Principal® | GEORGE MASON UNIVERSITY | Volgenau School of Engineering

Absolute Return Algorithm - Chinese Equities

Final Presentation

May 2018

George Mason University

Data Analytics Engineering

Acknowledgements

Dr.
James Baldo

Role:
Professor

School:
George Mason University

Dr.
F Berlin

Role:
Professor

School:
George Mason University

Mr.
Chase Grimm

Role:
Mentor

Data & Operations
Research Scientist

Company:
Principal Financial Group

Mr.
Ben Harlander

Role:
Mentor

Data & Operations
Research Scientist

Company:
Principal Financial Group

Dr.
Joe Byrum

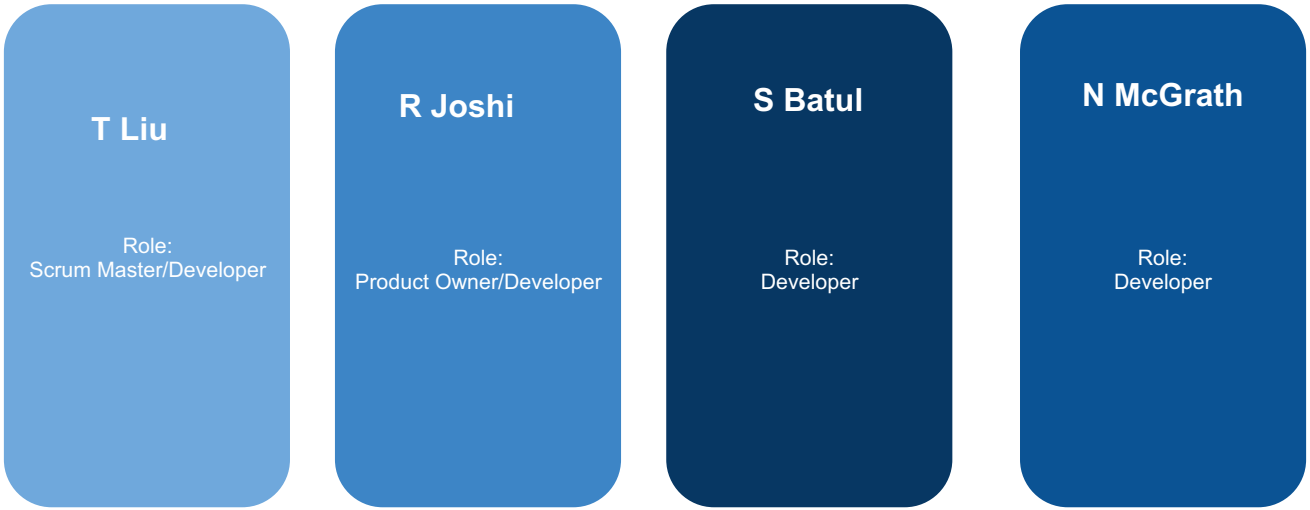
Role:
Project Sponsor

Chief Data Scientist

Company:
Principal Financial Group

Principal℠

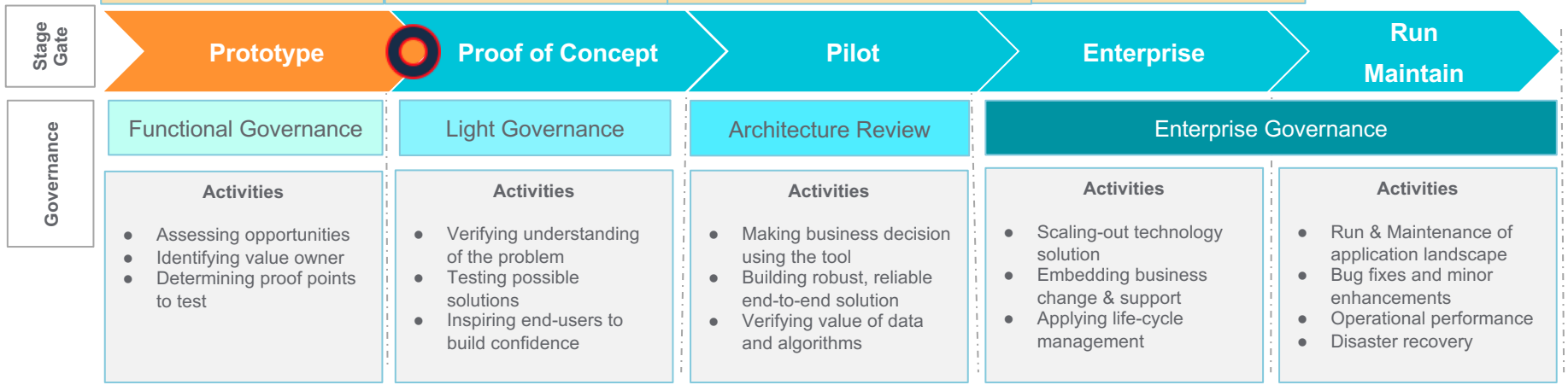
The Team



Principal℠

PGI Pipeline Flow

New Idea Deliverables	POC Deliverables	Pilot Deliverables	Enterprise Deliverables
• Insights and surprises into today's pain points • Theoretical viable solution • Organizational support to tackle problem	• Preliminary business value defined • Minimum viable solution • Organizational, Process and end user assessment	• Validated business case • Minimum scalable solution • Organizational change commitment • Validated with real data in a real application environment	• Fully scalable solution • Business Implementation • KPI's in place • Organizational change • Process changes implemented



