Absolute Return Algorithm: Chinese Equities

Volgenau School of Engineering
at
George Mason University

GEORGE MASON UNIVERSITY | Volgenau School of Engineering

Sponsored Project by
Principal Global Investor

Principal®

By -
T Liu
N McGrath
R Joshi
S Batul

Table of Contents

Abstract

Creating an Absolute Return Algorithm that Competes with the Highly Volatile Chinese Exchange Market.

Principal Global Investors is a global asset firm offering a boutique investment atmosphere to its clients by conducting independent research and, thus, providing a customized investment strategy to clients. Chinese clients have shown increased demand in an absolute return fund that uses Chinese based equities. An absolute return fund is a type of fund that seeks to shield clients from high volatility while producing positive returns over a certain period. This has proven a difficult task for many investment firms and the highly volatile Chinese exchange market adds more complexity to the problem.

This project will research the Chinese economy to find key macroeconomic indicators that could influence the equities market. The chosen macroeconomic indicators (i.e. Gross Domestic Product (GDP)) will be used, along with microeconomic indicators (i.e. Simple Moving Average (SMA)), to develop a multi-factor model that predicts short-term market movement and returns a yearly positive return on investment greater than the yearly return of an identified Exchange Traded Fund (ETF) tracking the FTSE China 50 Index, an index that tracks the largest and most liquid companies in China. The project will analyze five years of economic and financial data from China and the U.S. as the U.S. has close economic ties to China and could provide predictive insight into market movement. The model will account for seasonality of macroeconomic data and the timing of the release of data.

To create the model, the team will use both supervised and unsupervised learning methods to choose macroeconomic indicators. As well, unsupervised learning will be used to evaluate microeconomic indicators and find trading signals. Once signals have been identified, a set of investment strategies for buying and selling ETFs will be developed and evaluated for effectiveness. The results will stimulate further research into the absolute return algorithm problem space and could increase the market presence of PGI in China.

PA

Project Definition

Overview

The growth of the Chinese economy over the past twenty years has been remarkable by historical standards. In real GDP terms, it has grown from a value of 1877.43 billion Chinese Yuan in 1990, to a value of 827121.7 billion Chinese Yuan in 2017 (Huang, Lai, & Besler, February 2018). At the same time, there is an increasing demand from Chinese clients for Principal Global Investors, i.e. insurance companies, banks wealth management products, Chinese state-owned enterprises, for an absolute return solution on offshore China Equity (Principal Global Investors, July 2017). In this section, we aim to show the background of the project, the importance of having absolute return algorithm, the current situation of Chinese Economy and Chinese Market, the basic needs of Principal Global Investors, and the objectives we aim to meet for this project.

Background

Absolute Return Algorithm

An absolute return fund (Absolute Return, 2018) is a type of fund that seeks to shield investors from high volatility while producing positive returns over a certain time period. It is an investment vehicle that uses techniques such as “short selling, futures, options, derivatives, arbitrage, leverage and unconventional assets” to build a portfolio with positive overall returns. This has proven a difficult task for many investment firms.

Principal Global Investors (PGI) is a global asset firm working as part of the larger Principal Financial Group. Currently, PGI has over $360 billion assets under management and operates in over 70 countries (Principal Global Investors, August 2017). However, PGI offers a boutique atmosphere to its clients because PGI believes in empowering its smaller subsidiaries with the ability to conduct independent research. In this way, PGI can find unique investment insights and, thus, provide a customized investment strategy (Principal Global Investors, May 2016).

The Chinese component of PGI is experiencing an increase in demand from enterprise Chinese clients for an investment strategy that involves Chinese Equities. These clients require an absolute return strategy that can hedge against the fluctuating Chinese market by using key macro and micro economic indicators to predict short term market movement and return a positive investment over the course of a year. PGI currently has a five-factor model in place. This model utilizes fund flow, the economic revision index, the economic surprise index, the Buffet Model and a risk factor to determine market movement and the necessity to keep an Electronically Traded Fund (ETF) as is, move the ETF to cash, or move the ETF to another asset. However, this current model does not meet the needs of the client as it does not create a return that is equal to or greater than the returns of the ETF or the Index the ETF follows.

Chinese Economy

With a population of over 1.379 billion and a Gross Domestic Product (GDP) of $11.19 Trillion in 2017 (Regular Press Release Calendar of NBS in 2018, 2017), the Chinese economy is the largest agricultural and manufacturing economy in the world. The Chinese primary industry is agriculture and its secondary industry is construction and manufacturing. It has a tertiary or third major industry of services. This service sector is a promising avenue for research because China has a burgeoning middle class. By example, in 1990, the percent of China's population living on less than $1.90 a day was 66, in 2016, that percentage had decreased to just 1.9 percent (Regular Press Release Calendar of NBS in 2018, 2017) of the Chinese population. Empowered by economic growth, ingenuity and disposable income, the middle class has exerted its purchasing power. The middle class is buying real estate, going to the movies, buying tech gadgets and designer clothes and doing other activities that boost the services market. Changes in the middle class could lead to movements in the economy and the underlying equities markets.

Research in the services sector and other aspects of the economy could produce macroeconomic indicators of the Chinese equities market. Some of these indicators include the US-China Exchange rate, interest rates, inflation rates, real estate investments and sales, GDP, Manufacturing Purchasing Manager's Index (PMI), the non-manufacturing or services PMI, real estate investments and sales, employment rate and a host of others. Given the rebalancing of the Chinese market from agriculture to consumption (Connett, 2018), there is much turbulence and so the economic indicators available will have to be researched and tested, jointly and individually, to see if we can find predictive elements.

Chinese Stock Exchange

The Chinese stock market was founded more than 150 years ago and is currently divided into several exchanges where different types of socks traded. For this project, we will focus on ETFs traded on the Shanghai stock exchange as we seek to find an absolute return algorithm for Chinese based investors. Although Chinese A and B shares are traded on the Shanghai stock exchange, we will focus on Chinese A shares as these shares are part of mainland China and are traded in the local currency of Renminbi. These stocks can only be purchased by Chinese investors.

The Chinese stock market differs from the U.S and European markets in two key ways (China's Stock Markets vs U.S. Stock Markets, 2015):

The Chinese stock market is dominated by private investors or 'retail punters.' For most developed countries, the majority of investments in the market are made by large institutionalized corporations that spread credit across the market. These investors are also interested in long-term investments, making the market less volatile. However, 80 percent of China’s investments come from individual investors. “Many of these investors have utilized margin lending by brokerage firms in order to leverage thei accounts and enhance overall returns, despite margin interest rates as high as 20%. These dynamics have made the Chinese markets inherently more unstable than their global counterparts across the developed world.” (Kuepfer, 2017) The instability arises when sentiment in the market changes and investors quickly buy or sell stocks in large overall quantities as a reaction to the change in market sentiment.

The policies of the Chinese government as well as the preponderance of state-owned large cap companies has been known to have a large influence over the stock market. When new policy is released by the government, investors react by buying or selling large amounts of stock, this has led to turbulence in the market. As well, the stock market is turbulent ahead of an initial public offering (IPO) as the government has a tight control over IPOs and, effectively, sets the price. This results in investors pulling cash from the market prior to an IPO and then pushing the cash back in after (Noble, 2015). As well, the state-owned large cap companies dominate the market due to the tight parameters needed to enter the market and the tight parameters needed to exit the market. The presence of these corporations has inhibited the entrance of privately owned corporations. The overall state of the stock market is not an enticing environment for large enterprise investors thus making it an unstable market.

Given the above information, we need to create a multi-factor absolute return algorithm that finds and extorts the relationships among micro and macroeconomic indicators and the market. We must also find micro economic trends within the market and find a way to predict how the macroeconomic indicators influence the micro economic decision making of the investors. The intricate weaving and balancing of the consumer, the economy, and the stock market must be considered when working to develop a market timing model with sustained positive returns.

Primary User Stories

Based on PGIs need for a dynamic absolute return portfolio with an embedded market signal, we have developed the following primary user stories for our project:

- As an investor, I want a Chinese based portfolio of ETFs that provide at least an 8 percent yearly return on investment.
- As a consultant, I want to provide personalized investment advice based on sound macroeconomic and microeconomic indicators and based on historical and current movement in the U.S., Chinese and global stock markets.
- As a business owner of PGI, I want to increase business from Chinese clients by providing a portfolio of Chinese investments that meets the specific goals of my clients.
- As a data analyst, I want to assist consultants in predicting market movement by researching, selecting, and weighting macro and micro economic indicators of a major ETF index and based on signal, build a model that optimizes ETF portfolio.

Solution Space

New Vision

Our research seeks to build a multi-factor market timing model that allows prediction of weekly movement in an ETF portfolio. We will produce an absolute return portfolio with a yearly rate of return greater than 8 percent but at the least, greater than the stock index that the ETF follows.

The model includes a four-step process:

Signal. Create a trade signal based on research and analysis of macro, micro economic indicators and indices. The trade signal should be weighted and consider indicators that have a more prominent influence on market trends.

Trend. Identify an overall index or make combination of indices to select ETFs. The ETFs will come from the mainland China stock exchanges and will most likely include the two main ETFs of those exchanges. For this project that is the HSI and FTSE. The rate of return of the MXCN during the time period Jan 2010 to Apr 2017, was -8.79%, a benchmark for improvement.

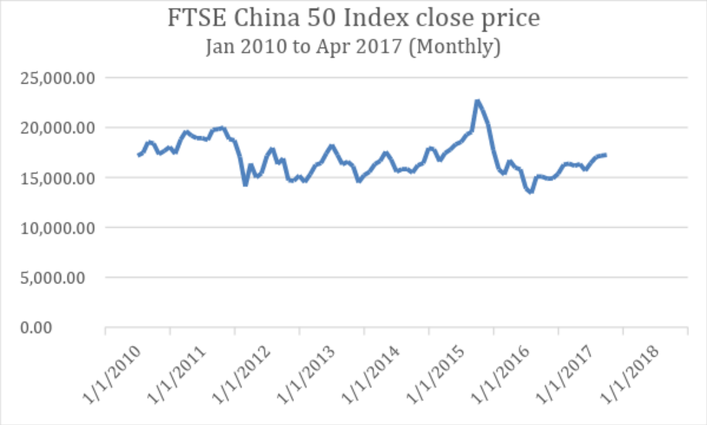


Figure 1. Time Series Performance of FTSE China 50 Index

Combination. Combine the signal and trade together to build a portfolio. The signal will tell us whether we should keep ETFs as they are, move to other assets or move to cash. This will be done on a weekly basis.

Final solution. Balance risk and annual return rate 8% (or more). Compare final model against previous model.

Product Vision

Scenario #1 - Our Vision for PGI

Our vision for this product is to provide PGI a vehicle to connect with Chinese clients and provide sustained growth in the stock market. We envision this product being used to build a formidable relationship with Chinese clients by providing an absolute return fund that is backed by research and data science.

Scenario #2 - Our Vision for Investors

Our vision for Chinese individual (private) investors and institutional investors is to provide a data science-based solution to maneuvering a volatile market. We envision this product as a guideline for choosing macro and micro economic indicators that could provide predictive insight into the equities market.

Dataset

Overview

This section provides insight into how we selected factors, the story behind our dataset, the descriptive statistics of our sample, and the sources of the sample. We apply three main filters to obtain our sample data. First, we chose Chinese domestic macroeconomic indicators, obtained from myriad sources but most frequently from the National Bureau of Statistics of China (NBSC). Second, we acquired related international macroeconomic indicators, to show multi-tiered and outside influences on the Chinese market. Data were gathered from the Federal Reserve and other sources. Third, we restricted our analysis to Chinese and World Economy from Feb 2013 to Feb 2018 as a five-year period will show several bull and bear markets.

Table 2 presents descriptive statistics of daily, monthly and quarterly macroeconomic indicators of China, with mean, standard deviation, median, minimum value, maximum value, range and skewness of the attribute. We find the smallest standard deviation of those attributes is 0.27, which is *Daily Exchanges* between USD and CNY; and the largest standard deviation is 2.04*10^13, which is M2. Table 3 presents descriptive statistics of daily, monthly and quarterly indicators from macroeconomic indicators of the United States, with mean, standard deviation, median, minimum value, maximum value, range and skewness of the attribute. We find the smallest standard deviation of those attributes is 0.29, which is inflation rate of United States; and the largest standard deviation is 4136, which is the *Balance of Trade.* Table 4 presents descriptive statistics of major stock indices from China Shanghai Stock Exchanges, and Shenzhen Stock Exchanges, with mean, standard deviation, median, minimum value, maximum value, range and skewness of the attribute.

These values allow us to test the correlation among pairs of different attributes and study the relationship between the features and response variables.

Table 1. Summary statistics of Chinese Macroeconomic Indicators

variable	***mean***	***std***	***median***	***min***	***max***	***range***	***skew***
ETF_VOLATILITY	24.86	6.25	23.85	15.09	58.4	43.31	1.36
DAILY_EXCH	6.40	0.27	6.27	6.04	6.96	0.92	0.52
LPR	4.83	0.63	4.30	4.30	5.75	1.45	0.51
CHINA_BOND_YTM	2.99	0.60	3.09	1.64	4.25	2.61	-0.11
CHIEPUINDXM	228.07	150.41	174.22	40.40	694.85	654.45	1.15
MANU_CI	99.07	0.36	99.03	98.44	99.77	1.33	0.13
M0 (*10^{12})	6.26	0.62	6.15	5.41	8.66	3.25	1.15
M1 (*10^{13})	3.89	0.73	3.55	2.96	5.36	2.39	0.56
M2 (*10^{14})	1.33	0.20	1.34	0.99	1.67	0.67	0.06
M3 (*10^{14})	1.33	0.20	1.31	1.00	1.67	0.66	0.06
RESERVES_DOLLARS (*10^{12})	3.50	0.33	3.53	3.02	4.01	0.99	0.03
CPI	115.05	2.73	114.80	110.30	119.70	9.40	-0.01
EXPORTS (*10^{11})	1.86	0.23	1.89	1.14	2.28	1.13	-1.26
IMPORTS (*10^{11})	1.49	0.19	1.49	9.36	1.83	0.89	-0.38
M_PMI	50.67	0.74	50.55	49.00	52.40	3.40	0.09
NONMAN_PMI	54.23	0.75	54.10	52.70	56.30	3.60	0.53
TSALE_AMOUNT	37884.24	28006.32	34314.19	0.00	110239.50	110239.50	0.54

TSALE_RATE	19.59	24.45	16.60	-16.70	87.20	103.90	0.55
PRIME_Industry_Value	14274.92	5258.37	14088.30	6725.80	24238.60	17512.80	0.13
Second_Ind_Value	69553.56	12324.10	67427.65	51321.70	102594.00	51272.30	1.07
Tertiary_Ind_Value	68332.12	30625.03	73497.25	8872.40	113080.50	104208.10	-0.76

Table 2. Summary statistics of US Macroeconomic Indicators

variable	***mean***	***std***	***median***	***min***	***max***	***range***	***skew***
US_INF_RATE	1.89	0.29	1.87	1.21	2.57	1.36	0.09
US_BAL_TRADE	-42279.10	4136.41	-41624.80	-56938.20	-34330.80	22607.32	-0.79
US_GDP	17894.79	954.68	18008.78	16259.70	19638.24	3378.54	-0.07
US_FED_DEBT	102.43	1.76	102.12	99.35	105.66	6.30	0.16
US_ST_INT	0.30	0.40	0.07	0.00	1.46	1.46	1.41
US_LT_INT	2.25	0.348	2.27	1.38	3.01	1.63	-0.21

Table 3. Summary statistics of Chinese Stock Indices

variable	***mean***	***std***	***median***	***min***	***max***	***range***	***skew***
SSE	2907.29	655.82	3052.78	1950.01	5166.35	3216.34	0.47
SSEA	3072.25	687.96	3221.93	2040.67	5410.86	3370.19	0.40
SSEB	346.17	43.49	345.41	255.03	536.09	281.06	1.38
SSECI	2807.14	363.26	2727.61	2241.23	4363.07	2121.84	1.94
SSEIN	2561.20	566.76	2418.50	1917.38	4837.49	2920.11	2.02
SSEMC	4368.10	860.76	4224.21	3039.15	8038.34	4999.19	1.65
SSECBI	189.65	17.71	190.78	161.10	214.26	53.16	-0.11
SSE180	6887.30	1579.65	7049.50	4545.14	11741.07	7195.93	0.36
SZSEC	10106.24	1962.23	10262.83	6998.19	18098.27	11100.08	0.93
CNIB100	6088.77	1018.59	6275.58	3817.17	10205.21	6388.04	0.29
CHINEXTC	2096.49	747.10	2282.50	789.01	4414.78	3625.77	0.08
SZ700I	6799.21	2132.67	7346.92	3349.42	14003.09	10653.67	0.18
SZ500LVI	1617.76	476.35	1804.27	817.63	3161.03	2343.40	-0.01
SZAI	6297.29	1225.74	6571.38	4304.93	10664.75	6359.82	0.27
CNILC	2890.63	656.69	2995.29	1940.60	5099.00	3158.40	0.47
CNIMS	3977.44	1046.61	4152.30	2346.73	8136.50	5789.77	0.73
CCTV50	5711.58	1296.95	5870.20	3737.46	10482.42	6744.96	0.60
SZMI	1719.09	498.94	1904.70	924.03	3205.32	2281.29	-0.04

Macroeconomic Indicators

China

In any given country, there are major measures of economic activity. This economic activity fluctuates within a given time period and these fluctuations in the economy set the tone for the market and help businesses make fiscal decisions. Since the activity occurs on the macro or broad level of an economy, we refer to the activity as macroeconomic activity and use the activity as macroeconomic indicators of market movement. (Young & McAuley, Apr 1994)

China has several macroeconomic indicators that were considered for data analysis as these indicators have been identified as having a potential impact on market activity and business decisions. The data are chosen for three reasons:

1. 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.
2. 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.
3. Research Value: Data have been researched and identified as having potential to help predict changes the economy.

