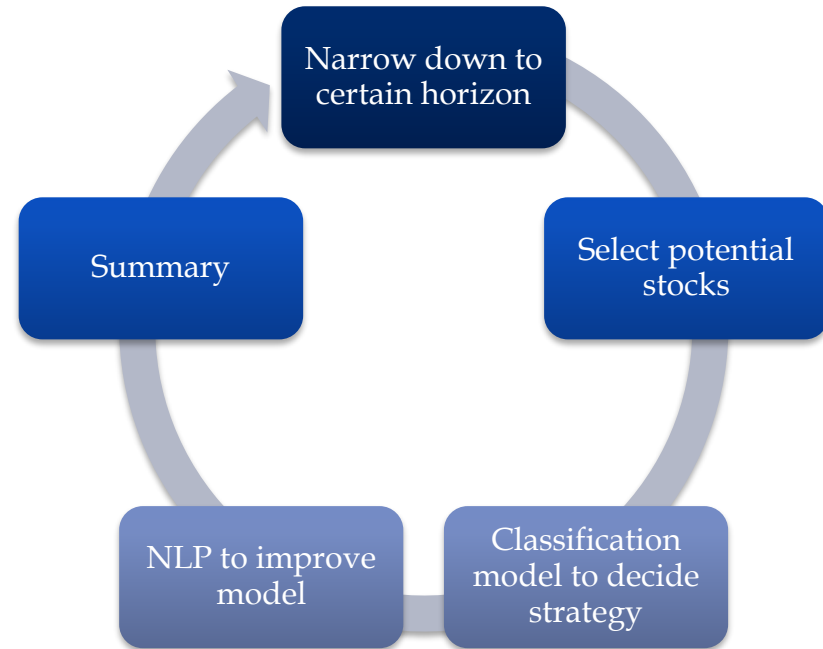
- Using the financial data set
- Focused on Earning release date related days
- Using the US market most liquidity stock
- Across more than 10 years investment horizon

Data Description

- Using more than 1000 stocks
- Using more than 39000+ entries
- Including financial fundamental information

Data Science related

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



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)
- EAR (Earning announcement return)

$$SUE_{i,q} = \frac{X_{i,q} - E(X_{i,q})}{\sigma_{i,q}}$$

$$EAR_{i,q} = \prod_{j=t-1}^{t+1} (1 + R_{i,j}) - \prod_{j=t-1}^{t+1} (1 + FF_j)$$

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 terminal, Scraping from Yahoo and Wall Street Journal
- Size: including financial data 39034 entries (based on the volatility, choose the most volatility ones), news related txt files

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.aspx4>

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 stocks to do further analysis
- Here, we choose most momentum stocks!

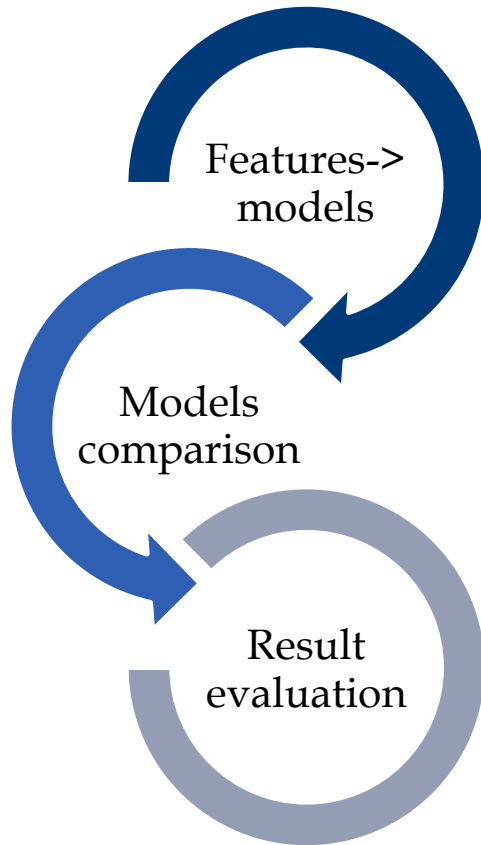
Statistical Description of the data

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

*Full code can be found in the R file attached

Step 3: Classification model

Final model

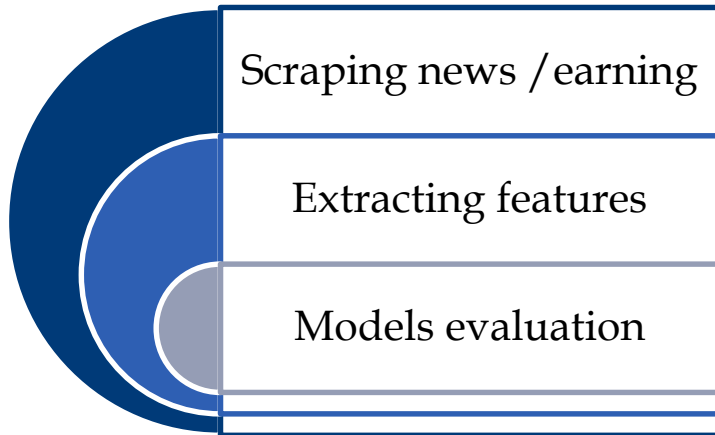


- After feature selection and running Gradient Boosting Machine, BP Neural network, Random Forest, Logistic regression, Supported Vector Machine,
- Model comparison:
 - SVM: $J_{\lambda}(\alpha) = \frac{1}{m} \sum_{i=1}^m L(K^{(i)T} \alpha, y^{(i)}) + \frac{\lambda}{2} \alpha^T K \alpha$
tuning on bandwidth and using kernel RBF
 - Random Forest: tuning on mtry
 - Logistic regression: $\beta = (X'WX)^{-1}X'WY$
tuning on bandwidth
- We summarized the prediction accuracy on test data set and runtime analysis

Models	Prediction accuracy	Run time analysis
SVM	72%	Tens of minutes +
GBM	66%	Tens of minutes
Neural Network	76%	Tens of minutes
Random Forest	82%	Tens of minutes

*Full code can be found in the R file attached

Step 4: Including news features



- **News and analytic articles**
- We would like to analyze if these articles 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

- Building up the linear regression model, we see that land area, population, percentage under poverty, per capita income and personal income are critical to our crime related analysis
- Further more, in the angle of public affairs, county basics and economics related are most important

Step 5: Strategy and Back test

- After building up the

*Full code can be found in the R file attached

Summary and Improvement

Summary

- Machine learning algorithms application, potential monetization

More data

- Use data not only for
- More types,

More Feature

- Using the weather information
- To analyze characteristics, including

Algorithm optimization

- It could be improved with other distributed algorithms
- Some data preparation technique/ subgroup analysis could be done

Thank You !

Contact Information: Jason LI

Dejian WANG

Yuxin ZHU

Ke HAN