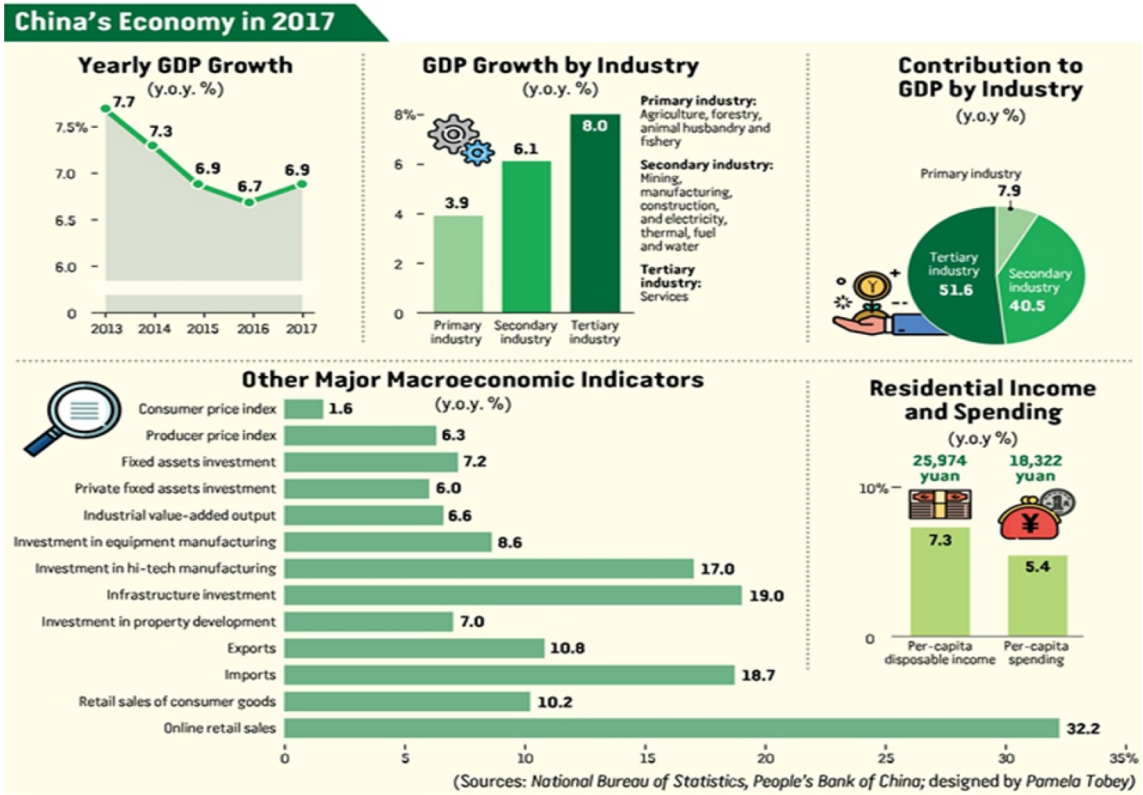
Principal℠

Agenda

- Problem Scope
- Research & Dataset
- Analytics & Results
- Summary & Future Work

Principal℠

China: A Booming Yet Complex Economy



Investment Opportunities

- Equity returns are considerable;
- Reasonably priced when compared with peers;
- Industrial revolution and internet revolution provide exciting investment opportunities;
- Consumption upgrade.

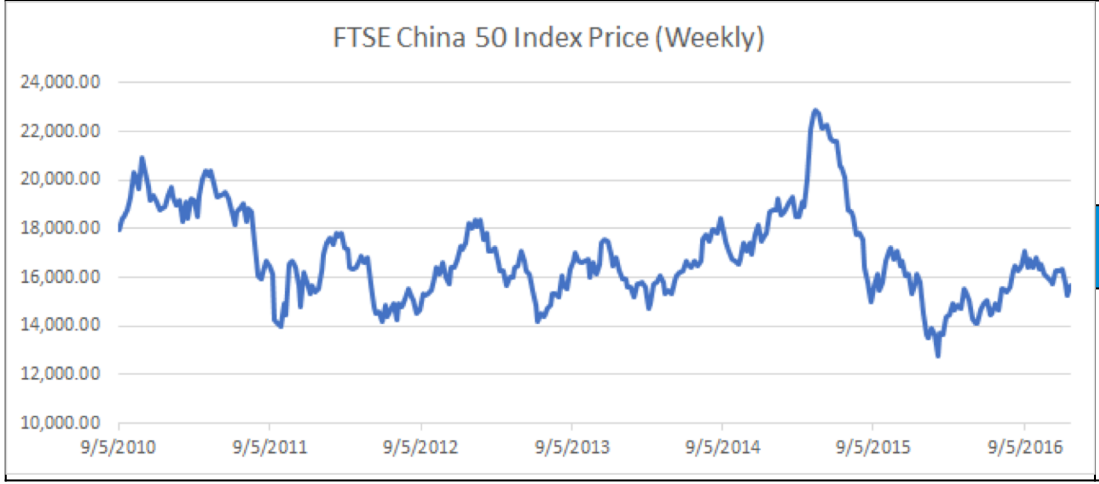
Investment Risks

- High degrees of investment risk;
- Low levels of transparency;
- Unfamiliarity with local opportunities;
- Deleveraging and state-reform programs;
- Preponderance of 'retail punters' in market.

Principal℠

Why is there a Need for a New Absolute Return Algorithm?

Market Volatility



Current Model Methodology

Factors
Fundflow
Earning Revision
Economic Surprise
Buffet Model
Risk Factor

ETF & index
MSCI China Index
YANG UltraPro Long 3x
YINN UltraPro Short -3x

Current Model Performance

Annualized Return 2010 - 2016
Benchmark: -1.20%
Portfolio 1: 7.39%
Portfolio 2: 5.86%
MSCI: -9.64%
FTSE: -17.09%

Principal℠

Project Objectives

- Research the Chinese market and economy to select possible macro and micro economic features that could have predictive power
- Create a multi-factor regression/classification model based on weekly movement
- Identify a group of factors which have predictive power on market movement and also make economic sense

Agenda

- Problem Scope
- Research & Dataset
- Analytics & Results
- Summary & Future Work

Principal℠

Macroeconomic Indicators

Chinese Market

Daily
- Volatility Index (VIX)
- Exchange Rate
- Government Bond Yield to maturity (YTM)

Monthly
- Policy Uncertainty Index
- Purchasing Managers Index (PMI)
- Consumer Price Index (CPI)
- Cash Assets

Quarterly
- GDP
- Unemployment Rate

Why we pick those indicators?

United States Market

- US GDP
- US Short and Long Term interest rate
- Inflation Rate
- Federal Debt
- Balance of Trade

Why choose US not other countries?

Why picking those Indicators?

Feature Selection

- **Timeliness**
  - Time range: 2013 to 2018
  - Seasonality pattern
- **Empirical Value**
  - Economy condition
  - Market condition
- **Research Value**
  - Predictive power

Relation between US and Chinese

- **US and China are each other's largest trading partner**
  - China Trade War
- **Exchange Rate of USD and CNY**
  - Volatility
- **America and China Need Each Other**
  - Broad and consequential relationship

Microeconomic Indicators

Stock Indices	ETF related Indicators
Shanghai Stock Exchanges • SSE index • SSE A-Share index • SSE B-Share index **Shenzhen Stock Exchanges** • SZSE index • SZSE A-Share index • SZSE B-Share index **Hong Kong Stock Exchanges** • Hang Seng index • Hang Seng mainland index • Hang Seng Industry index	**Moving Average Lines** • Obvious uptrend or downtrend **Trading Volume** • Little trading volume shows wider bid-ask spreads **Candlesticks** • Obvious bullish or bearish **Sentiment** • Social media sentiment, manual trading **Tracking Index** • FTSE China 50

Data Cleansing: Missing Values

Generally, all stock markets in the world close on holidays and weekends. In China and Hong Kong, there are about 252 trading days in a year, however, the holidays are different from the two regions.