Macroeconomic data were pulled from several sources including the Federal Reserve of the United States (FRED), the National Bureau of Statistics of China (NBSC), the Organization for Economic Co-Operation and Development (OECD), Sina Finance, and the People’s Bank of China. Data were conditioned to extrapolate most relevant indicators, cleaned to impute or delete missing values and formatted into a weekly value for comparison and analysis.

The People’s Bank of China, under the guidance of the State Council of China formulates and implements monetary policy. It sets goals for Gross Domestic Product (GDP), Consumer Price Index (CPI), Money Flow (M2) and provides guidance on the Interest rate, bond maturity rate, and exchange rate. These indicators have, therefore, been selected for inclusion in our study (Focus Economics, 2018). Data are provided by the People’s Bank of China, Sina Finance, and FRED.

The Purchasing Managers Index, both manufacturing and non-manufacturing (services) will be used in our study. These indicators are used as early indicators of the health of the manufacturing and services industries. Data are provided by the NBSC.

As China moves to a consumer-based economy and strengthens its middle class, the real estate index has become ever more important. Data on the real estate market is provided by NBSC.

Microeconomics is “the study of buyers and sellers in individual markets within a larger economy” (Young & McAuley, Apr 1994) that seeks to solve the problem of scarce resource allocation. Thus, for our project, we will create variables from the Chinese Indices data that will allow us to monitor the supply and demand of stocks and the consumer or private investor reaction to movements within the market. We will create a 10, 20, and 200-day simple moving average. These averages are divisible by 5 which reflects the 5-day stock market week. We will also analyze and create variables for the stock volume, and the stock open, close, high and low values.

International (USA as proxy)

The US dollar is considered the global currency and is the commonly accepted currency for international trade (Amadeo, 2018). The US dollar makes up almost 64 percent of all known central bank foreign exchange reserves and more than 85 percent of forex trading involves the US dollar. The most important fact about the US dollar is that 39 percent of the world’s debt is issued in dollars. From the above data we can see the importance of US dollar and it is the world’s largest economy and is on track to break GDP of US 20 trillion dollars in 2018. China is the second largest economy in world with GDP forecasted to reach more than 13 US trillion dollars in 2018. United States is China’s largest trading partner and based on that we use US economic variables as a proxy for global economic activity (Nguyen & Sayim, 2016).

There has been many research done in past to study if US economic variables affect the Chinese stock market and one such study was done by J.C. Goh et al. at Singapore University. They provide us with three reasons why investigating the effect of US economic variables on the Chinese market is important (Jareno & Negrut, 2015).

1. Studying the relation sets up solid benchmark for investors like PGI whose goal is to focus on the Chinese market and increase their returns (Valukonis, 2014).
2. Researching the effect of US variables on the Chinese stock market would improve our understanding of return predictability across different countries (Goval & Welch, 2004). Once we understand how the US economic variables affect China, we can research variables from different countries and try to incorporate them in future models (Goh, Jiang, & Tu, 2011).
3. This study can also provide us with more information about the importance of US factors that could have an effect on the cross-sectional returns of the Chinese stock market.

The data for US indicators was obtained from Federal Reserve Bank of Saint Louis and OECD.org. We have selected six US indicators to see if there is any relation with China.

1. GDP
2. Inflation Rate
3. Balance of Trade
4. Federal Debt
5. Long Term Treasury Interest Rate
6. Short Term Treasury Interest Rate

Microeconomic Indicators

Microeconomic indicators focus on business itself. Just as sales are determined by the supply and demand of products. In our project, the microeconomic indicators are those indicators related to the stock market and the ETF trading market.

Chinese Indices

In China, there are three major stock exchanges: Shanghai, Shenzhen, and Hong Kong. The Shanghai Stock Exchange is the 2nd largest stock market in Asia, and Hong Kong is the 3rd largest. The major stock indices (Major Stock Indices Definition, 2018) summarize the performance of major stocks, classified by the exchange on which they trade, by region. Investors use stock indices as a benchmark to tell the overall performance of the stocks. In the beginning of our research, we selected 8 indices from Shanghai Stock Exchange, 10 indices from Shenzhen Stock Exchange, and 5 from Hong Kong Stock Exchange, by different industries. We listed the indices of interest in the lexicon appendix.

ETF(exchange trade fund):

The specific ETFs are what the project owner will trade in the portfolio. Like stocks, ETFs also experience price changes during the day. There are several trading indicators for those ETF investors review and use (Ho, 2012).

The first one is Moving Average Lines: Investors can easily tell if an ETF is in an uptrend or a downtrend based on where it's trading in relation to these two lines.

The second one is week number: In sales, prices change with the season; this is called seasonality. The same condition exists for ETF prices. Since we want to do weekly predictions, week number is a good choice to show seasonality.

The third one is trading volume: ETFs with low trading volume means it has few assets under management, and it tends to have wider bid-ask spreads than more liquid ones.

Lastly, candlesticks (The History of Japanese Candlesticks, 2010). The real body of candlestick is a great starting point, because we can get a lot of information from it. When bodies become larger, it shows an increase in momentum, which leads to a bullish decision. When bodies become smaller, it shows slowing momentum, which lead to a bearish decision.

For the response variables, we chose FTSE China 50 index as the ETF tracking index. The FTSE China 50 Index measures the performance of Large Cap securities and is selected by a Hong Kong.

Field Descriptions

Our final dataset consists of both the Chinese and USA indicators. Below are the descriptions of the various data fields (See Appendix A for details):

- Survey.OECD (numeric dataset, benchmark: 100): an indicator provided by the Organization of Economic Cooperation and Development (OECD). It monitors the key economic indicators or a country and provides a level of confidence in a market based on Consumer Opinion Surveys, Confidence Indicators, and Composite Indicators.Export/Import (numeric dataset, in billion): this fields are also numeric dataset and represent the value of the goods of China
- PMI.MANU/PMI.NONMANU: The Purchasing Managers' Index (PMI) is an indicator of the economic health of the manufacturing sector. The PMI is based on five major indicator they are new orders, inventory levels, production, supplier deliveries and the employment environment in various manufacturing and non-manufacturing companies of China

- M0 : Narrow Money. This field indicates the various currency details that are being held by the banking institutions of China. It is a measure where both the cash and liquid assets are being combined and held by the central bank of China.
- M1 : It is a metric for the money supply of a country that includes physical money in form of coin or paper. Most of the money that is in liquid portion of the money supply are being measured and indicated in this field.
- Policy.uncertainty : Is a class that includes the economic risk involved for various companies while investing in the Chinese stock market that eventually leads in delay and lower amount of investment percentage. Policy uncertainty may also indicate uncertainty over electoral outcomes that will influence political leadership.
- MANU.CI (manufacturing confidence indicators)/MANU.EX (manufacturing exchange) : It is a statistical indicator based on the results from various company firms. It also considers other factors such as various construction services, retail trade and consumers.
- Building sale/Building sale rate (numeric dataset, in percentage): This only indicates the total sale involved in the construction sector of the industries. It is an overall gives the value of total buildings sold and the complete commercialized residential growth rate.
- CPI (Consumer Price Index) : It is the benchmark for increase in the Chinese economy. It records the main products that are being purchased every year/ monthly basis based upon the industries.

- CPI.Residence/Transport/ Household/Health/Education : This fields usually indicates the various CPI involved in the residence, transport and other sectors in the economic development of China.
- SSE.10D.MA/SSECBI.10D.MA : The SSE Composite Index also known as SSE Index is a stock market index of A shares that are traded at the Shanghai Stock; it shows the 10-day simple moving average.
- Exchange. Rate : This field indicates the exchange rate between the US and the Chinese financial .
- ETF (exchange traded fund) : It is an investment fund that tracks an index, commodity, bond or basket of assets. It can be traded on a stock exchange just like other stocks and experiences price changes throughout the day based on sales (Exchange-Traded Funds, 2018).
- ETF.VOLATILITY : This field describes the update performed by the CBOE Volatility Index—VIX for short. It measures what is known as implied volatility, which is calculated based on the price of the ETF. VIX allows the traders to bet on what value of the ETF will be on the same date in future years. CBOE lists a number of weekly and monthly VIX futures contracts whose values fluctuate based on where traders believe the level of the VIX will be at the contract's expiration date.
- BOND.YTM (BOND YIELD TO MATURITY) : China Bond Yield to maturity (YTM) is the total return anticipated on China bond if the bond is held until it matures..
- GDP.US/GDP.CHINA : The real economic growth rate measures economic growth, in relation to gross domestic product (GDP), from one period to another, adjusted for

inflation - in other words, expressed in real as opposed to nominal terms. The real economic growth rate is expressed as a percentage that shows the rate of change for a country's GDP from one period to another, typically from one year to the next. this basically gives the rate between description of GDP of United States and China.

- CN.Prime.Ind.Value/CN.Tertiary.Ind.Value/CN.Second.Ind.Value : This field indicates the various Chinese primary, secondary and tertiary industries values. The Chinese primary industry is agriculture and its secondary industry is construction and manufacturing. It has a tertiary or third major industry of services
- US_INF_RATE : A measure of the rising price of goods and services in the US and the subsequent fall of purchasing power of the USD.
- US_BAL_TRADE : The balance of trade (BOT) is the difference between the value of a country's imports and its exports for a given period. The balance of trade is the largest component of a country's balance of payments (BOP). Economists use the BOT as a measure of the relative strength of a country's economy. The balance of trade is also referred to as the trade balance or the international trade balance.
- US_FED_DEBT : The net accumulation of the federal government's annual budget deficits: It is the total amount of money that the U.S. federal government owes to its creditors.
- US_ST_INT : Short-term interest rates are the rates at which short-term borrowings are affected between financial institutions or the rate at which short-term government paper is issued or traded in the market. Short-term interest rates are generally averages of daily rates, measured as a percentage. Short-term interest rates are based

on three-month money market rates where available. Typical standardized names are "money market rate" and "treasury bill rate”.

- US_LT_INT : Long-term interest rates refer to government bonds maturing in ten years. Rates are mainly determined by the price charged by the lender, the risk from the borrower and the fall in the capital value. Long-term interest rates are generally averages of daily rates, measured as a percentage. These interest rates are implied by the prices at which the government bonds are traded on financial markets, not the interest rates at which the loans were issued. In all cases, they refer to bonds whose capital repayment is guaranteed by governments. Long-term interest rates are one of the determinants of business investment. Low long-term interest rates encourage investment in new equipment and high interest rates discourage it. Investment is, in turn, a major source of economic growth
- FTSE.Price response variable): The FTSE China 50 Index measures the performance of Large Cap securities and is selected by a Hong Kong-listed process.

Data Context

The goal of this project is to use micro and macroeconomic indicators, both Chinese and worldwide,to create a market signal that predicts the movement of a major Chinese index and gives weekly instructions for trading the ETFs that follow that specific index.

Macroeconomic data are provided on a daily, weekly, monthly or even yearly basis. Therefore, depending on the frequency of the data, macroeconomic indicators will be used in several ways:

1. Yearly data will be used to set the tone for the type of market we could be entering. For instance, for the macro economic indicator of GDP, if GDP experiences or is forecasted to experience a decline, we could be entering a bear market or a market in which prices

are declining, encouraging sales. Likewise, if the GDP increases or is forecasted to increase, we could be entering a bull market or a market in which prices are increasing, encouraging purchases.

2. Monthly data will be measured based on time series release date and its immediate impact on the equities market. It will be used to help predict shorter term market movement and shorter-term investment strategies. For instance, an increase in the Non-manufacturing (Services) PMI from 50 to above 50 could indicate that consumers are happy and purchasing power is increasing.
3. Daily data will be used as a weekly indicator of variations in the macro economy. The data will be used to make short term investment decisions. For example, a decrease in the interest rate is used to encourage consumer/business spending and bank lending. Therefore, this could be a boost to the economy and we may be able to predict upward movement in the equities market based on an increase in the interest rate.
4. The actual release date of yearly and monthly data will also be included so that we may explore whether there are immediate changes in the stock market when certain indicators are released to the public.

Micro economic indicators are classified as either a Western indicator, popularized by the Western world, or an Eastern indicator, popularized by Japan or other.

Western indicators include the 10, 20, and 50-day simple moving average (SMA). A simple moving average is calculated by totaling the closing price of a security over a set time and dividing the total by the number of time periods. They are used to evaluate or predict market trends. Thus, the act of a short-term SMA crossing a long-term SMA signifies an uptrend or a downtrend in the market.

We will classify our Eastern indicator as a candlestick. The candlestick indicator was created by rice farmers in Japan in the 16th century. Rice was traded, used and sold as the major currency in Japan. Rice futures were sold at an astonishing rate to the degree that at one point 130 thousand barrels were sold on the market while only 30 thousand were physically in circulation. The candlestick used historic price moves and weather to predict rice sales. It has gained popularity in the United States over the last 25 years (The History of Japanese Candlesticks, 2010). Candlestick indicators are created by using the close and open price of a stock, coupled with the high and low cost of a stock. Once the candlestick is created, it is used to find relationships amongst other candlesticks. Data can be evaluated on a daily, weekly or even monthly basis. For our analysis, we will evaluate the chosen Chinese index on a weekly basis.

We will also analyze stock volume and seasonality.

Most of the US indicator data was collected by the Organization for Economic Co-operation and Development (OECD) which was founded in 1961 with the goal of promoting different policies that would lead to improvement of social and economic well-being of people all around the world. Their goal is to help governments and countries to find solutions to problems which affect people’s day to day life. And they do this by analyzing data and predicting trends and in this process, they collect and share the data with public. We obtained our short and long-term interest data was obtained from OECD.

Second source for our US indicator data is The Federal Reserve Bank of St. Louis who collects and publishes data to help Bank president make important economic decisions. Per them “This site offers a wealth of economic data and information to promote economic education and enhance economic research. The widely-used database FRED is updated regularly and allows 24/7 access to regional and national financial and economic data”. Data for GDP, Inflation Rate, Balance of Trade, Federal Debt and Interest Rate was obtained from them.

The US Indicator data provides us with valuable information about the US economy and how it can be used in relation to Chinese economy and predict if they influence the Chinese market. Let’s take the example of Balance of Trade which tells us the difference in value between US’s imports and exports. Currently China is our largest trading partner and using this relation to predict market movement can be beneficial.

Data Conditioning

Data conditioning is an often undervalued but important step in the data mining process. Since data gathering methods are loosely controlled, resulting in out-of-range values, missing values, impossible combinations (e.g., Gender: Male, Pregnancy: Yes), etc. Analyzing data without conditioning can produce misleading results. Thus, the representation and quality of data is first and foremost before running analysis.

Data preprocessing methods are divided into the following categories (Indian Agricultural Statistics Research Institute, 2012): data cleansing, data integration, data transformation and data reduction.

- Data cleansing is to clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistency.
- Data integration combines data from multiple sources into a coherent dataset. These sources may include multiple databases, data cubes, or flat files. We combined all our time series data by date.

- Data transformation aims to transform or consolidate data into forms appropriate for data mining, which may involve Normalization, Smoothing, Aggregation and Generalization. Complex data analysis and mining on huge amounts of data may take a very long time, which makes analysis impractical or infeasible.
- Data reduction techniques help reduce representation of the dataset without lacking original information and help produce the quality knowledge. Data reduction is commonly understood as reducing the volume or reducing the dimensions, which means fewer number of attributes.