Date	SSE	HSI
12/29/2017	3307.172	29919.15
12/28/2017	3296.385	29863.71
12/27/2017	3275.783	29597.66
12/26/2017	3306.125	#N/A
12/25/2017	3280.839	#N/A
12/24/2017	#N/A	#N/A
12/23/2017	#N/A	#N/A
12/22/2017	3297.063	29578.01
12/21/2017	3300.059	29367.06
12/20/2017	3287.606	29234.09
12/19/2017	3296.538	29253.66
12/18/2017	3267.922	29050.41
12/17/2017	#N/A	#N/A
12/16/2017	#N/A	#N/A
12/15/2017	3266.137	28848.11
12/14/2017	3292.439	29166.38
12/13/2017	3303.037	29222.1
12/12/2017	3280.814	28793.88
12/11/2017	3322.196	28965.29

Hong Kong has Christmas Holiday, but mainland China doesn't.

$$10-day\ SMA = \frac{p_D + p_{D-1} + \cdots + p_{D-9}}{10}$$

$$= \frac{1}{10}\sum_{i=0}^{9} p_{D-i}$$

Date	10-day MA SSE	10-day MA HSI
12/29/2017	3294.17	29544.76
12/28/2017	3289.81	29420.66
12/27/2017	3288.99	29346.82
12/26/2017	3290.88	29296.65
12/25/2017	3285.17	29221.89
12/24/2017	3286.82	29213.96
12/23/2017	3288.85	29214.98
12/22/2017	3287.96	29168.19
12/21/2017	3290.75	29100.11
12/20/2017	3289.59	29066.74
12/19/2017	3289.87	29042.83
12/18/2017	3288.93	28955.15
12/17/2017	3289.52	28848.40
12/16/2017	3290.08	28770.45
12/15/2017	3291.59	28778.49
12/14/2017	3296.42	28810.73
12/13/2017	3296.92	28766.27
12/12/2017	3296.04	28701.16
12/11/2017	3301.30	28741.21

Data Transformation: Indicators

Lag Days

Lag Days = Data Public Release Date - Data Observation Date

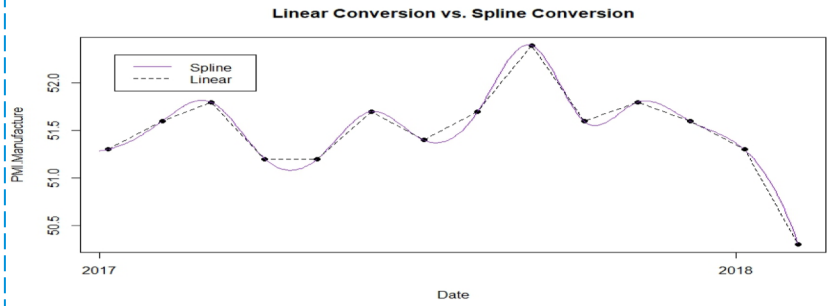
Subject	Release Date	Observation Date
GDP (Q4 2017)	January 26, 2018	December 31, 2017

Why those transformations are important?

Weekly Timeframe

Spline Interpolation

- Smoother than linear equation
- No duplicate data on adjacent weeks
- Reflects real world fluctuation



Principal℠

Why those transformations are important?

Lag Days

- 75% of our indicators have lags
- Nearly every issue in economic can be subject to time-lag
- Economic indicators affect the market after the data released

Weekly Timeframe

- Balance trading fees and weekly returns
- Keep a certain turnover rate
- Vital for both short-term and long-term analysis

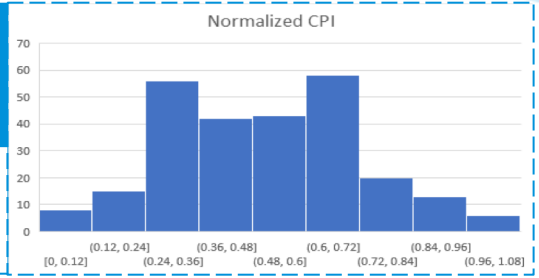
Data Transformation: Final Datasets

Original Dataset

· Different distributions
· Different scales

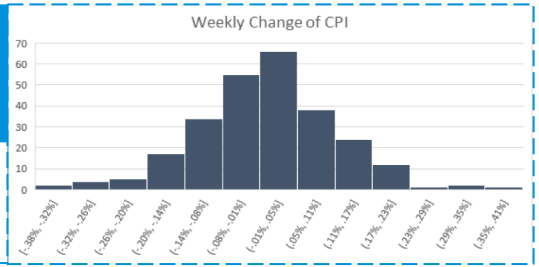
Dataset #1
Normalized Dataset

· Normal distribution to fit most algorithms
· Same scale from 0 to 1



Dataset #2
Weekly Change Dataset

· Normal distribution to fit most algorithms
· Same scale in percentage



Principal℠

Agenda

- Problem Scope
- Research & Dataset
- Analytics & Results
- Summary & Future Work

Principal℠

Unsupervised Learning

What is Unsupervised Learning?

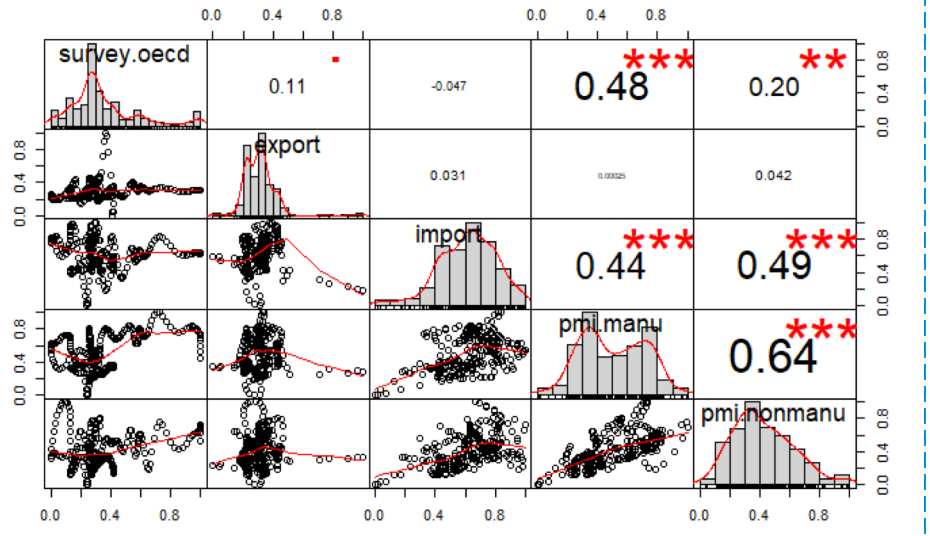
- **Machine learning algorithm;**
- **Unsupervised learning happens when you only have predictor variables;**
- **No need to consider response variables as results;**
- **Goal**
  - Model the structure or distribution of the data
  - Reduce dimensions of the data
  - Learn more about the data

Unsupervised Learning methods

- **Correlation Check**
  - Data reduction option #1
  - Evaluate the relation between indicators
  - Aim to remove highly correlated indicators
- **Principal Component Analysis (PCA)**
  - Data reduction option #2
  - Transform correlated indicators into uncorrelated variables called principal components.
- **Clustering**
  - Identify the data points into different groups by similarities

Unsupervised Learning: Correlation Check

Correlation Check



This graph provides the following information,

- Correlation coefficients: the strength of the relationship.
- P-value: the significance of the relationship.
- Histogram with kernel density estimation and rug plot.
- Scatter plot with fitted line.

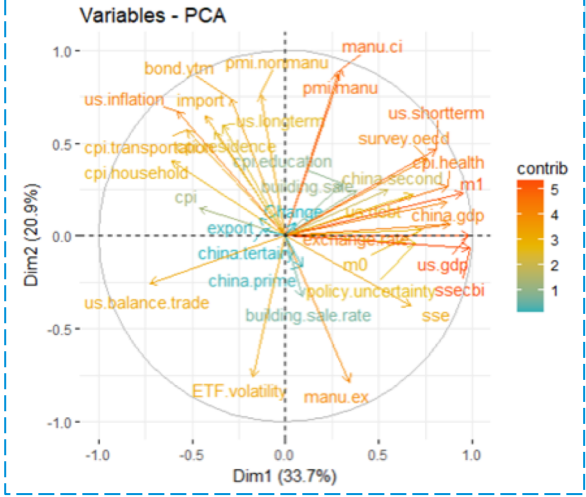
Principal℠

Unsupervised Learning: PCA

Principal Component Analysis (PCA)

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     3.3879 2.6205 1.8776 1.45717 1.40975 1.10160 1.01113
Proportion of Variance 0.3376 0.2020 0.1037 0.06245 0.05845 0.03569 0.03007
Cumulative Proportion  0.3376 0.5396 0.6432 0.70569 0.76415 0.79984 0.82991
                          PC8     PC9     PC10    PC11    PC12    PC13    PC14
Standard deviation     0.9793 0.8410 0.79213 0.74501 0.68361 0.67140 0.60363
Proportion of Variance 0.0282 0.0208 0.01845 0.01632 0.01374 0.01326 0.01072
Cumulative Proportion  0.8581 0.8789 0.89737 0.91369 0.92744 0.94070 0.95141
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.55642 0.49030 0.46054 0.44525 0.38886 0.37269 0.29058
Proportion of Variance 0.00911 0.00707 0.00624 0.00583 0.00445 0.00409 0.00248
Cumulative Proportion  0.96052 0.96759 0.97383 0.97966 0.98411 0.98819 0.99068
                          PC22    PC23    PC24    PC25    PC26    PC27    PC28
Standard deviation     0.25929 0.24794 0.20861 0.19105 0.17952 0.16277 0.12357
Proportion of Variance 0.00198 0.00181 0.00128 0.00107 0.00095 0.00078 0.00045
Cumulative Proportion  0.99265 0.99446 0.99574 0.99682 0.99776 0.99854 0.99899
```

PC1 to PC8 tell 86% of the original information;
PC1 to PC11 tell 91% of the original information;
PC1 to PC21 tell 99% of the original information.



Reduce Overfitting

Overfitting: the model learns noise in the training dataset and performs very well on it, but gives high errors and huge variance on testing dataset.

Principal℠

Dataset Separation

Training Dataset

- From the week of 2/8/2013 to the week of 1/19/2017
- 207 weeks

Testing Dataset

- From the week of 2/2/2017 to the week of 2/2/2018
- 52 weeks

one week buffer between training and testing dataset

Principal℠

Training Model: Regression Algorithms

Algorithm	Description	Advantage
Linear Regression	Use linear approach to predict a quantitative response variable with a set of predictor variables. The “best fit” line through all data points.	Easy to understand. You clearly see what the biggest drivers of the model are.
Regression Tree	Use decision tree to represent the recursive partition with a set of predictor variables.	It’s easy to understand what variables are important in making the prediction.
Random Forest	Takes advantage of many trees, with rules created from subsamples of features. Each tree is weaker than a full decision tree, but by combining them we get better overall performance.	A sort of “wisdom of the crowd”. Tends to result in very high quality models. Fast to train.
Gradient Boosting	Use the model in a stage-wise form of an ensemble of weak prediction models, and it generalizes them by optimization	Can fit heterogeneous data; High-performing.