Data Cleansing

Missing Value

Most the daily data we have are related to the stock market. Generally, all stock markets in the world close on holidays and weekends. In the US, the NYSE and NASDAQ average about 252 trading days a year. However, in China, the holidays are quite different from what we have, so we cannot use US trading days to judge. To impute missing values, we used the SMA (simple moving average) for 10 days, which could treat both weekends and 7-day holidays well.

$$10 - day\ SMA = \frac{p_D + p_{D-1} + \cdots + p_{D-9}}{10}$$
$$= \frac{1}{10}\sum_{i=0}^{9} p_{D-i}$$

After 10-day SMA conversion, we could get the true daily data, instead of the original trading days' data.

The attribute, *ETF Volatility*, is a daily measure. The root of ETF volatility is based on the Volatility Index, which is the fear index for the ETF trading market. It has the same trading days as the ETFs. We chose VLOOKUP function in excel to fill the missing values based on the proximity match, which is to find the closest previous value to take over the missing value.

Most monthly indicators had no missing values. However, we found one indicator with missing values over 50%, so we decide to delete this one.

Data Transformation

Lag Days

Lag Days = Data Public Release Data - Data Observation Date

A recognition lag (Recognition Lag, 2018) is the time lag between when the financial data are accounted for, and when the economic data are released to the public. We found that the more frequently the data are gathered, the fewer number of lag days. We discovered the lag days in two ways. One is to refer to the public news press as it is the most authoritative way for the public to retrieve information on data releases. The second is to locate the data resource, and find the date the data were released and the date the date were 'refreshed' or appended to the website. These methods are not a perfect science.

For example, the United States has a public news press to announce the GDP for the 4th quarter and annual of the year 2017 on January 26th, 2018. The data range is from October 1st to December 31st. So here we have a 26-day lag for this information.

Lag days were created to ensure that any trends found between indicators and the predictor variables are honest and reflect the actual date the data were released to the public. This is important in measuring the impact of a variable on the predictor variable. We do not want to assume that a variable had an impact on a predictor variable based on the observation date, when in fact, the data had not been released to the public on the date the impact was measured or perceived.

Triangular Conversion

Daily Economic Indicators. We go by the mean of triangular distribution (hi-tech, 2012) and converted ETF Volatility and China Bond YTM (Yield-to-Maturity) in Excel. For a certain week, we could find the minimum, maximum and approximate mode, and based on a random number between 0 and 1, we created our weekly dataset.

Spline Conversion

Monthly and Quarterly Indicators. We did Cubic Interpolation (Oscar, July 2014) to convert quarterly basis to weekly. Cubic Interpolation is a kind of spline interpolation, which is widely used for the financial data. We refer to the slides from Data & Statistical Services of Princeton University, and create the dataset in R.

Data Reduction

Complex data analysis and mining on huge amounts of data may take a very long time, making such analysis impractical or infeasible. There are several methods that have facilitated in analyzing a reduced volume or dimension of data and yet yield useful knowledge. Mining on the reduced dataset should be more efficient yet produce the same (or almost the same) analytical results.

Correlation Check to remove multicollinearity

Multicollinearity is a state that attributes are highly correlated with others, dependent on others. We may get duplicated information from highly correlated attributes. It is hard to do analysis with correlated attributes.

Using highly correlated features in a regression or classification model is not recommended. The first step in removing redundant features is to find the correlation coefficient. The basic rule of highly correlated cut-off is 0.7, which means if the correlation coefficient locates between (0.7, 1) or (-1, -0.7), those two features are highly correlated, we need to decide which one will be delete.

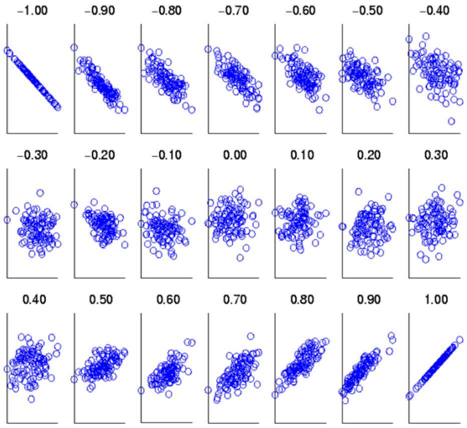


Figure 2. Examples of Correlation Coefficients

Data Quality Assessment

Completeness

Definition	The proportion of stored data against the potential of 100% complete (DAMA UK Working Group, October 2013)
Measure	A measure of the absence of blank (null or empty string) values or the presence of non-blank values
Scope	0-100% of critical data to be measured in any data item, record, data set or

	database
Unit of Measure	Percentage
Example	For our response variable, FTSE China 50 index, from our data source Investing.com, we have two options, one is FTSE China 50 based on USD, the other one is FTSE China 50 based on HKD. If we want to download daily data, we could only get data from Oct 15th, 2013 to Feb 7th, 2018 based on USD, convert to percentage, it is 1576 days/1826 days (our desired time range) = 86.31% completeness has been achieved for this data item. However, we could reach 100% completeness based on HKD.

Uniqueness

Definition	Nothing will be recorded more than once based upon how that thing is identified.
Measure	Analysis of the number of things as assessed in the real-world compared to the number of records of things in the data set. The real-world number of things could be either determined from a different and perhaps more reliable data set or a relevant external comparator.
Scope	Measured against all records within a single data set
Unit of Measure	Percentage

Example	Shanghai stock exchanges has 259 indices (Shanghai Stock Exchanges, 2018) by different sections. Those sections are size, sector, style, thematic, strategy, bond, fund, customization, and dividend point indices. However, when we look up those indices, there are 310 indices we may get. There should be some duplicated ones. Therefore, this indicator uniqueness is 259/310 = 83.55%

Accuracy

Definition	The degree to which data correctly describes the real-world object or event being described.
Measure	The degree to which the data mirrors the characteristics of the real-world object or objects it represents.
Scope	Any "real world" object or objects that may be characterized or described by data, held as data item, record, data set or database.
Unit of Measure	The percentage of data entries that pass the data accuracy rules.
Example	In Chinese culture, the date format, entering to any database, is YYYY-MM-DD or DD/MM/YYYY rather than the US MM/DD/YYYY format, causing the representation of days and months to be reversed. As a result, 09/08/YYYY really meant September 8th or August 9th is the problem.

	The representation of date, we got data from Chinese sources– whilst valid in its US context–means that we have to check the adjacent dates to make sure data have an accurate date, other than simply transform the format by as.Date() function in R.

Atomicity

Definition	Atomicity is one of the features from ACID of database system, which means databases systems dictating where a transaction must be all-or-nothing. That is, the transaction must either fully happen, or not happen at all.
Measure	The definition of what constitutes an atomic transaction is decided by its context or the environment in which it being implemented.
Scope	Database system feature
Unit of Measure	0 or 1, which represents the transaction happens for not.
Example	In our project, business logic dictates that we cannot make more than one trade in a single week, to maintain a low turnover rate. If one happens without the other, problems can occur. For example, high turnover rate may cause high trading fee which cannot be afforded by the business.

Conformity

Definition	Are there expectations that data values conform to specified formats? If so, do all the values conform to those formats? Maintaining conformance to specific formats is important in data representation, presentation, aggregate reporting, search, and establishing key relationships.
Measure	Conformity means the data is following the set of standard data definitions like data type, size and format.
Scope	Maintaining conformance to specific formats is important
Unit of Measure	Percentage
Example	Possible values for exchange rate between CNY and USD are from 6 to 7, the data for this value cannot be a negative value, or go higher/lower than a certain value.

Overall Quality

Data quality refers to the overall utility of a dataset as a function of its ability to be easily processed and analyzed for other uses, usually by a database, data warehouse, or data analytics system. Quality data is useful data. To be of high quality, data must be consistent and unambiguous. Data quality issues are often the result of database merges or systems/cloud integration processes in which data fields that should be compatible are not due to schema or format inconsistencies. Data that is not high quality can undergo data cleansing to raise its quality.

Data quality activities involve data rationalization and validation. When data is of excellent quality, it can be easily processed and analyzed, leading to insights that help the organization make better decisions. High-quality data is essential to business intelligence efforts and other types of data analytics, as well as better operational efficiency.

Other Data Sources

Please refer to Appendix A- Lexicon to find other data sources and definition sources.

Analytics and Algorithms

Overview

Various cleaning techniques were used to finalize the dataset used for modeling. Next, myriad regression and classification modeling techniques were tested to find the best model that would determine the indicators that have the most impact on our dependent variable. A small amount of noise was added to the data to check the robustness of each methodology.

Cleaning

Normalization

Normalization means adjusting values measured on different scales to a common scale. It will be easier for us to compare data from of different magnitudes. We use the scale() function in R to run Z-score normalization for us.

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

where max(x) is the maximum value of sample dataset, and min(x) is the minimum value of sample dataset.

Another way of normalization is to remove skewness of data. as stated in Wikipedia," In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean."

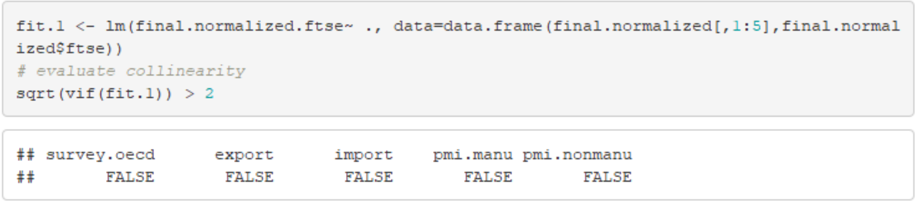
For right skewed data, we choose logarithm transformation. The logarithm, x changes to log base 10 of x, is a strong transformation and can be used to reduce right skewness. For left skewed data, we choose square transformation. The square, x changes to x^2, has a moderate effect on distribution shape and it could be used to reduce left skewness.

Re-check the multicollinearity

In the previous data conditioning part, we viewed the correlation plot to find multicollinearity. Looking at correlations only among pairs of predictors, however, is limiting. It is possible that the pairwise correlations are small, and yet a linear dependence exists among three or even more variables, for example, if $x_3=2x_1+5x_2$+error. That's why many regression analysts often rely on what are called variance inflation factors (VIF) to help detect multicollinearity.

As the name suggests, a VIF (Detecting Multicollinearity Using Variance Inflation Factors, 2018) quantifies how much the variance is inflated. Recall that we learned previously that the standard errors — and hence the variances — of the estimated coefficients are inflated when multicollinearity exists. So, the variance inflation factor for the estimated coefficient bk —denoted VIFk —is just the factor by which the variance is inflated.

For example, in our final dataset, we tried to figure out the paired VIF of several set of our final data. We divided our dataset into six blocks and got the VIFs as following,



```
fit.1 <- lm(final.normalized.ftse~ ., data=data.frame(final.normalized[,1:5],final.normal
ized$ftse))
# evaluate collinearity
sqrt(vif(fit.1)) > 2
```

```
## survey.oecd      export      import    pmi.manu pmi.nonmanu
##       FALSE       FALSE       FALSE       FALSE       FALSE
```

It is used to explain how much multicollinearity (correlation between predictors) exists in a regression analysis. Multicollinearity is dangerous because it can increase the variance of the regression coefficients. A rule of thumb for interpreting the variance inflation factor (Interpreting the Variance Inflation Factor, 2018):

1: not correlated

Between 1 and 4: moderately correlated

Greater than 4: highly correlated

Exactly how large a VIF must be before it causes issues is a subject of debate. What is known is that the more your VIF increases, the less reliable your regression results are going to be. In general, a VIF above 10 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above.

Features

Accordingly, a column representing week number was added to the dataset to determine if seasonality is major factor.

The below are the data features we have after cleansing:

- Survey.OECD:): an indicator provided by the Organization of Economic Cooperation and Development (OECD). It monitors the key economic indicators or a country and provides a level of confidence in a market based on Consumer Opinion Surveys, Confidence Indicators, and Composite Indicators.Export/Import (numeric dataset, in billion): this fields are also numeric dataset and represent the value of the goods of China

- PMI.MANU/PMI.NONMANU: The Purchasing Managers' Index (PMI) is an indicator of the economic health of the manufacturing sector. The PMI is based on five major indicator they are new orders, inventory levels, production, supplier deliveries and the employment environment in various manufacturing and non-manufacturing companies of China
- M0: Narrow Money. This field indicates the various currency details that are being held by the banking institutions of China. It is a measure where both the cash and liquid assets are being combined and held by the central bank of China.
- M1:. It is a metric for the money supply of a country that includes physical money in form of coin or paper. Most of the money that is in liquid portion of the money supply are being measured and indicated in this field.
- Policy.Uncertainty: Is a class that includes the economic risk involved for various companies while investing in the Chinese stock market that eventually leads in delay and lower amount of investment percentage. Policy uncertainty may also indicate uncertainty over electoral outcomes that will influence political leadership
- Manu.Ex : Manufacturing exchange. It is a statistical indicator based on the results from various company firms, especially manufacturing firms.
- Building sale rate: It is an overall gives the value of total buildings sold and the complete commercialized residential growth rate.

- CPI: It is the benchmark for increase in the Chinese economy. It records the main products that are being purchased every year/ monthly basis based upon the industries.
- CPI.Residence/Transport/Household/Educatio: CPI sectors. This fields usually indicates the various CPI involved in the residence, transport and other sectors in the economic development of China.
- SSE.10D.MA:The SSE Composite Index also known as SSE Index is a stock market index of A shares that are traded at the Shanghai Stock; it shows the 10-day simple moving average .
- Exchange. Rate :This field indicates the exchange rate between the US and the Chinese financial
- ETF: It is an investment fund that tracks an index, commodity, bond or basket of assets. It can be traded on a stock exchange just like other stocks and experiences price changes throughout the day based on sales (Exchange-Traded Funds, 2018).
- ETF.VOLATILITY : This field describes the update performed by the CBOE Volatility Index—VIX for short. It measures what is known as implied volatility, which is calculated based on the price of the ETF. VIX allows the traders to bet on what value of the ETF will be on the same date in future years. CBOE lists a number of weekly and monthly VIX futures contracts whose values fluctuate based on where traders believe the level of the VIX will be at the contract's expiration date.
- GDP.CHINA: The real economic growth rate measures economic growth, in relation to gross domestic product (GDP), from one period to another, adjusted for inflation - in other words, expressed in real as opposed to nominal terms. The real economic growth rate is expressed as a percentage that shows the rate of change for a country's GDP from

one period to another, typically from one year to the next. this basically gives the rate between description of GDP of United States and China.

- CN.Prime.Ind.Value/CN.Tertiary.Ind.Value/CN.Second.Ind.Value : This field indicates the various Chinese primary, secondary and tertiary industries values. The Chinese primary industry is agriculture and its secondary industry is construction and manufacturing. It has a tertiary or third major industry of services
- US_INF_RATE : A measure of the rising price of goods and services in the US and the subsequent fall of purchasing power of the USD.
- US_BAL_TRADE : The balance of trade (BOT) is the difference between the value of a country's imports and its exports for a given period. The balance of trade is the largest component of a country's balance of payments (BOP). Economists use the BOT as a measure of the relative strength of a country's economy. The balance of trade is also referred to as the trade balance or the international trade balance.
- US_FED_DEBT : The net accumulation of the federal government's annual budget deficits: It is the total amount of money that the U.S. federal government owes to its creditors.

- US_ST_INT : : Short-term interest rates are the rates at which short-term borrowings are affected between financial institutions or the rate at which short-term government paper is issued or traded in the market. Short-term interest rates are generally averages of daily rates, measured as a percentage. Short-term interest rates are based on three-month money market rates where available. Typical standardized names are "money market rate" and "treasury bill rate

- US_LT_INT: Long-term interest rates refer to government bonds maturing in ten years. Rates are mainly determined by the price charged by the lender, the risk from the borrower and the fall in the capital value. Long-term interest rates are generally averages of daily rates, measured as a percentage. These interest rates are implied by the prices at which the government bonds are traded on financial markets, not the interest rates at which the loans were issued. In all cases, they refer to bonds whose capital repayment is guaranteed by governments. Long-term interest rates are one of the determinants of business investment. Low long-term interest rates encourage investment in new equipment and high interest rates discourage it. Investment is, in turn, a major source of economic growth
- FTSE.Price: The FTSE China 50 Index measures the performance of Large Cap securities and is selected by a Hong Kong-listed process.

Analytics

Data Separation

Time series data presents unique challenges when determining the robustness of a model. It is important to ensure that the final model can function in different financial environments and can work in any time frame. Therefore, we will use four measures of robustness when building our model. For each separation method, several supervised and unsupervised learning techniques will be employed.

The first technique we will employ is the 80/20 split. We will split the data into a training set that consists of 4 years of data or 80 percent of the data. The timeframe of the data is 2/8/2013 to 1/19/2017. The test set will consist of 1 year of data or 20 percent of the data. The timeframe of the test data is 2/2/2017-2/2/2018. There is a one-week buffer between the training set and the test set. The one-week buffer was created so that data is not used in both the training and testing set, creating bias.

The second technique will also employ an 80/20 split. However, the differentiating factor is that this second 80/20 split will be a random set of data points derived from the dataset. The data will not be separated based on a specific timeframe. Splitting the data in this way will give us an idea of whether our data will work over any given time frame. We will split the data into a training set that consists of 80 percent of the data. The test set will consist of 20 percent of the data. A seed will be set on the data so that the results can be reproduced.

The third technique we will employ is ten-fold cross validation. Data will be partitioned into ten equal parts, with 9 folds of the data being used in the training set and 1-fold of the data being reserved as the validation data for testing the model. The unique feature of 10-fold cross validation is that each partition of the data will be used exactly one time as the validation data. The results of each of the validation folds are averaged.

To ultimately test the robustness of our training models, we decided to add noise to the final dataset. Adding small values to datum within a training dataset gives a level of understanding of how well the model performs when the specific environment of the data has changed. For instance, a model that is robust will have the same if not very similar results with and without the presence of noise. There are two datasets used for modeling, one is based on actual dataset with normalization between 0 and 1, the other one is based on weekly changes calculated by original dataset. Given this, we built several rules for adding noise.

Table 4. Rule of Noise Adding

To normalized dataset
● Add certain value (e.g -0.0005, -0.0001, 0.0001, 0.0005) to normalized dataset ● Add random value (e.g. random between -0.0001 and 0.0001) to normalized dataset

To original dataset before weekly change,

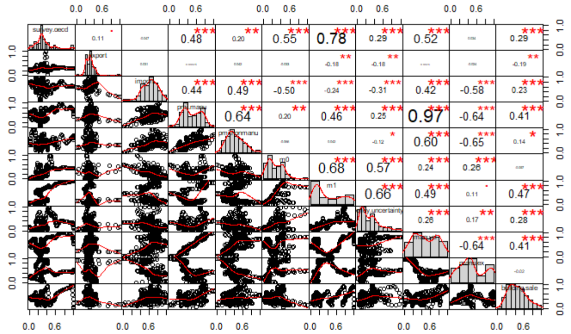
- Add certain value (e.g. -0.5, 0.5) to original dataset, then calculate weekly change
- Add random value (e.g random between -0.5 and 0.5) to original dataset, then calculate weekly change

After that, we have five different datasets for normalized data to test robustness for our best regression model, and three different datasets for weekly changes to test robustness for our best regression model.

Unsupervised learning

Correlation Significant Check

A scatterplot matrix was created to attempt to visually locate significant relationships to include in the model. The matrices did not show significant relationships and so all variables were included in the model for analysis.



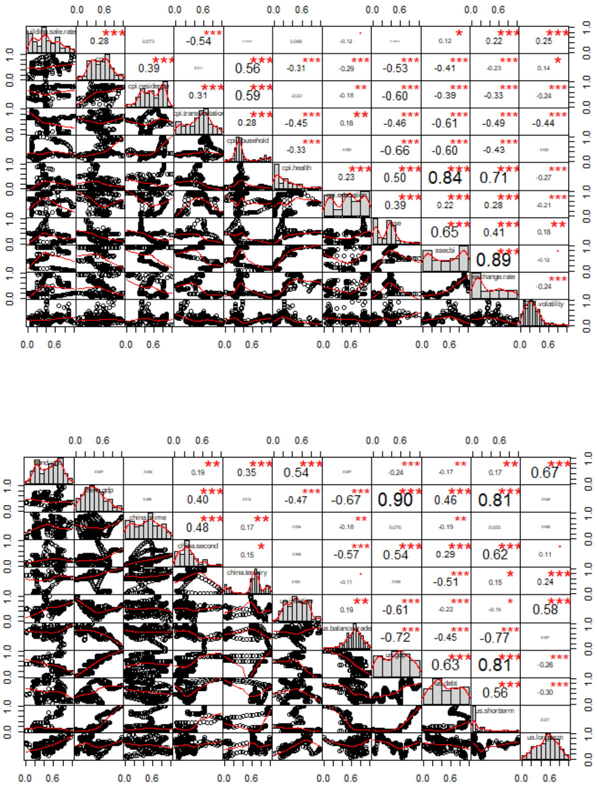


Figure 3. Scatter Plot and Correlations

Principal Component Analysis

Principal Component Analysis (PCA) was run on the model. In simple words, PCA is a method of extracting important variables from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. With fewer variables, visualization also becomes much more meaningful. PCA is more useful when dealing with 3 or higher dimensional data. The results were placed into a scree plot.

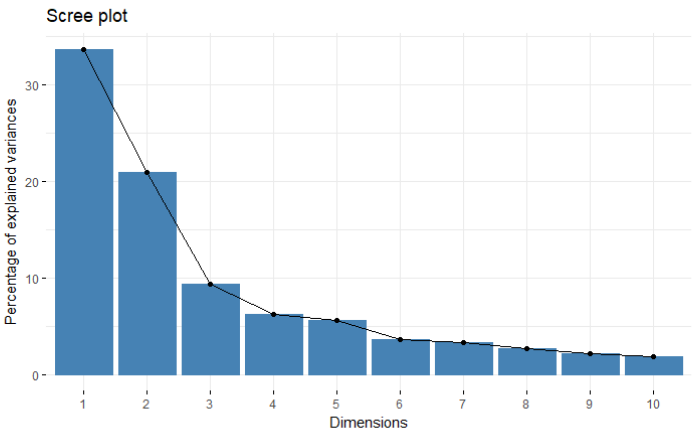


Figure 4. Scree Plot

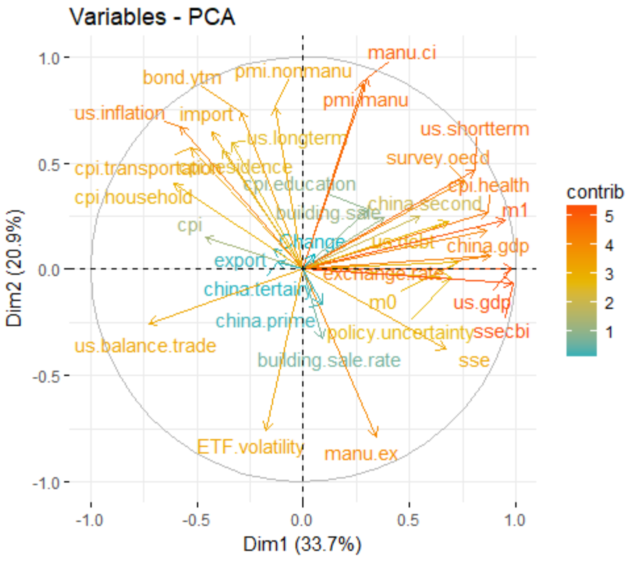


Figure 5. Important Variables chosen by PCA

This variable graph of PCA result shows the features contributions based on the first two principal components, which are called Dim1 and Dim2 here. The red color shows the most important feature and the green color shows the least important feature. It will be a proof of our regression model variables importance.

K-means Clustering

The goal of K-means Clustering algorithm is to find groups in the data based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are we could get labels for the training data, and each data point is assigned to a single cluster.

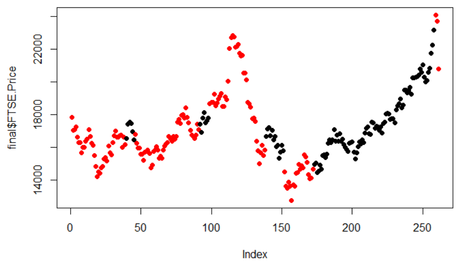


Figure 6. K-means Clustering

From this graph, we see the negative/positive weekly change in tracking index price was not decided by the similarities of features, or we may say, the negative/positive weekly change was not decided by the similarities among those existed features based on K-means Clustering.

Supervised Learning

The FTSE indicator is identified as the dependent variable. We used both classification and regression models to find the features that have the best chance of predicting the price or change in price of the FTSE. We ran two datasets: 1. Normalized dataset 2. Dataset with percent change from week to week.

Regression Models

Multiple Linear Regression Model

A multiple linear regression is an extension of simple linear regression. The MLR is used to determine the relationship between an outcome or response variable and multiple predictor or input variables. In this sense, every value of the predictor variables is associated with a value of the dependent variable (Multiple Linear Regression, 1998). The formula is as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i \text{ for } i = 1,2, \ldots n.$$

Three MLR models were used, a full model which used all predictor variables, a reduced model which used 16 variables, and a backwards or stepwise which used stepwise variable selection and resulted in the use of 19 predictor variables. Each model was run on the training data and test data for the 4/1 split, 80/20 random split, and the 10-fold cross validation concept. Figure 7, below, shows the coefficient plots of the 4/1 split for the full, reduced, and backwards model. Variable names have been omitted for proprietary reasons. Of the three models, the backward model performed best based on RSS value and RMSE.

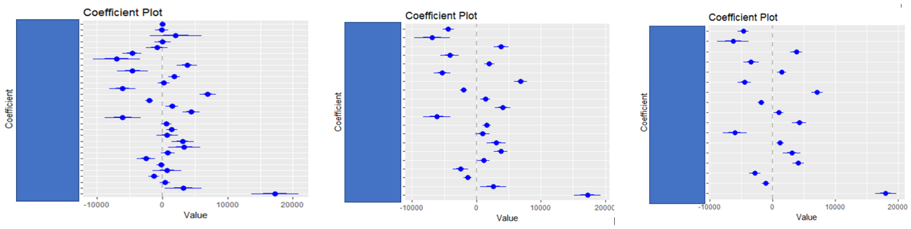


Figure 7. Coefficient Plot

Regression Tree

Regression Trees are part prediction trees machine learning where we predict a response variable (FTSE in our case) from independent variables. When we have a prediction tree with continuous response it is called regression tree. Wei-Yin Loh at Wiley describes Regression Tree as, “The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. Thus, the partitioning can be represented graphically as a decision tree. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values”. (Loh, Eltinge, Cho, & Li, 2017) One of the benefits of using regression trees is that it is very good in predicting the top features which is what we are trying to do. In training data we first find the best feature which predicts the dependent variable and then we find a split on the feature.

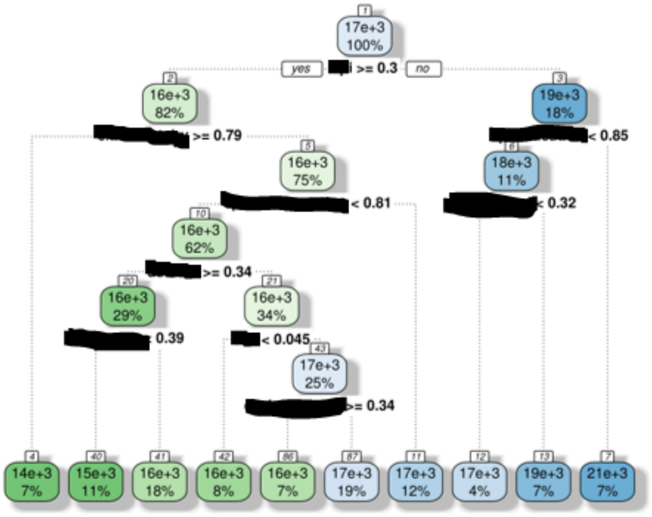


Figure 8. Regression Tree FTSE close price prediction

Figure 8 represents a regression Tree created for this project, top indicators have been removed for proprietary reasons.

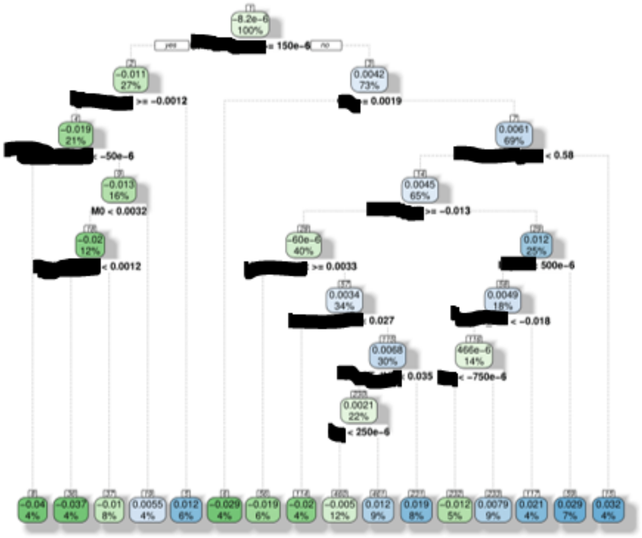


Figure 9 Regression Tree FTSE weekly change prediction

Random Forest

Random forest is considered as an ensemble learning technique that are usually used for regression and classification techniques where various mean prediction regressions of the individual trees are determined.

In this model predictions were made by combining decisions from a sequence of base models. More formally we can write this class of models as the following formula:

$$g(x)=f0(x)+f1(x)+f2(x)+...$$

Where the final model g is the sum of simple base models fi. Here, each base classifier is known as decision tree. this technique of arranging various models is known as model assembling.

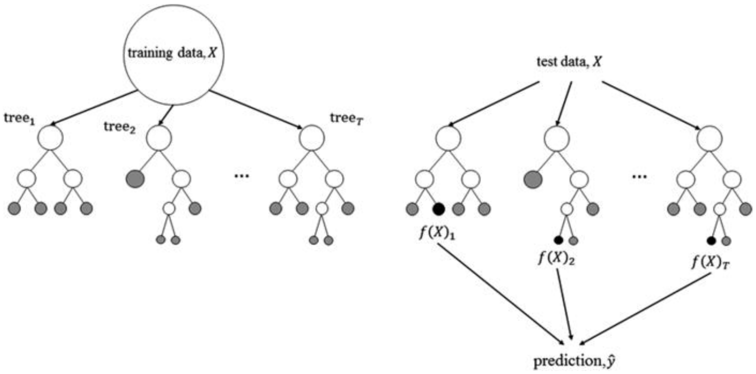


Figure 10. Steps for Random Forest

After running the random regression and testing the error rate for all the three models, we plotted variable importance. Figure 11 (below) shows the error rate of the data for the following three models.

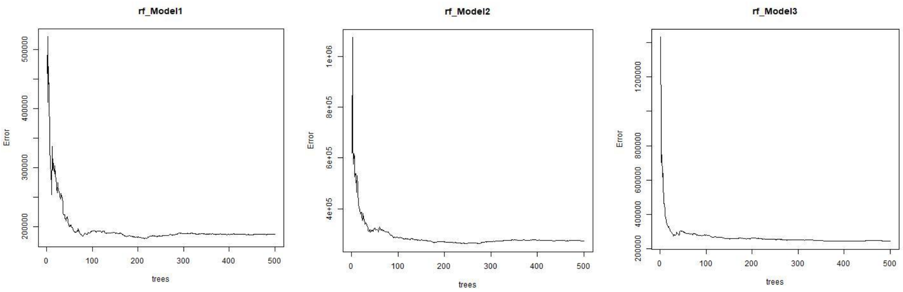
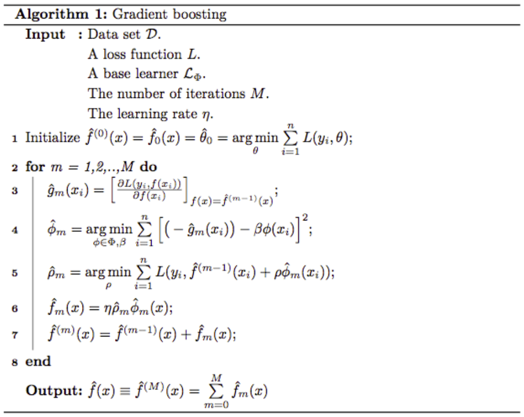


Figure 11. Error Rate

Gradient Boosting Regression

Gradient boosting (Barranco, May 2017) is a machine learning tool for boosting or improving model performance.



```
Algorithm 1: Gradient boosting
  Input  : Data set $\mathcal{D}$.
           A loss function $L$.
           A base learner $\mathcal{L}_{\Phi}$.
           The number of iterations $M$.
           The learning rate $\eta$.
  Initialize $\hat{f}^{(0)}(x) = \hat{f}_0(x) = \hat{\theta}_0 = \arg\min_{\theta} \sum_{i=1}^{n} L(y_i, \theta)$;
  for $m$ = 1,2,..,$M$ do
      $\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}^{(m-1)}(x)}$;
      $\hat{\phi}_m = \arg\min_{\phi \in \Phi, \beta} \sum_{i=1}^{n} \left[\left(-\hat{g}_m(x_i)\right) - \beta\phi(x_i)\right]^2$;
      $\hat{\rho}_m = \arg\min_{\rho} \sum_{i=1}^{n} L(y_i, \hat{f}^{(m-1)}(x_i) + \rho\hat{\phi}_m(x_i))$;
      $\hat{f}_m(x) = \eta\hat{\rho}_m\hat{\phi}_m(x)$;
      $\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{f}_m(x)$;
  end
  Output: $\hat{f}(x) \equiv \hat{f}^{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x)$
```

Figure 12. Gradient Boosting Algorithm

First, you develop an initial model called the base learner using whatever algorithm of your choice (linear, tree, etc.). Next, gradient boosting looks at the error and develops a model using what is called loss function. The loss function is the difference between the current accuracy and the desired prediction whether it's accuracy for classification or error in regression. This process of making additional models based only on the misclassified ones continues until the level of accuracy is reached. Gradient boosting is also stochastic. This means that it randomly draws from the sample as it iterates over the data. This helps to improve accuracy and or reduce error.