Validation Methods

Model Selection

Validating using R^2
- Relative measurement
- Range from 0 to 1
- Model with larger R^2 is better

Validating using RMSE/MAE
- Absolute measurement
- Indicates how close the observed data points are to the model's predicted values
- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)
- Model with smaller RMSE/MAE is better

Validating using AIC/BIC
- Relative measurement
- Model with smaller AIC/BIC is better

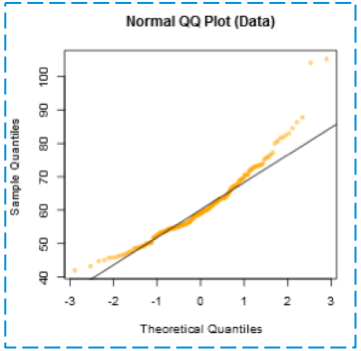
Robustness Test

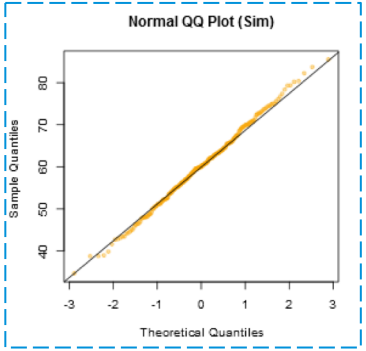
Add small numbers (noise) to two datasets and see if top variables remain unchanged

Normal QQ Plot of Residuals

residual = observed value - predicted value

Assumptions:
Sum of residuals is zero
Best fitted model residuals are normal distributed



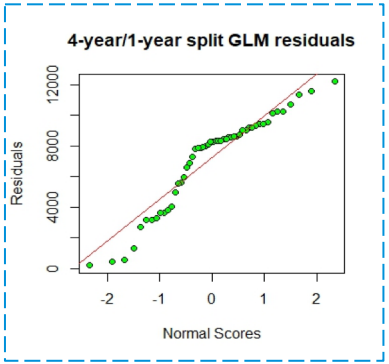
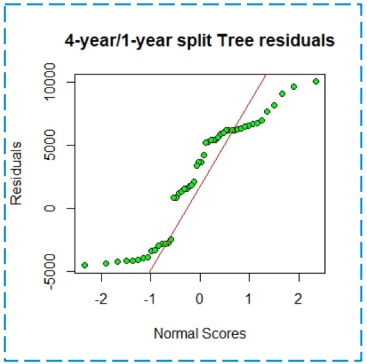
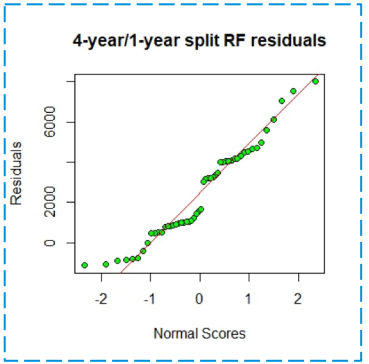
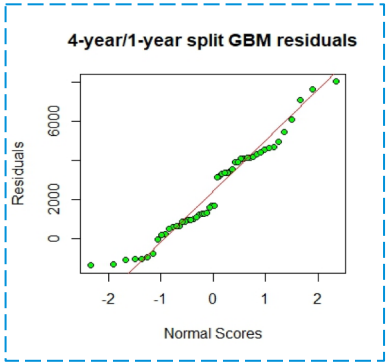


Straight line: normal distribution line
The more residual points locate in line, the better results.

Principal℠

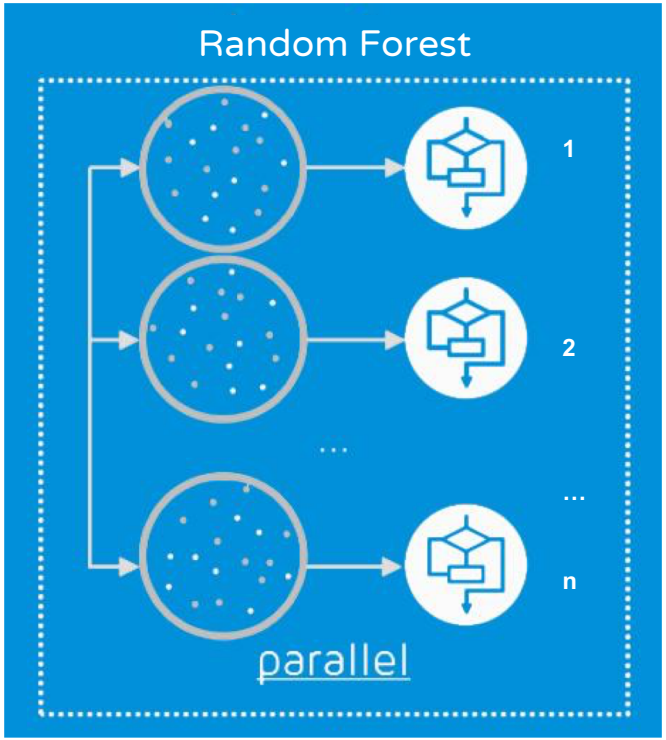
Validation: Residual Plot and RMSE

Normalized Dataset

Linear Regression	Regression Tree	Random Forest	Gradient Boosting
			
RMSE=7755.86	RMSE=5147.54	RMSE=3344.82	RMSE=3603.56

Principal℠

What is Random Forest?



Algorithms

- The numbers from 1 to n mean decision tree 1 to n
- Each tree has a loss function, which is parallel
- Random Forest Algorithm choose several trees with lower loss results
- Make predictions based on those selected trees
- Final prediction is voted from those predictions

In programming, we choose number of trees (n), number of variables tried in each tree (mtry) based on best regression tree result, and get best result by minimal loss function.

Robustness Test: Results after adding noise

Random Forest Robustness Test

Normalized data

var	%incRMSE
CPI	18.51
Survey.OECD	14.83
SSE	14.60
PMI.Manufacture	14.46
CPI.Household	12.82

Normalized data + .0005

var	%incRMSE
CPI	17.52
SSE	15.91
PMI.Manufacture	14.72
Survey.OECD	12.40
Week Number	11.93

Normalized data - .0005

var	%incRMSE
CPI	17.90
SSE	14.35
Survey.OECD	13.09
Manufacture.CI	11.76
Exports	11.63

Normalized data + .0001

var	%incRMSE
CPI	17.57
SSE	14.13
Survey.OECD	13.02
PMI.Manufacture	12.39
Week Number	11.81

Normalized data - .0001

var	%incRMSE
CPI	17.61
SSE	14.50
Survey.OECD	13.00
Manufacture.CI	12.11
CPI.Household	12.01

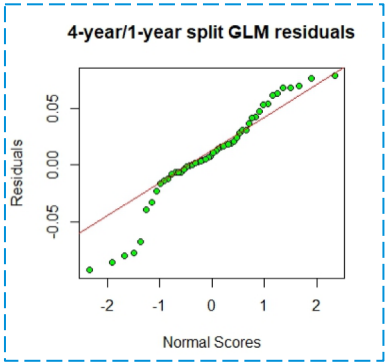
Normalized data random noise

var	%incRMSE
CPI	16.97
SSE	14.66
Survey.OECD	12.79
PMI.Manufacture	11.72
CPI.Household	11.12

Validation: Residual Plot and RMSE

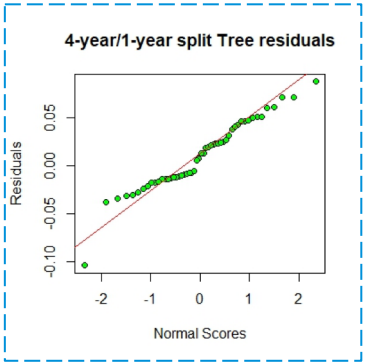
Percentage Weekly Change Dataset