Gradient boosting uses weak decision trees, that are increasingly focused on hard examples, which is a high-performing model, but with some limits. A small change in the feature set or training set can create radical changes in the model, and not easy to understand predictions.

```
control <- trainControl(method = "repeatedcv",number=10,repeats=10)
gbm.train1 <- train(ftse~., data=training1, method='gbm',trControl=control)
```

```
Stochastic Gradient Boosting

208 samples
 29 predictor
```

```
No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 187, 186, 188, 188, 187, 188, ...
Resampling results across tuning parameters:

 interaction.depth n.trees RMSE  Rsquared MAE
 1      50  767.2777 0.8334973 604.1640
 1      100  629.3569 0.8802369 500.5266
 1      150  575.0051 0.8985565 454.8025
 2      50  611.2360 0.8880863 478.7978
 2      100  510.9346 0.9169155 401.4268
 2      150  480.2912 0.9252673 378.5971
 3      50  561.2744 0.9049783 437.0193
 3      100  477.6859 0.9263526 374.8565
 3      150  457.6645 0.9313707 359.9159

Tuning parameter 'shrinkage' was held constant at a value of 0.1

Tuning parameter 'n.minobsinnode' was held constant at a value of 10
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were n.trees = 150, interaction.depth
 = 3, shrinkage = 0.1 and n.minobsinnode = 10.
```

The printout shows the values for each potential model. At the bottom of the printout are the recommended parameters for our model. We take the values at the bottom to create our model for the test data. Based on the grid search (Synced, 2018), we choose interaction. Depth=3, number of trees = 150, and learning rate = 0.1.

Ridge Regression

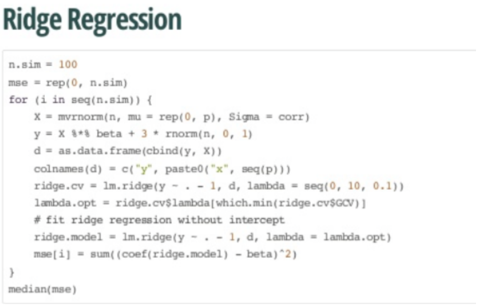
Ridge Regression is a technique used for creating various models and to predict the number of variables in a dataset when the dataset has many relations between the predictor variables.

Specifically, the ridge regression estimate β is defined as the value of β that minimizes

$$\sum_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Ridge regression calculates a ‘shrinkage’ term. This term is usually controlled by another term known as lambda (which has to be calculated separately).

Two interesting implications of this term are the facts that when λ = 0, then the ridge regression is equivalent to least square regression and if the value of λ= infinity, then all the coefficients are shrunk to the value of zero. Hence the values usually lie between the values of zero to infinity.



Ridge Regression

```
n.sim = 100
mse = rep(0, n.sim)
for (i in seq(n.sim)) {
    X = mvrnorm(n, mu = rep(0, p), Sigma = corr)
    y = X %*% beta + 3 * rnorm(n, 0, 1)
    d = as.data.frame(cbind(y, X))
    colnames(d) = c("y", paste0("x", seq(p)))
    ridge.cv = lm.ridge(y ~ . - 1, d, lambda = seq(0, 10, 0.1))
    lambda.opt = ridge.cv$lambda[which.min(ridge.cv$GCV)]
    # fit ridge regression without intercept
    ridge.model = lm.ridge(y ~ . - 1, d, lambda = lambda.opt)
    mse[i] = sum((coef(ridge.model) - beta)^2)
}
median(mse)
```

Figure 13. Ridge Regression R Code Algorithm

We have initially split the data set into two fractions, then use one portion to fit β and the other to evaluate how well Xβ predicted the observations in the second portion.

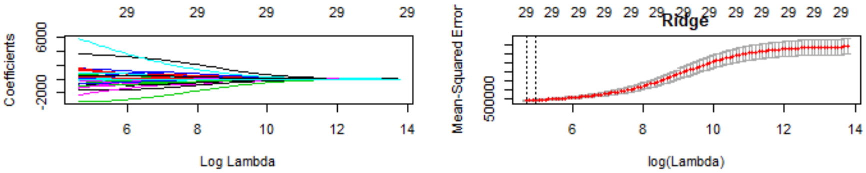


Figure 14. Ridge Regression Results

LASSO Regression

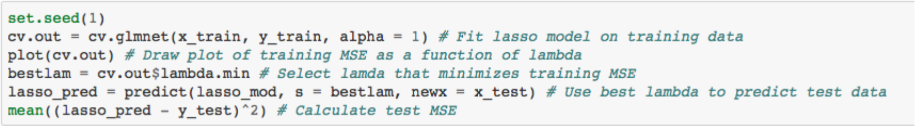
Least absolute shrinkage and selection operator (LASSO) is a type of regression analysis method that helps in performing the variable selection and the regularization technique simultaneously.

Lasso regression is usually used for predicting the best accuracy of the variables while training and testing a dataset.

The below formula can be used for calculating the lasso regression model:

$$\begin{aligned}\hat{\beta}^{\text{lasso}} &= \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda\sum_{j=1}^{p}|\beta_j| \\ &= \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}\end{aligned}$$

Lasso regression is different from ridge regression. It can usually predict the variable selection in the linear model more accurately, for example as λ increases, more coefficients are set to zero (less variables are selected), and among the nonzero coefficients, more shrinkage takes place.



```
set.seed(1)
cv.out = cv.glmnet(x_train, y_train, alpha = 1) # Fit lasso model on training data
plot(cv.out) # Draw plot of training MSE as a function of lambda
bestlam = cv.out$lambda.min # Select lamda that minimizes training MSE
lasso_pred = predict(lasso_mod, s = bestlam, newx = x_test) # Use best lambda to predict test data
mean((lasso_pred - y_test)^2) # Calculate test MSE
```

Figure 15. LASSO Regression R Code

Results: Normalized Data

Multiple Linear Model	***Regression Tree***

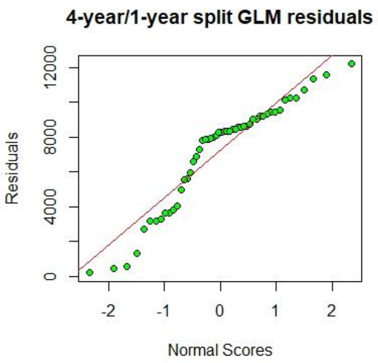
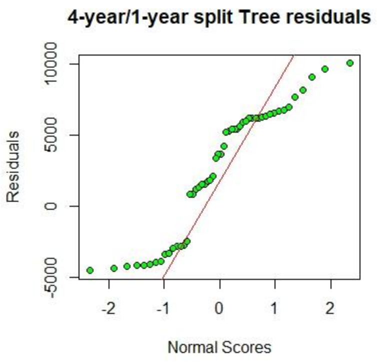
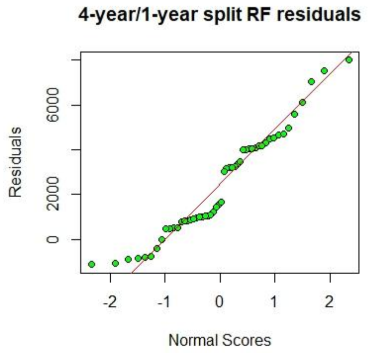
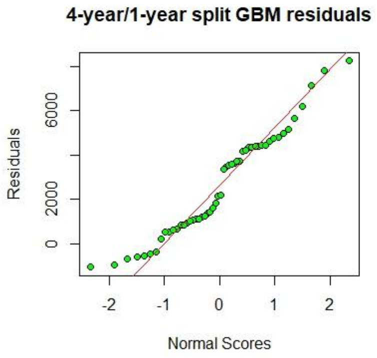
	
Random Forest	**Gradient Boosting**
	

Figure 16. Residuals Normal Q-Q Plot (normalized data; 4-year and 1-year split)

Based on 4-year and 1-year data split, we got the Residuals Normal Q-Q plot as above. From the figure we could find, all of the training model residuals were not located in a straight normal distribution line. However, the latter two for Random Forest and Gradient Boosting look better than the first two.

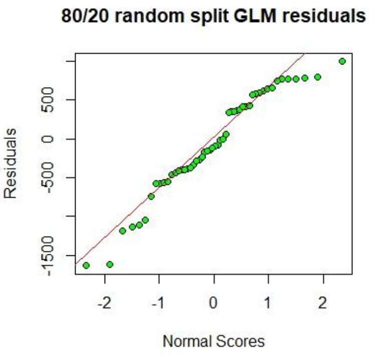
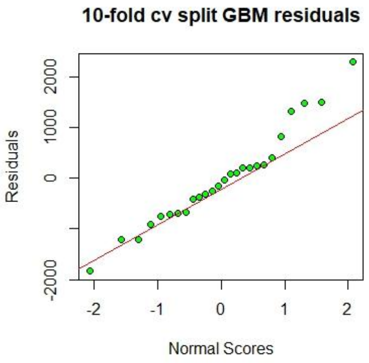
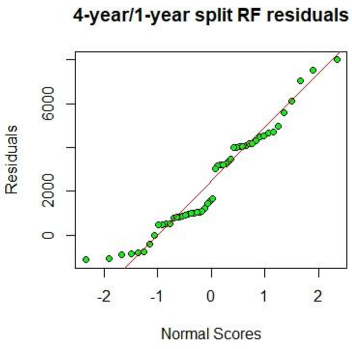
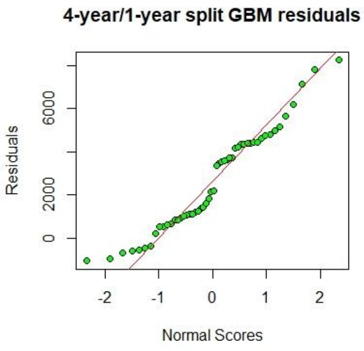
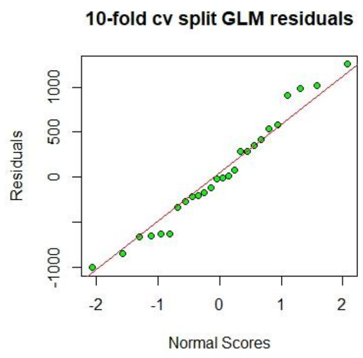
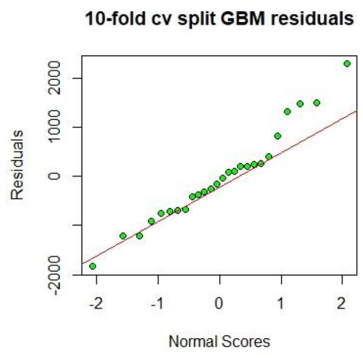
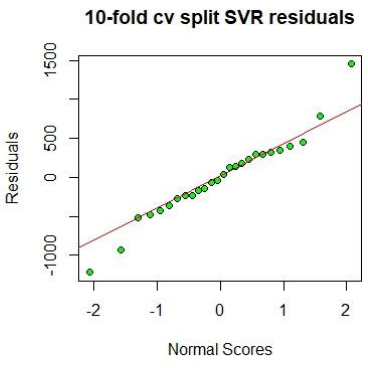
Multiple Linear Model	***Regression Tree***
 80/20 random split GLM residuals	 10-fold cv split GBM residuals
Random Forest	**Gradient Boosting**
 4-year/1-year split RF residuals	 4-year/1-year split GBM residuals

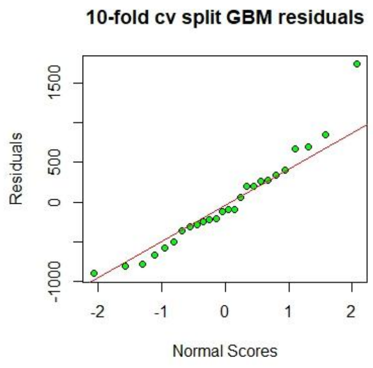
Figure 17. Residuals Normal Q-Q Plot (normalized data; 80% and 20% random split)

Based on 80% and 20% random split, we got the Residuals Normal Q-Q plot as above. From the figure we could find, all of the training model residuals were not located in a straight normal distribution line. However, the latter two for Random Forest and Gradient Boosting look better than the first two.

<table>
<tr><th><em><strong>Multiple Linear Model</strong></em></th><th><em><strong>Regression Tree</strong></em></th></tr>
<tr><td>


</td><td>


</td></tr>
<tr><td><em><strong>Random Forest</strong></em></td><td><strong>Gradient Boosting</strong></td></tr>
<tr><td>


</td><td>


</td></tr>
</table>

Figure 18. Residuals Normal Q-Q Plot (normalized data; 10-fold cross validation split)

Based on 10-fold cross validation split, we got the Residuals Normal Q-Q plot as above. From the figure we could find all of the training model residuals were located in a straight normal distribution line with a few outliers. This is inconsistent with the results of the 4 year/1 year and 80/20 split.

Table 5. Model Training Result (normalized data)

	4-year and 1-year split	*80% and 20% random split*	*10-fold cross validation split*
Multiple linear model	7755.86	645.44	593.76
Regression Tree	5147.55	1061.24	840.36
Random Forest	3344.22	482.97	518.80
Gradient Boosting	3603.56	528.18	661.70

Based on comparison of RMSE, Random Forest model is the best among four regression training models.

Table 6. Model Training Variable Importance (normalized data)

Multiple Linear	***Regression Tree***	***Random Forest***	***Gradient Boosting***
VAR1	VAR4	VAR4	VAR4
VAR2	VAR1	VAR1	VAR1
VAR3	VAR4b	VAR6	VAR9
VAR4a	VAR2	VAR7	VAR4c
VAR5	VAR6	VAR8	VAR6

Table 6 (above) shows that the variables with strongest predictive power are the same among all four training models.

Table 7 Best Training Model - Random Forest Robustness Test

Normalized data	***Normalized data +.0001***	***Normalized data +.0005***	***Normalized data -.0001***	***Normalized data -.0005***	***Normalized data w/ random noise***
VAR4	VAR4	VAR4	VAR4	VAR4	VAR4
VAR1	VAR1	VAR6	VAR1	VAR1	VAR7
VAR6	VAR6	VAR1	VAR7	VAR6	VAR6
VAR7	VAR5	VAR4b	VAR6	Var5	VAR1
VAR8	VAR4b	VAR8	VAR8	VAR8	VAR8

Table 7 (above) shows how the important variables for the random forest change as we add slight randomness to the training data.

Results: Percentage Change Dataset

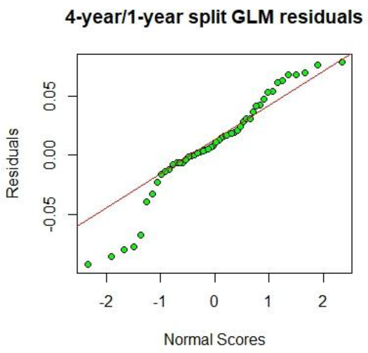
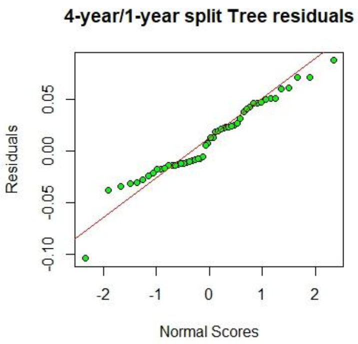
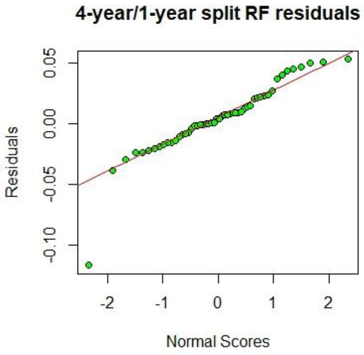
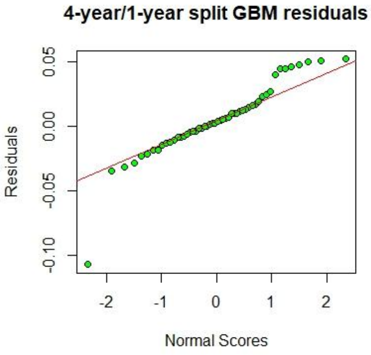
Multiple Linear Model	***Regression Tree***
	
Random Forest	**Gradient Boosting**
	

Figure 19. Residuals QQ Plot for Regression Model Training Results (percentage change data)

Figure 21 (above) shows Residual Normal Q-Q plots for the 4year and 1-year split. The first two training model residuals were not located in a straight normal distribution line. However, the latter two for Random Forest and Gradient Boosting look better than the first two, although there is one outlier. Note, the x-axis plots the theoretical quantiles, which are the quantiles from the standard Normal distribution with mean 0 and standard deviation 1.