Linear Regression



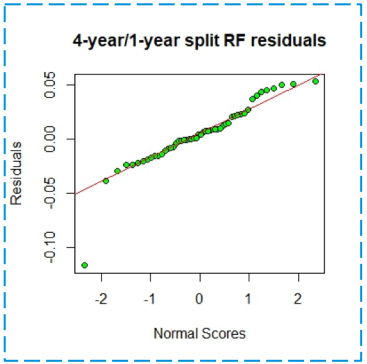
RMSE=4.74%

Regression Tree



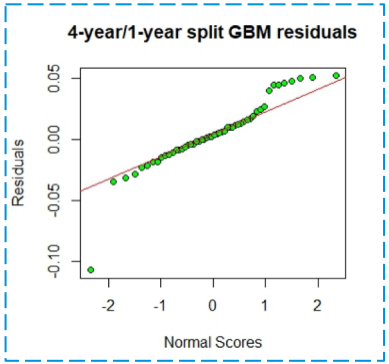
RMSE=3.66%

Random Forest



RMSE=2.79%

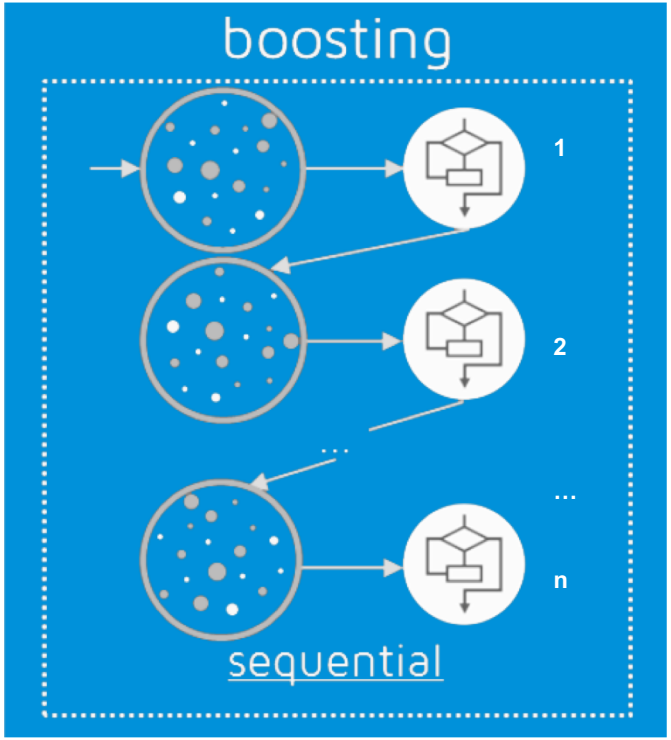
Gradient Boosting



RMSE=2.74%

Principal℠

What is Gradient Boosting?



Algorithms

- The numbers from 1 to n mean decision tree 1 to n
- Each tree has a loss function, and a prediction calculated by learning rate
- Losses from tree 1 to tree n are gradient reducing, which is sequential
- Boosting Algorithm chooses the n^{th} tree with lowest loss result, and makes final prediction based on the n^{th} tree

In programming, we choose the learning rate (alpha), the number of trees (n) by grid search, and get best result by minimal loss function.

Robustness Test: Results after adding noise

Gradient Boosting Robustness Test

Weekly change data

var	rel.inf
SSE	10.145769
CPI.Residence	9.058734
M1	9.003031
US_LT_INT	7.387070
Manufact.CI	6.274988

Weekly change data -.5%

var	rel.inf
SSE	10.505489
US_LT_INT	8.240447
CPI.Residence	7.915105
M1	7.037673
M0	6.511761

Weekly change data +.5%

var	rel.inf
SSE	9.266028
CPI.Residence	9.028783
M1	7.858548
US_LT_INT	7.788937
Manufact.CI	6.254241

Weekly change data Random noise

var	rel.inf
M1	9.355416
US_INF_RATE	9.058754
US_LT_INT	8.451384
SSECBI	7.492808
SSE	6.747347

Agenda

- Problem Scope
- Research & Dataset
- Analytics & Results
- Summary & Future Work

Principal℠

Conclusion

- **Based on our extensive research we can conclude that even though Chinese market could be volatile it has very good potential for investors.**
- **Based on our research and analytical models we can say that the following indicators would lead to significant results in future models:**

1. **For normalized dataset**
   a. Major index of Shanghai Stock Exchange (SSE)
   b. Consumer Opinion Survey from OECD (Survey.OECD)
   c. Consumer Price Index (CPI)

2. **For weekly change dataset**
   a. Weekly change of major index of Shanghai Stock Exchange (SSE)
   b. Weekly change of US Long Term Interest Rate (US.Long)
   c. Weekly change of M1 Cash asset (M1)

- **Using big data to research market can help explain the chinese market**

Future Work: Indicators

Separate the Chinese indicators from US indicators

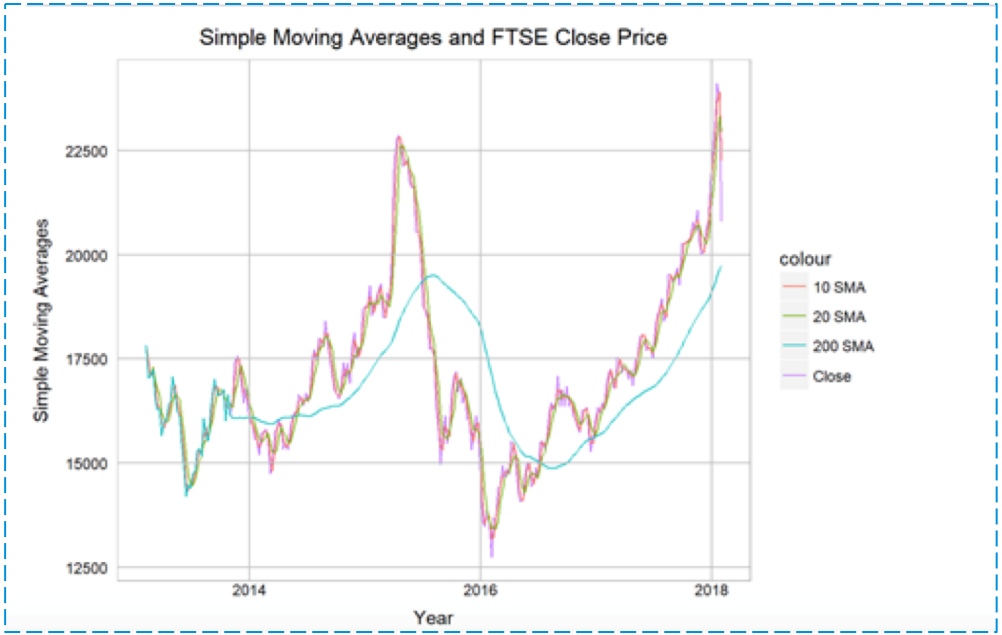
- E.g. use four Chinese indicators and two US indicators in model, other than put it all together

Research and add indicators from multiple countries

Research and add US Manufacturing PMI indicators

- PMI tells power of manufacturing sector
- PMI tells information about the imports and exports between China and the US

Future Work: Simple Moving Average



If weekly price went higher than all 10-day, 20-day and 200-day SMA, mark signal 1; otherwise, mark signal 0.

```
##                     signal 0 signal 1
##   Weekly Change 0 63.82979         0
##   Weekly Change 1 36.17021       100
```

If 10-day SMA went higher than 20-day SMA, mark signal 1; otherwise, mark signal 0.

```
##                     signal 0 signal 1
##   Weekly Change 0 62.18487 32.39437
##   Weekly Change 1 37.81513 67.60563
```

Suggestion: to build complicated rules based on statistical learning, or to try clustering method

Principal℠

Future Work: Candlesticks

K-means Clustering

- L-O: difference between weekly lowest price and weekly open price
- H-O: difference between weekly highest price and weekly open price
- C-O: difference between weekly close price and weekly open price (real body)

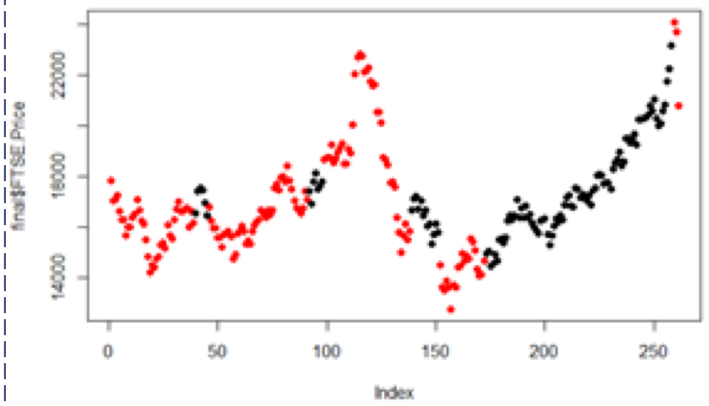
```
##                   cluster 0 cluster 1
##   weekly change 0  88.70968   7.29927
##   weekly change 1  11.29032  92.70073
```


Principal℠

Future Work: Clustering Regime

K-means Clustering

The goal of K-means Clustering algorithm is to find groups in the data based on the similarities that are provided.



After k-means Clustering
(Label with cluster 1 and 2)

Suggestions:

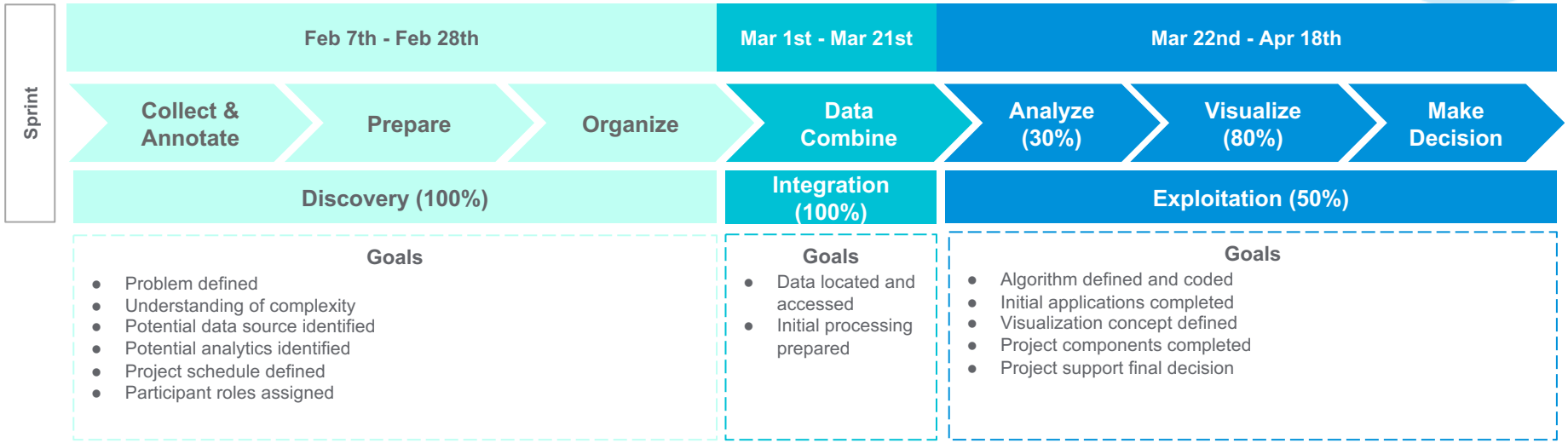
- Run algorithms within different cluster to make results more accurate;
- A new interesting topic: market regime detection.

Questions?

Back-up Slides

(under construction)

Chinese Equities Project: Pipeline Flow



Principal℠

Criteria for Feature Selection

Timeliness

- Data that are provided in a timely, consistent, and highly regular manner can be used as early indicators of market movement.
- This is relevant given the high fluctuations in the Chinese equities market.

Empirical Value

- Several of the indicators can be used on its own or as a source for other measures beyond its reflection of a certain market condition.

Research Value

- Data have been researched and identified as having potential to help predict changes the economy

ETF tracking index: FTSE

What is the FTSE?

- The FTSE China 50 Index measures the performance of Large Cap securities and is selected by a Hong Kong-listed process.

Why choose the FTSE?

- Currently, 6 ETFs track the FTSE China 50 Index with more than $5B in Exchange Traded Products assets with an average expense ratio of 1.00%.

Which ETF tracks FTSE?

TICKER	FUND NAME	LEVERAGE	MARKET
FXI	iShares China Large-Cap ETF	1X	Bull
YINN	Direxion Daily FTSE China Bull 3X Shares	3X	Bull
XPP	ProShares Ultra FTSE China 50	2X	Bull
YANG	Direxion Daily FTSE China Bear 3X Shares	3X	Bear
FXP	ProShares UltraShort FTSE China 50	2X	Bear
YXI	ProShares Short FTSE China 50	1X	Bear

Data Preprocessing Methods

Data Cleansing	Data cleansing routines work to clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.