Table 8. Model Training Result (percentage change data)

Multiple Linear	***Regression Tree***	***Random Forest***	***Gradient Boosting***
0.0474	0.0366	0.0279	0.0274

The RMSE of Random Forest and Gradient Boosting were quite similar as we would expect, given the Q-Q plots. We chose Gradient Boosting as the best model since it, slightly, outperformed Random Forest.

Table 9. Model Training Variable Importance (percentage change data)

Multiple Linear	***Regression Tree***	***Random Forest***	***Gradient Boosting***
VAR1	VAR3b	VAR2b	VAR6
VAR2a	VAR4	VAR6	VAR2b
VAR2b	VAR2	VAR3b	VAR1
VAR3	VAR5	VAR2c	VAR8
VAR2c	VAR2b	VAR7	VAR3a

Table 9 (above) shows that the variables with the strongest predictive power are the similar among all four training models.

Table 10. Best Training Model - Gradient Boosting Robustness Test

Percentage change data	***Percentage change data +.5%***	***Percentage change data - .5%***	***Percentage change data w/ random noise***
VAR6	VAR6	VAR6	VAR1
VAR2b	VAR2b	VAR8	VAR10
VAR1	VAR1	VAR2b	VAR8
VAR8	VAR8	VAR1	VAR11
VAR3a	VAR3a	VAR9	VAR6

Table 10 (above) shows the results of the robustness check for our best performing model, Gradient Boosting. Not all variables that were present in the Gradient Boosting model prior to adding noise, are present in the model with noise added. As well, some new variables have been identified when noise was added.

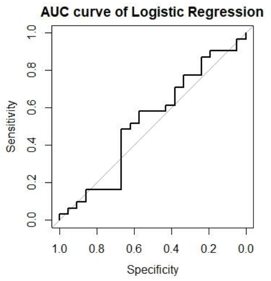
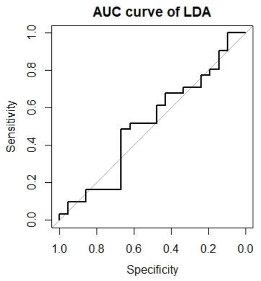
Classification Models

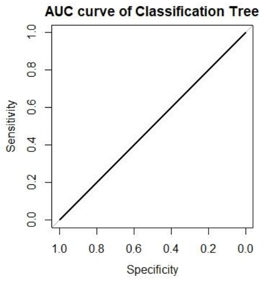
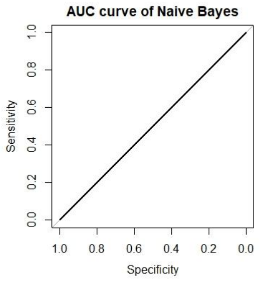
Classification aims to group data per its similarities, which is also a kind of prediction method that assigns each data point to a predefined category. A classification model attempts to categorize some conclusion from observed values. There are several classification models in this project: Logistic Regression, Linear Discriminant Analysis (LDA), Classification Tree and Naive Bayes were used to train the dataset.

Logistic regression is a variation of linear regression model and is a simple classification algorithm which learns to make binary decisions. LDA is a pattern recognition algorithm to find a linear combination of features and then separate the data into two or more classes. Classification Tree is a type of machine learning algorithm used for classifying loose data in by binary classifier.

Results: Normalized Data

Table 11. AUC results for Classification models (normalized data)

	Logistic Regression	***LDA***
ROC Curve		
AUC	0.5238	0.5131
	Classification Tree	***Naive Bayes***

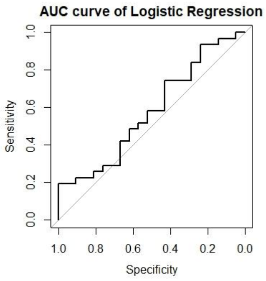
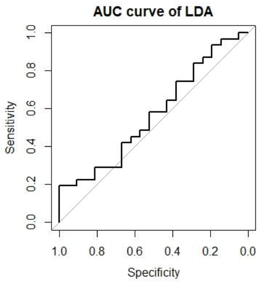
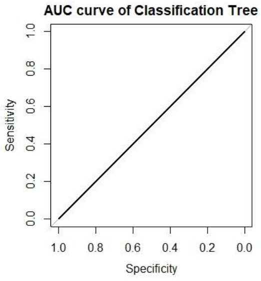
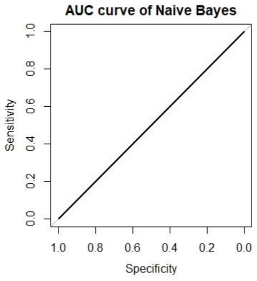
ROC ***Curve***		
AUC	0.5	0.5

Area under the curve (AUC) was used in order to determine which of the training models predicts the classes best. The Receiver Operating Characteristic(ROC) Curve, plots the true positive rates are plotted against false positive rates. The closer the AUC is to 1, the better the results will be. The closer ROC Curve move to the top left, the better the results will be. Our results of four classification models, based on dataset 1, are just slightly above 0.5, which is not a good training model result.

Results: Percentage Change Dataset

Table 12. AUC results for Classification models (percentage change data)

	Logistic Regression	***LDA***

ROC Curve		
AUC	0.5776	0.56
	Classification Tree	***Naive Bayes***
ROC Curve		
AUC	0.5	0.5

Similarly, our results of the four classification models, based on dataset 2, are just slightly above 0.5, which is not a good training model result.

Summary

In this project, we sought to find significant variables, both on the micro and macro level, that could be used to predict the price of a Chinese ETF. We ran regression and classification supervised learning models to find significant predictor variables given a known dependent variable. To test model robustness, we added noise to our model, tested our model with a random data split, and used ten-fold cross validation. As well, we began creating unsupervised learning techniques to uncover patterns within our microeconomic indicators.

Model Robustness

We first sought to test model robustness by splitting the data into a random set and by using cross fold validation. However, we found that since future data would be included in the robustness checks, it was difficult to determine how the model was affected by these changes in the training and testing data. For this reason, we chose to add noise to our best performing model to test model robustness.

Classification Models

The classification model we ran did not provide results that were significant enough to use to predict movement in the price of the FTSE. We hypothesize the following reasons why classification model fails,

There is no obvious similarities among attributes to do. When there often are similar inputs that are associated to different outputs (Dorard, 2013). As we could see from the unsupervised learning part, the k-means cluster may not work well on the dataset, which means it's hard for the algorithm to find similarities based on those attributes. When that happens in a classification problem, it becomes difficult to clearly separate classes.

The training dataset is significantly different from testing dataset. Our dataset was chosen from the 5-year period from Feb 2013 to Feb 2018. Our training dataset was from Feb 2013 to Feb 2017, testing dataset was from Feb 2017 to Feb 2018, with one-week buffer between them. In this case, we ended up creating a model on the training dataset which had significantly different data than the testing dataset. The Chinese Economy could have rapid changes in one year and so this is not ideal.

High dimensionality for implementation. The unsupervised results show that, even if we generated PCA for dimension reduction, there would still be more than twenty dimensions in the final dataset. In business cases, we sometimes start with bi-variable or tri-variable, even though these variables make no sense. More than twenty variables are way too many for classification model to see the actual light of the real-world (Srivastava, 2016).

Regression Models: Normalized Dataset

For the regression models, we chose the Random Forest model based on how well the residuals plotted to the prediction line on the QQ plots and we chose Random Forest based on the RMSE value. However, the significant variables selected by the Random Forest model included many of the significant variables captured by the other regression models. The most significant variables remained relevant even after adding noise to the model.

Regression Models: Percentage Change Dataset

As financial institutions often seek to find trends in data based on a percentage change from one time period to the next. We developed a ‘percent change’ dataset.   It will allow the product owner the ability to identify factors that are changing in the same way as the predictor variable. The Gradient Boosting algorithm performed best on the percent change dataset based on the QQ-plot and the RMSE. However, all models produced many of the same leading indicators.

Future Work

From our results, we can see that the top indicators are Chinese indicators with one US indicator identified. It is obvious that the Chinese indicators would have the most effect on the market and they would outperform the US indicators. A possible extension of this research can be to separate the Chinese indicators from US indicators and perform analytics. By doing this we will have top Chinese indicators and top US indicators separately and use these to create our model. One option for creating model would be to use four Chinese indicators and two US indicators and see what results we get.

Another suggestion for future work would be to research the US Manufacturing PMI variable and add it to the list of indicators. PMI tells us how powerful the manufacturing sector is for US. It is also a good indicator for learning about the imports and exports between two countries . Once we have a good grip on the US indicators, the next step would be to add indicators from different countries like Japan, Korea, and Russia to our model.

A clear, immediate next step in our project would be to create a signal which would indicate where to allocate available cash or assets. One technique we recommend is using 10, 20 and 200 day moving average price and compare it with the daily price and if the daily price is more then we buy or else we sell. Moving averages are used by traders to help reduce the amount of “noise” and get an idea if the price is going to go up or down. We could do the ground work and created R code which would predict the signal (0 or 1) using the 10, 20 and 200-day simple moving average. Weekly change will be positive if the weekly price is greater than all SMAs, and weekly change will be negative when weekly price is less than or equal to all the SMAs. In the below figure we plotted simple moving average along with the FTSE close price and it is evident that the SMA resonates the closing price closely.

Deleted: and we would buy

Deleted: or sell/hold

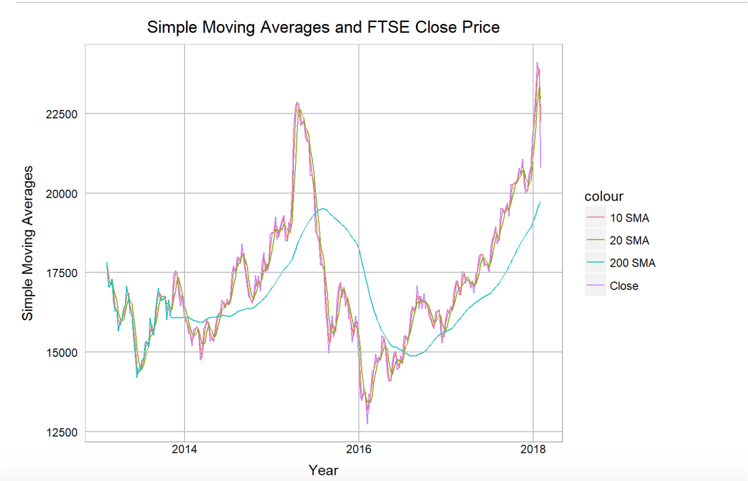


Figure 20. Simple Moving Average Lines

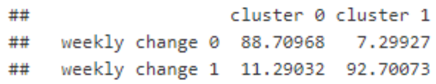
Here are the signals we were able to predict –

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

We can see in our probability table that we are 63% confident about signal being 0 for weekly change 0 and 100% sure about signal been 1 for weekly change 1.

Another alternative to SMA we suggest for future groups to consider is to use week ‘t’ to predict response variable for week ‘t+1’. This would need more research done in a future project.

One last method we researched and tried to implement to predict the signals is using candlesticks. We were able to use candlesticks to predict the signal similar to SMA method. K-means clustering is used to group candlesticks into groups and these groups are used to create a signal of when to trade. When a body type changes, the signal to trade will be set to 0 or 1. There is a probability matrix that can be created to predict the likelihood of a price change based on a body change.



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

A final recommendation for future work is sentiment analysis. Sentiment is the overall attitude of investors towards a particular ETF or the trading market. We could gather and track investors attitude from social media, like twitter. There is much to be discovered in this area.

Appendix A - Lexicon

General Terminologies

Word	***Abbr***	***Definition***
Fund Flow		Fund flow is the net of all cash inflows and outflows in and out of various financial assets. Fund flow is usually measured on a monthly or quarterly basis; the performance of an asset or fund is not taken into account, only share redemptions, or outflows, and share purchases, or inflows. Net inflows create excess cash for managers to invest, which theoretically creates demand for securities such as stocks and bonds.
Economic Surprise Index	ESI	The Citigroup Economic Surprise Indices are objective and quantitative measures of economic news. They are defined as weighted historical standard deviations of data surprises (actual releases vs Bloomberg survey median)
Chinese A Shares		China's "A Share" market refers to stocks that trade on the Shanghai and Shenzhen exchanges. these stocks are strictly off limits to non-Chinese investors
Chinese B Shares		Here's where it starts to get confusing. Some Chinese companies are listed in Shanghai and Shenzhen, but their shares trade in U.S. dollars. These stocks, known as "B shares," were historically designed to give Chinese companies a way to raise capital from overseas.
Hong Kong H Shares		"H Shares" are also Chinese companies, but these securities trade on the Hong Kong Stock Exchange, rather than on the mainland, and they are priced in Hong Kong dollars
Chinese Stocks in New York		These are companies that are headquartered in mainland China but have chosen to list their shares on the New York Stock Exchange or Nasdaq. There are currently over 100 such Chinese companies listed in the U.S
Shanghai-Hong Kong Stock Connect		connects the Shanghai Stock Exchange and the Hong Kong Stock Exchange. The reasoning behind the connection was to open up the Chinese markets to additional investors by way of Hong Kong.
Earnings Revision Index	ERI	Earnings estimates play a huge role in determining equity prices, but analyst revisions to bottom-line forecasts often have significantly more influence over future stock performance
Volatility Index	VIX	The Volatility Index, known by its ticker symbol VIX, is a popular measure of the stock market's expectation of volatility implied by S&P 500 index options, calculated and published by the Chicago Board Options Exchange (CBOE). It is colloquially referred to as the fear index or the fear gauge.
Assets Under Management	AUM	In finance, assets under management (AUM), sometimes called funds under management (FUM), measures the total market value of all the financial assets which a financial institution such as a mutual fund, venture capital firm, or brokerage house manages on behalf of its clients and themselves.
Exchange-Traded Fund	ETF	An exchange-traded fund (ETF) is an investment fund traded on stock exchanges, much like stocks. An ETF holds assets such as stocks, commodities, or bonds and generally operates with an arbitrage mechanism designed to keep it trading close to its net asset value, although deviations can occasionally occur. Most ETFs track an index, such as a stock index or bond index. ETFs may be attractive as investments because of their low costs, tax efficiency, and stock-like features.
Bull		A bull market is a period of generally rising prices. The start of a bull market is marked by widespread pessimism. This point is when the "crowd" is the most "bearish". The feeling of

		despondency changes to hope, "optimism", and eventually euphoria, as the bull runs its course. This often leads the economic cycle, for example in a full recession, or earlier.
Bear		A bear market is a general decline in the stock market over a period of time. It is a transition from high investor optimism to widespread investor fear and pessimism. According to The Vanguard Group, "While there's no agreed-upon definition of a bear market, one generally accepted measure is a price decline of 20% or more over at least a two-month period."
Long		In finance, a long position in a financial instrument, means the holder of the position owns a positive amount of the instrument.
Short		In finance, short selling (also known as shorting or going short) is the practice of selling securities or other financial instruments that are not currently owned (usually borrowed), and subsequently repurchasing them ("covering"). In the event of an interim price decline, the short seller profits, since the cost of (re)purchase is less than the proceeds received upon the initial (short) sale.
ETF tracking index		ETFs can track an index. Those indices are designed to track a particular market "index", whether it is the S&P 500, Russell 2000, or MSCI EAFE; also called "index fund".
Liquidity		Liquidity describes the degree to which an asset or security can be quickly bought or sold in the market without affecting the asset's price. Market liquidity refers to the extent to which a market, such as a country's stock market or a city's real estate market, allows assets to be bought and sold at stable prices. Cash is considered the most liquid asset, while real estate, fine art and collectibles are all relatively illiquid.
Risk Aversion		A risk-averse investor dislikes risk, and therefore stays away from high-risk stocks or investments and is prepared to forego higher rates of return.
Risk Seeking		Risk seeking is the search for greater volatility and uncertainty in investments in exchange for anticipated higher returns. Risk seekers might pursue investments such as small-cap stocks and international stocks, preferring growth investments over value investments.
Listed Company		A listed company, publicly traded company, publicly held company, public company, or public corporation is a corporation whose ownership is dispersed among the general public in many shares of stock which are freely traded on a stock exchange or in over the counter markets.

World Indices

Symbol	***Name***	***Description***	***Country***	***Description Source***
MERV	MERVAL BUENOS AIRES	The Argentina Merval Index, a basket weighted index, is the market value of a stock portfolio, selected according to participation in the Buenos Aires Stock Exchange, number of transactions of the past 6 months and trading value.	Argentina	https://www.bloomberg.com/quote/MERVAL:IND
AXJO	S&P/ASX 200	The S&P/ASX 200 index is a market-capitalization weighted and float-adjusted stock market index of stocks listed on the Australian Securities Exchange	Australia	https://en.wikipedia.org/wiki/S%26P/ASX_200

AORD	ALL ORDINARIES	All Ordinaries is the oldest index of shares in Australia. It is made up of the share prices for 500 of the largest companies listed on the Australian Securities Exchange(ASX)	Australia	https://en.wikipedia.org/wiki/All_Ordinaries
BVSP	IBOVESPA	It is a gross total return index weighted by free float market cap & is comprised of the most liquid stocks traded on the Sao Paulo Stock Exchange. It has been divided 10 times by a factor of 10	Brazil	https://www.bloomberg.com/quote/IBOV:IND
GSPTSE	S&P/TSX Composite index	The S&P/TSX Composite Index is the benchmark Canadian index, representing roughly 70% of the total market capitalization on the Toronto Stock Exchange (TSX) with about 250 companies included in it	Canada	https://en.wikipedia.org/wiki/S%26P/TSX_Composite_Index
BATSK	BATS 1000 Index	BATS 1000 Index measures 1000 of the largest stocks in 10 sectors. The index combines aspects of equal and market capitalization weightings.	United States	https://www.forbes.com/2009/09/28/bats-exchange-financials-intelligent-investing-index/
IPSA	IPSA SANTIAGO DE CHILE	The IPSA Index is a Total Return Index and is composed of the 40 stocks with the highest average annual trading volume in the Santiago Stock Exchange (Bolsa de Comercio de Santiago)	Chile	https://www.bloomberg.com/quote/IPSA:IND
FTSE	FTSE 100	The Financial Times Stock Exchange 100 Index is a share index of the 100 companieslisted on the London Stock Exchange with the highest market capitalisation.	United Kingdom	https://en.wikipedia.org/wiki/FTSE_100_Index
KLSE	FTSE Bursa Malaysia KLCI	The FTSE Bursa Malaysia KLCI Index comprises of the largest 30 companies by full market capitalization on Bursa Malaysia's Main Board.	Malaysia	https://www.bloomberg.com/quote/FBMKLCI:IND
TA100	TA-125	The TA-125 Index is TASE's most significant index and considered as the Israel Economy BenchMark Index. TA-125 financial products are the most popular among TASE Indices.	Israel	https://www.tase.co.il/Eng/MarketData/Indices/MarketCap/Pages/IndexMainDataMarket.aspx?IndexId=137
CASE30	EGX 30 INDEX	The EGX 30 Index is a free-float capitalization weighted index of the 30 most highly capitalized and liquid stocks traded on the Egyptian Exchange. EGX 30 constituents are reviewed and changed twice a year (end of January and end of July).	Egypt	https://www.bloomberg.com/quote/EGX30:IND
JN0U.FGI	FTSE/JSE TOP 40 USD	The FTSE/JSE Top40 Index is a capitalization weighted index. Companies included in this index are the 40 largest companies by market capitalization included in the FTSE/JSE All Shares Index.	Africa	https://www.bloomberg.com/quote/TOP40:IND

STOXX50E	ESTX50 EUR P	The EURO STOXX 50 is a stock index of Eurozone stocks designed by STOXX, an index provider owned by Deutsche Börse Group.	European Country	https://en.wikipedia.org/wiki/Euro_Stoxx_50
N100	EURONEXT 100	The Euronext 100 Index is the blue chip index of the pan-European exchange, Euronext NV. It comprises the largest and most liquid stocks traded on Euronext.	European Country	https://en.wikipedia.org/wiki/Euronext_100
BFX	BEL 20	The BEL 20 is the benchmark stock market index of Euronext Brussels. In general, the index consists of a minimum of 10 and a maximum of 20 companies traded at the Brussels Stock Exchange.	European Country	https://en.wikipedia.org/wiki/BEL_20
FCHI	CAC 40	The CAC 40 is a benchmark French stock market index. The index represents a capitalization-weighted measure of the 40 most significant values among the 100 highest market caps on the Euronext Paris.	France	https://en.wikipedia.org/wiki/CAC_40
GDAXI	DAX	The Deutscher Aktien index (German stock index) is a blue chip stock market indexconsisting of the 30 major German companies trading on the Frankfurt Stock Exchange. Prices are taken from the Xetra trading venue	Germany	https://en.wikipedia.org/wiki/DAX
BSESN	S&P BSE SENSEX	The S&P BSE SENSEX is a free-float market-weighted stock market index of 30 well-established and financially sound companies listed on Bombay Stock Exchange	India	https://en.wikipedia.org/wiki/BSE_SENSEX
JKSE	Jakarta Composite Index	The Jakarta Stock Price Index is a modified capitalization-weighted index of all stocks listed on the regular board of the Indonesia Stock Exchange	Indonesia	https://www.bloomberg.com/quote/JCI:IND
N225	Nikkei 225	The Nikkei 225 is a stock market index for the Tokyo Stock Exchange (TSE).	Japan	https://en.wikipedia.org/wiki/Nikkei_225
KS11	KOSPI Composite Index	The KOSPI Index is a capitalization-weighted index of all common shares on the Korean Stock Exchanges. The Index was developed with a base value of 100	Korea	https://www.bloomberg.com/quote/KOSPI:IND
MXX	IPC MEXICO	The Índice de Precios y Cotizaciones (IPC) is an index of 35 stocks that trade on the Bolsa Mexicana de Valores	Mexico	https://en.wikipedia.org/wiki/Indice_de_Precios_y_Cotizaciones
NZ50	S&P/NZX 50 INDEX GROSS	The index is designed to measure the performance of the 50 largest, eligible stocks listed on the Main Board (NZSX) of the NZX by float-adjusted market capitalization.	New Zealand	http://us.spindices.com/indices/equity/sp-nzx-50-index

MICEXINDEXCF.ME	MICEX IND	The MOEX Russia Index (formerly MICEX Index) is the main Ruble-denominated benchmark of the Russian stock market.	Russia	https://en.wikipedia.org/wiki/MOEX_Russia_Index
STI	STI Index	The Straits Times Index comprises of the stocks of 30 representative companies listed on the Singapore Exchange. The index is calculated based on market-value weighted stock market index.	Singapore	http://www.moneycontrol.com/live-index/straitstimes
TWII	TSEC weighted index	The TWSE, or TAIEX, Index is capitalization-weighted index of all listed common shares traded on the Taiwan Stock Exchange. The index has a base value of 100	Taiwan	https://www.bloomberg.com/quote/TWSE:IND
MSCITW	MSCI Taiwan	The MSCI Taiwan Index is designed to measure the performance of the large and mid cap segments of the Taiwan market.	Taiwan	https://www.msci.com/documents/10199/6f36d84d-425d-4e1f-8d56-e65c455ebda1
GSPC	S&P 500	The index includes 500 leading companies and captures approximately 80% coverage of available market capitalization.	United States	http://us.spindices.com/indices/equity/sp-500
DJI	Dow 30	The Dow Jones Industrial Average is a price-weighted measure of 30 U.S. blue-chip companies. The index covers all industries except transportation and utilities.	United States	http://us.spindices.com/indices/equity/dow-jones-industrial-average
IXIC	Nasdaq	The NASDAQ Composite is a stock market index of the common stocks and similar securities listed on the NASDAQ stock market.	United States	https://en.wikipedia.org/wiki/Nasdaq_Composite
NYA	NYSE COMPOSITE (DJ)	The NYSE Composite is a stock market index covering all common stock listed on the New York Stock Exchange, including American depositary receipts, real estate investment trusts, tracking stocks, and foreign listings.	United States	https://en.wikipedia.org/wiki/NYSE_Composite
XAX	NYSE AMEX COMPOSITE INDEX	An index made up of stocks that represent the NYSE Amex equities market. The NYSE Amex Composite Index is a market capitalization-weighted index, so the weight of each stock depends on the price of the shares and how many are outstanding.	United States	https://www.investopedia.com/terms/n/nyse-amex-composite-index.asp
RUT	Russell 2000	The Russell 2000 Index is a small-cap stock market index of the bottom 2,000 stocks in the Russell 3000 Index.	United States	https://en.wikipedia.org/wiki/Russell_2000_Index
VIX	Vix	The CBOE Volatility Index is a popular measure of the stock market's expectation of volatility implied by S&P 500 index options, calculated and published by the Chicago Board Options Exchange (CBOE).	United States	https://en.wikipedia.org/wiki/VIX

Chinese Indicators

Indicator	*Abbr*	*Description*	*Description Source*	*Data Source*	*Lags*
ChinaBond: Yield to Maturity	YTM	China Bond Yield to maturity (YTM) is the total return anticipated on China bond if the bond is held until it matures.	https://www.investopedia.com/terms/y/yieldtomaturity.asp	People's Bank of China	0 day
Consumer Expectation, Consumer Confidence, Consumer Satisfaction	CEI CCI CSI	In China, the consumer confidence index is based on a survey of 700 individuals over 15 years old from 20 cities all over the country. This composite index covers the consumer expectation and consumer satisfaction index, thus measures the consumers' degree of satisfaction about the current economic situation and expectation on the future economic trend.	https://tradingeconomics.com/China/consumer-confidence	Sina Finance http://finance.sina.com.cn/mac/#boom-4-0-31-2	25-30 days
China Consumer Price Index	CPI	In China, the Consumer Price Index or CPI measures changes in the prices paid by consumers for a basket of goods and services.	https://tradingeconomics.com/China/consumer-price-index-cpi	Sina Finance http://finance.sina.com.cn/mac/#price-0-0-31-2 https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	10-15 days
China GDP	GDP	The gross domestic product (GDP) measures of national income and output for a given country's economy. The gross domestic product (GDP) is equal to the total expenditures for all final goods and services produced within the country in a stipulated period of time. The data includes GDP From Agriculture, Construction and Manufacturing.	https://tradingeconomics.com/China/gdp	Sina Finance http://finance.sina.com.cn/mac/#nation-0-0-31-1	20-30 days
China Interest Rate		In China, interest rates decisions are taken by The Peoples' Bank of China Monetary Policy Committee.	https://tradingeconomics.com/China/interest-rate	Sina Finance http://finance.sina.com.cn/mac/#fininfo-3-0-31-2	0 day
Keqiang Index		Keqiang index, created by The Economist to measure China's economy using three indicators, as better economic indicator than official numbers of GDP	https://en.wikipedia.org/wiki/Li_Keqiang_index	DataYes https://rs.datayes.com/indicator/1010000244	25-30 days
China Loan Prime Rate		In China, the prime loan rate is the weighted average rate of interest charged on loans by three major banks to private individuals and companies.	https://tradingeconomics.com/China/bank-lending-rate	Bank of China	0 day

Coincident Index *Leading Index* *Lagging Index*		The China Economic Monitoring and Analysis Center publishes monthly diffusion indexes of official statistics relating to economic indicators.	CEMAC website	Sina Finance http://finance.sina.com.cn/mac/#boom-0-0-31-2	20-25 days
National Public Expenditure		Public expenditure is the value of goods and services bought by the State and its articulations.	http://www.economicswebinstitute.org/glossary/pubexp.htm	Ministry of Finance	1 month
National Public Finance Income		Revenue earned and reported on the monetary statements of a government. In most business situations, a company's financial income will not usually be the same as its taxable income that is reported annually on its income tax return.	http://www.businessdictionary.com/definition/financial-income.html	Ministry of Finance	1 month
China Producer Price Index	PPI	In China, the Producer Price Index measures the average change in price of goods and services sold by manufacturers and producers in the wholesale market during a given period.	https://tradingeconomics.com/China/producer-prices	Sina Finance http://finance.sina.com.cn/mac/#price-3-0-31-2	10-15 days
ETF Volatility Index	ETF_VOLATILITY	CBOE China ETF Volatility Index, Index, Daily, Not Seasonally Adjusted	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	International Monetary Fund via Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	12-15 days
Daily Exchange Rate	DAILY_EXCH	China / U.S. Foreign Exchange Rate, Chinese Yuan to One U.S. Dollar, Daily, Not Seasonally Adjusted	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	10-15 days
Recession Indicator 1	CHNRECD	OECD based Recession Indicators for China from the Period following the Peak through the Trough, +1 or 0, Daily, Not Seasonally Adjusted	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	14-15 days
Recession Indicator 2	CHNRECDM	OECD based Recession Indicators for China from the Peak through the Trough, +1 or 0, Daily, Not Seasonally Adjusted	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	15-20 days

Recession Indicator 3	CHNRECDP	OECD based Recession Indicators for China from the Peak through the Period preceding the Trough, +1 or 0, Daily, Not Seasonally Adjusted	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	20-25 days
Narrow Money	M0	M0 comprises currency issued by the PBC less the amount held by banking institutions. It is a measure of the money supply which combines any liquid or cash assets held within a central bank and the amount of physical currency circulating in the economy. In some parts of the world, the M0 supply is referred to as narrow money.	https://www.investopedia.com/terms/m/m1.asp#ixzz56uFF8mGE	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	10-15 days
M1	M1	M1 is a metric for the money supply of a country and includes physical money — both paper and coin — as well as checking accounts, demand deposits and negotiable order of withdrawal (NOW) accounts. The most liquid portions of the money supply are measured by M1 because it contains currency and assets that can be converted to cash quickly. "Near money" and "near, near money," which fall under M2 and M3, cannot be converted to currency as quickly.	https://www.investopedia.com/terms/m/m1.asp#ixzz56uFF8mGE	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	45 days
M2	M2	M2 is a measure of the money supply that includes all elements of M1 as well as "near money." M1 includes cash and checking deposits, while near money refers to savings deposits, money market securities, mutual funds and other time deposits. These assets are less liquid than M1 and not as suitable as exchange mediums, but they can be quickly converted into cash or checking deposits.	https://www.investopedia.com/terms/m/m1.asp#ixzz56uFF8mGE	Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	30 days
Reserves_Dollars	Reserves_Dollars	Total Reserves excluding Gold for China in US Dollars		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	
Exports	Exports	Goods, Value of Exports for China		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	10-12 days
Imports	Imports	Goods, Value of Imports for China		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	

Manufacturing Confidence Indicators	Manu_CI	Business Tendency Surveys for Manufacturing: Confidence Indicators: Composite Indicators: OECD Indicator for China, Normalised (Normal=100), Monthly, Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	1 week
Exch Rate based on Manu CPI	EXCH_MANUCPI	Real Effective Exchange Rates Based on Manufacturing Consumer Price Index for China, Index 2010=1, Monthly, Not Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	15-20 days
Consumer Opinion Surveys: EC	COS_EC	Consumer Opinion Surveys: Confidence Indicators: Composite Indicators: European Commission and National Indicators for China, Net Percent, Monthly, Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	20-30 days
Consumer Opinion Surveys: OECD	COS_OECD	Consumer Opinion Surveys: Confidence Indicators: Composite Indicators: OECD Indicator for China, Normalised (Normal=100), Monthly, Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	30-45 days
Total Share Prices	TOTAL_SHR_PR	Total Share Prices for All Shares for China, Index 2010=1, Monthly, Not Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	1 month-45 days
Consumer Price Index	CPI	Consumer Price Index: All Items for China, Index 2010=100, Monthly, Not Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	10 days
Consumer Price Index Food	CPI_FOOD	Consumer Price Index: Total Food Including Restaurants for China, Index 2010=1, Monthly, Not Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	8-10 days
China-U.S. Exchange Rate	EXCHUS	China / U.S. Foreign Exchange Rate, Chinese Yuan to One U.S. Dollar, Monthly, Not Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	8-10 days

Economic Policy Uncertainty Index	CHIEPUINDXM	Economic Policy Uncertainty Index for China, Index, Monthly, Not Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	8-10 days
Import Price Index	CHNTOT	Import Price Index: China - All commodities, Index Dec 2003=100, Monthly, Not Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	8-10 days
Discount Interest Rates	INT_RATES	Interest Rates, Discount Rate for China, Percent per Annum, Monthly, Not Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	
M3	M3	M3 for China, National Currency, Monthly, Seasonally Adjusted		Federal Reserve https://fred.stlouisfed.org/tags/series?t=China%3Bcpi	8-10days
Manufacturing Purchasing Managers'' Index (%)	M_PMI			http://www.stats.gov.cn/english/	
Non-Manufacturing Business Index (%)	NONMAN_PMI			http://www.stats.gov.cn/english/	
Consumer Price Index (The same month last year=100)	CPI			http://www.stats.gov.cn/english/	
Total Sale of Commercialized Residential Buildings Sold, Accumulated(100 million yuan)	TSALE_AMOUNT			http://www.stats.gov.cn/english/	20-25 days
Total Sale of Commercialized Residential Buildings Sold, Accumulated Growth Rate(%)	TSALE_RATE			http://www.stats.gov.cn/english/	20-25 days