Data Integration	Data integration combines data from multiple sources into a coherent dataset. These sources include multiple databases, data cubes, or flat files.
Data Transformation	Data transformation aims to transform or consolidate data into forms appropriate for mining, which can involve *Normalization*, *Smoothing*, *Aggregation* and *Generalization*.
Data Reduction	Data reduction techniques are helpful in analyzing reduced representation of the dataset without compromising the integrity of the original data and producing the quality knowledge.

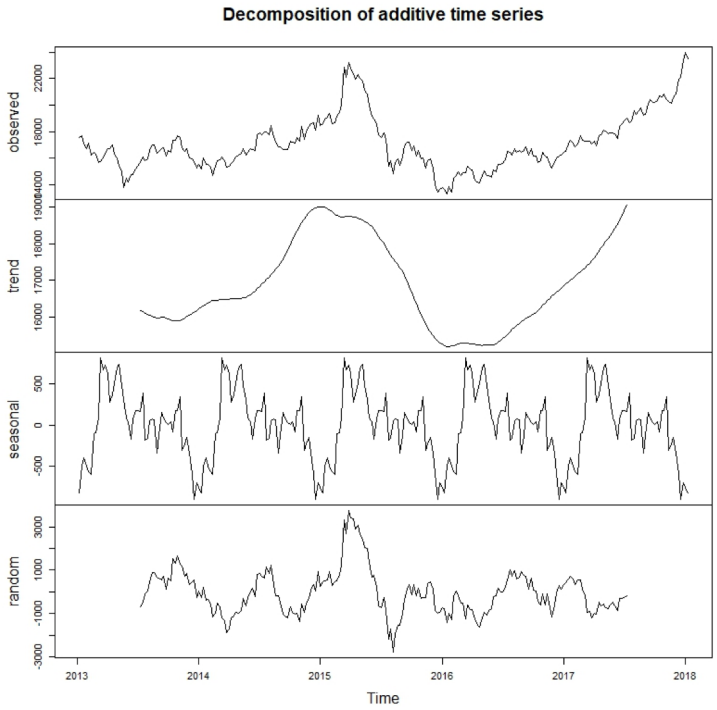
Basic Time Series Analysis for FTSE

Seasonality

- A seasonal pattern exists when a series is influenced by seasonal factors
- Our response variable is influenced by the week

Decomposition

- A time series consists of a trend, a seasonality and a randomness
- Decomposition means separating the time series into these three components



original time series of tracking index price

overall trend after removal of seasonality and randomness

weekly pattern

randomness

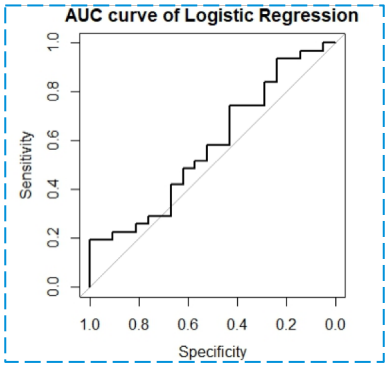
Principal℠

Training Model: Classification Algorithms

Classification Algorithms	Description	Advantage
Logistic Regression	Logistic regression is the classification counterpart to linear regression. Use a logistic curve to limit prediction within two values outcome	The predictor variables don't have to be normally distributed
Linear Discriminant Analysis	Based upon the concept of searching for a linear combination of predictor variables that best separates two classes	Can be used to determine which variable discriminates between two or more classes
Classification Tree	It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.	Robust to outliers, scalable, and able to naturally model non-linear decision boundaries
Naive Bayesian	Based on Bayes' theorem with the independence assumptions between predictors. Bayes theorem provides a way of calculating the posterior probability.	Easy to build; Useful for very large datasets

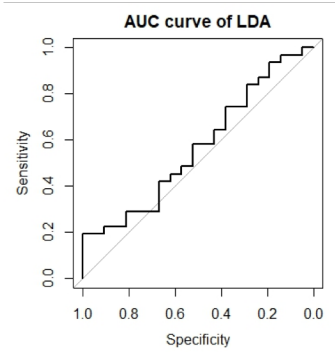
Classification Algorithms: Results

Logistic Regression



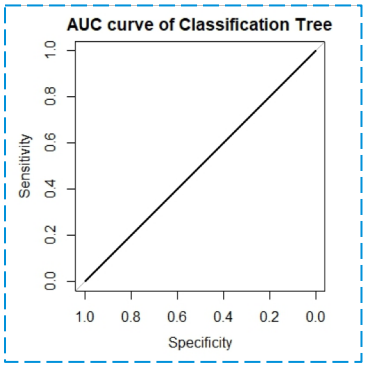
AUC=0.5776

LDA



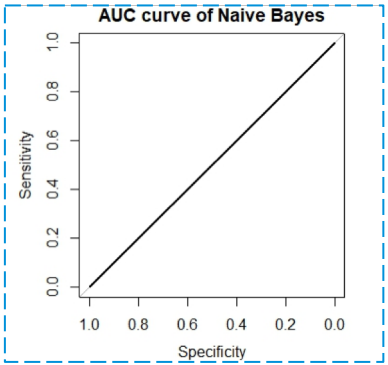
AUC=0.56

Classification Tree



AUC=0.5

Naive Bayes



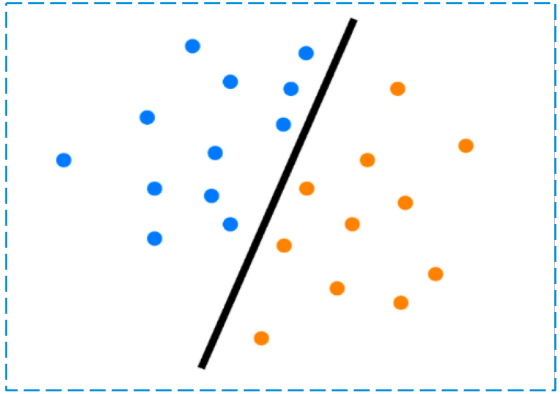
AUC=0.5

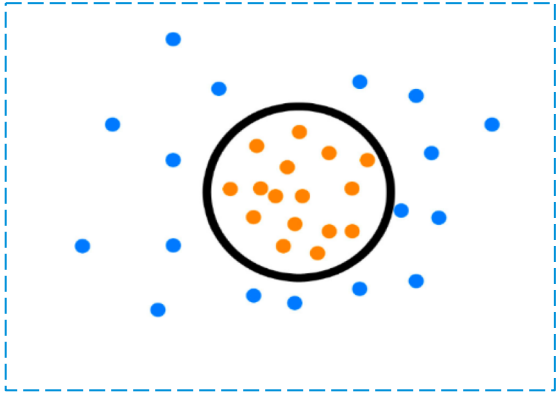
Top variables:
CPI related features and GDP related features

Principal℠

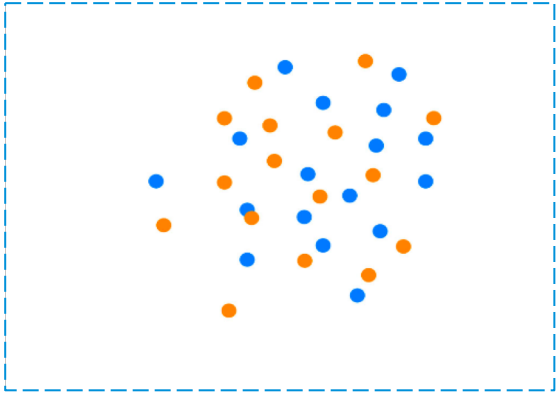
Why Classification Model Fails?

Dataset expected to be:





Our dataset was like:



- No obvious similarities among attributes to do classification,
- Training dataset significantly different from testing dataset,
- High dimensions for implementation

Suggestions

Try Sentiment Classification on how policy, news report, disasters and consumer behaviors influence the price change.