ETF

Ticker	*Full Name*	*Market*	*Stock Exchange*	*Equity*	*Index*	*Curency*
ASHR	X-trackers Harvest CSI 300 China A-Shares ETF	Long	NYSE	China	CSI 300 Index	USD
CHAU	Direxion Dly CSI 300 CHN A Shr Bl 2X ETF	Long	NYSE Arca	China	CSI 300 Index	USD
ASHS	Xtrackers Harvest CSI 500 CHN A SmCp ETF	Long	NYSE Arca	China	CSI 500 Index	USD
FXI	iShares China Large-Cap ETF	Long	NYSE Arca	China	FTSE China 50 Index-USD NET	USD
XPP	ProShares Ultra FTSE China 50	Long	NYSE Arca	China	FTSE China 50 Index-USD NET	USD
YINN	Direxion Daily FTSE China Bull 3X ETF	Long	NYSE Arca	China	FTSE China 50 Index-USD NET	USD
YANG	Direxion Daily FTSE China Bear 3X Shares	Short	NYSE Arca	China	FTSE China 50 Index-USD NET	USD
YXI	ProShares Short FTSE China 50	Short	NYSE Arca	China	FTSE China 50 Index-USD NET	USD
FXP	ProShares UltraShort FTSE China 50	Short	NYSE Arca	China	FTSE China 50 Index-USD NET	USD
CNXT	VanEck Vectors ChinaAMC SME-ChiNext ETF	Long	NYSE Arca	China	SME-ChiNext 100 Index	USD
TAO	Guggenheim China Real Estate ETF	Long	NYSE Arca	China	AlphaShares China Real Estate Index	USD
HAHA	CSOP China CSI 300 A-H Dynamic ETF	Long	NYSE Arca	China	CSI 300 Index	USD

USA Indicators

Indicator	*Description*	*Abbr*	*Description Source*
GDP Growth Rate	The real economic growth rate measures economic growth, in relation to gross domestic product (GDP), from one period to another, adjusted for inflation - in other words, expressed in real as opposed to nominal terms. The real economic growth rate is expressed as a percentage that shows the rate of change for a country's GDP from one period to another, typically from one year to the next.		https://www.investopedia.com/terms/r/realeconomicrate.asp?ad=dirN&qo=serpSearchTopBox&qsrc=1&o=40186

Inflation Rate	Inflation is the rate at which the general level of prices for goods and services is rising and, consequently, the purchasing power of currency is falling. Central banks attempt to limit inflation, and avoid deflation, in order to keep the economy running smoothly.		https://www.investopedia.com/terms/i/inflation.asp?ad=dirN&qo=investopediaSiteSearch&qsrc=0&o=40186
Balance of Trade	The balance of trade (BOT) is the difference between the value of a country's imports and its exports for a given period. The balance of trade is the largest component of a country's balance of payments (BOP). Economists use the BOT as a measure of the relative strength of a country's economy. The balance of trade is also referred to as the trade balance or the international trade balance.		https://www.investopedia.com/ask/answers/061515/what-impact-does-balance-trade-have-gdp-calculations.asp?ad=dirN&qo=serpSearchTopBox&qsrc=1&o=40186
Fedral Debt	The federal or national debt is simply the net accumulation of the federal government's annual budget deficits: It is the total amount of money that the U.S. federal government owes to its creditors.		https://www.investopedia.com/updates/usa-national-debt/
Treasury Long Term Rate	Market yield on U.S. Treasury securities at 10-year constant maturity, quoted on investment basis. Long-term interest rates refer to government bonds maturing in ten years. Rates are mainly determined by the price charged by the lender, the risk from the borrower and the fall in the capital value. Long-term interest rates are generally averages of daily rates, measured as a percentage. These interest rates are implied by the prices at which the government bonds are traded on financial markets, not the interest rates at which the loans were issued. In all cases, they refer to bonds whose capital repayment is guaranteed by governments. Long-term interest rates are one of the determinants of business investment. Low long-term interest rates encourage investment in new equipment and high interest rates discourage it. Investment is, in turn, a major source of economic growth. (https://data.oecd.org/interest/long-term-interest-rates.htm)	US_LT_INT	https://www.federalreserve.gov/datadownload/Choose.aspx?rel=H15

Short term interest rates	Market yield on U.S. Treasury securities at 3-month constant maturity, quoted on investment basis Short-term interest rates are the rates at which short-term borrowings are effected between financial institutions or the rate at which short-term government paper is issued or traded in the market. Short-term interest rates are generally averages of daily rates, measured as a percentage. Short-term interest rates are based on three-month money market rates where available. Typical standardised names are "money market rate" and "treasury bill rate".	US_ST_INT	https://www.federalreserve.gov/datadownload/Choose.aspx?rel=H15 https://fred.stlouisfed.org http://www.oecd-ilibrary.org/finance-and-investment/short-term-interest-rates/indicator/english_2cc37d77-en
Interest Rate	Interest rate is the amount charged, expressed as a percentage of principal, by a lender to a borrower for the use of assets. Interest rates are typically noted on an annual basis, known as the annual percentage rate (APR).		https://www.investopedia.com/terms/i/interestrate.asp?ad=dirN&qo=serpSearchTopBox&qsrc=1&o=40186

Appendix B - Risks and Mitigations during five sprints

Sprint 1: Problem Definition

Risk: failure to choose proper factors/market indices

Probability: high

Impact: low

Mitigation: same trend indices fluctuate within a certain range

Risk: not enough investment knowledge

Probability: medium

Impact: medium

Mitigations: learning-by-doing; refer to PGI

Risk: analytic methods fail

Probability: medium

Impact: medium

Mitigations: Apply multiple techniques; Integrate

Risk: failure to achieve an annual return of 8 percent.

Probability: medium

Impact: high

Mitigation: create new model

Sprint 2: Dataset

Risk: failure to access certain database

Probability: low

Impact: high

Mitigation: prepare a list for PGI to get internal data

Risk: Failure to choose final dataset

Probability: Medium

Impact: Low

Mitigation: Important features are quite same; Refer to PGI

Sprint 3: Analytics and Algorithms

Risk: Analytic Methods failure

Probability: Medium

Impact: High

Mitigation: Choose the most reasonable result and check the feature importance; Refer to current model to build our model in Excel based on important features; Try new ideas like sentiment analysis

Risk: We may continuously come up with new ideas

Probability: High

Impact: High

Mitigation: If the new ideas can be realised and coded, add to our model; If not, refer to PGI

Risk: Time limits we cannot meet all objectives provided by PGI

Probability: High

Influence: Low

Mitigation: Finish key milestone 1- factors and 2- ML algorithms; Wrap-up our work and provide our results to next semester capstone project team

Risk: Potential regression/classification methods fail

Probability: Low

Influence: Medium

Mitigation: Prepare several ML methods; Prepare other methods other than regression/classification

Risk: Fail to interpret ML algorithm results to PGI

Probability: High

Influence: Low

Mitigation: Use more visualizations than algorithm formulas; Use simple words to present our works

Risk: Time limits for final presentation, we cannot show all works

Probability: High

Influence: Medium

Mitigation: Manage presentation time effectively (60% - problem and objectives, 30% - analytics methods, and 10% - conclusion); Choose the most important parts to present; Hard to explain or additional information in back-up slides

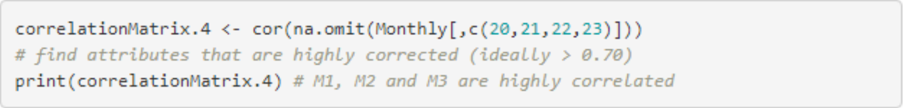
Risk: Continuously come up with new ideas

Probability: High

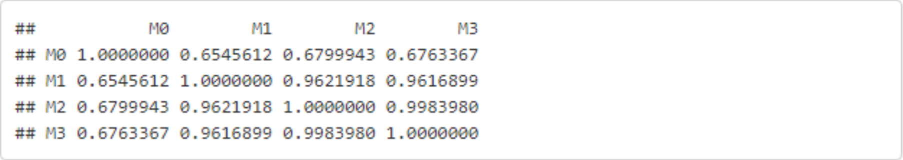
Influence: Low

Mitigation: Stop coding or visualizing new ideas after April; Convey our ideas to next semester capstone project team

Appendix C - Outputs of Multicollinearity check

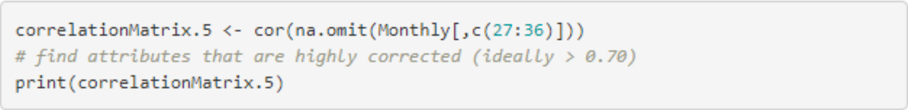


```
correlationMatrix.4 <- cor(na.omit(Monthly[,c(20,21,22,23)]))
# find attributes that are highly corrected (ideally > 0.70)
print(correlationMatrix.4) # M1, M2 and M3 are highly correlated
```

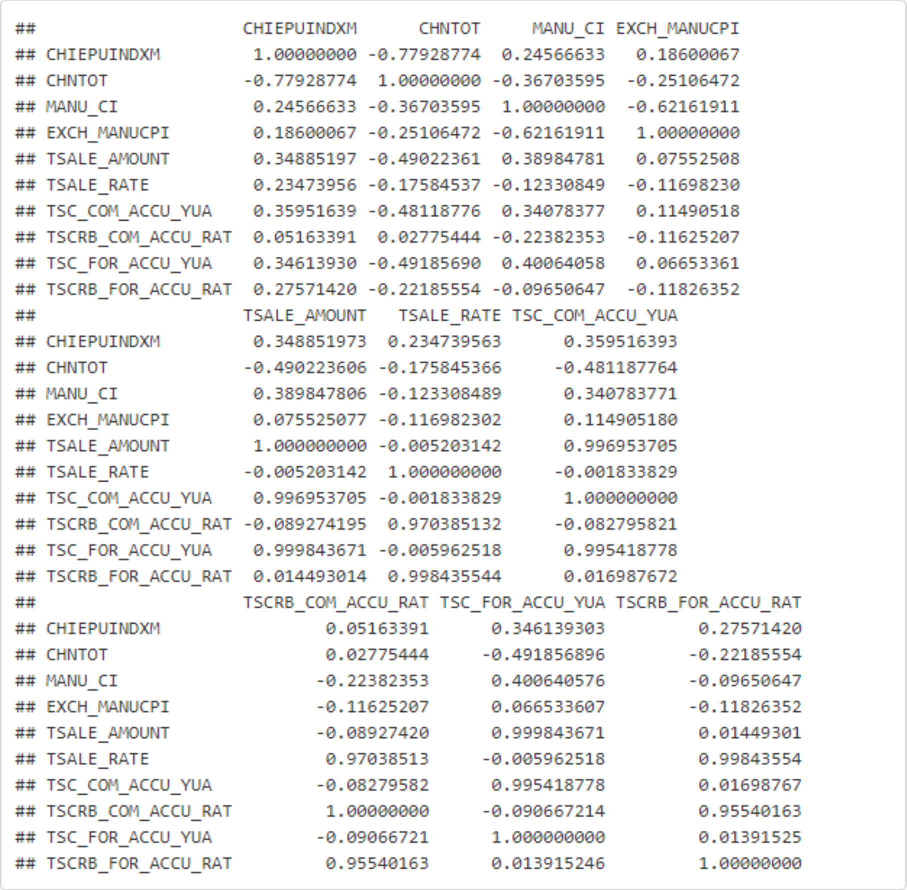


```
##           M0        M1        M2        M3
## M0 1.0000000 0.6545612 0.6799943 0.6763367
## M1 0.6545612 1.0000000 0.9621918 0.9616899
## M2 0.6799943 0.9621918 1.0000000 0.9983980
## M3 0.6763367 0.9616899 0.9983980 1.0000000
```

We checked the correlation among four attributes: M0, M1, M2 and M3, and found M1, M2 and M3 are highly correlated.

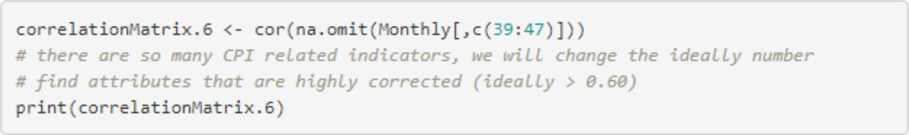


```
correlationMatrix.5 <- cor(na.omit(Monthly[,c(27:36)]))
# find attributes that are highly corrected (ideally > 0.70)
print(correlationMatrix.5)
```

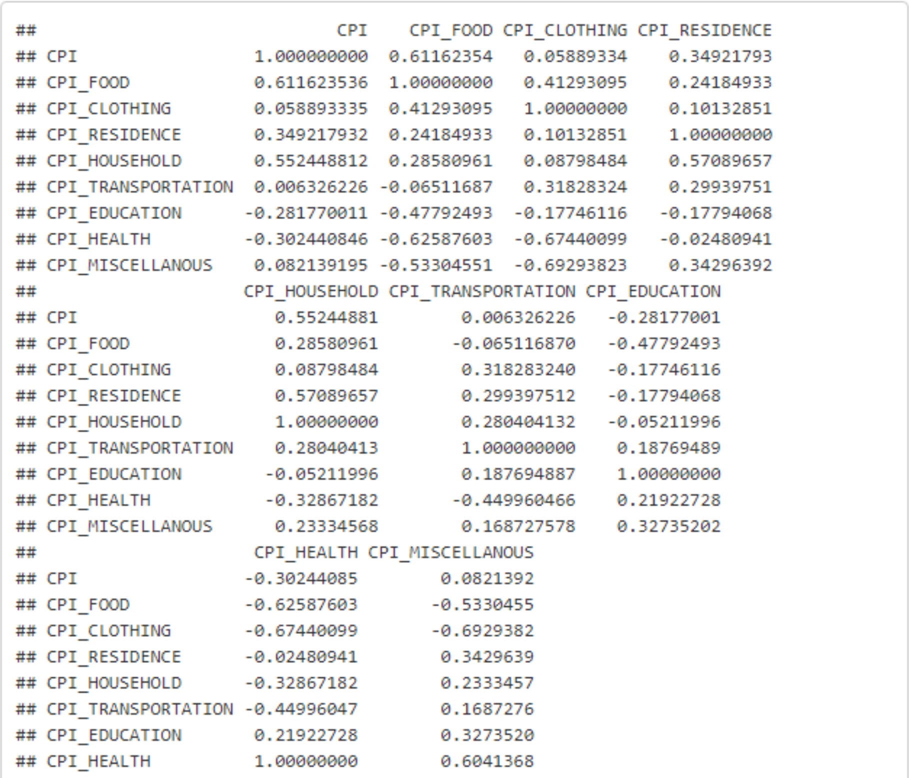


```
##                    CHIEPUINDXM      CHNTOT     MANU_CI EXCH_MANUCPI
## CHIEPUINDXM         1.00000000 -0.77928774  0.24566633   0.18600067
## CHNTOT             -0.77928774  1.00000000 -0.36703595  -0.25106472
## MANU_CI             0.24566633 -0.36703595  1.00000000  -0.62161911
## EXCH_MANUCPI        0.18600067 -0.25106472 -0.62161911   1.00000000
## TSALE_AMOUNT        0.34885197 -0.49022361  0.38984781   0.07552508
## TSALE_RATE          0.23473956 -0.17584537 -0.12330849  -0.11698230
## TSC_COM_ACCU_YUA    0.35951639 -0.48118776  0.34078377   0.11490518
## TSCRB_COM_ACCU_RAT  0.05163391  0.02775444 -0.22382353  -0.11625207
## TSC_FOR_ACCU_YUA    0.34613930 -0.49185690  0.40064058   0.06653361
## TSCRB_FOR_ACCU_RAT  0.27571420 -0.22185554 -0.09650647  -0.11826352
##                    TSALE_AMOUNT   TSALE_RATE TSC_COM_ACCU_YUA
## CHIEPUINDXM         0.348851973  0.234739563      0.359516393
## CHNTOT             -0.490223606 -0.175845366     -0.481187764
## MANU_CI             0.389847806 -0.123308489      0.340783771
## EXCH_MANUCPI        0.075525077 -0.116982302      0.114905180
## TSALE_AMOUNT        1.000000000 -0.005203142      0.996953705
## TSALE_RATE         -0.005203142  1.000000000     -0.001833829
## TSC_COM_ACCU_YUA    0.996953705 -0.001833829      1.000000000
## TSCRB_COM_ACCU_RAT -0.089274195  0.970385132     -0.082795821
## TSC_FOR_ACCU_YUA    0.999843671 -0.005962518      0.995418778
## TSCRB_FOR_ACCU_RAT  0.014493014  0.998435544      0.016987672
##                    TSCRB_COM_ACCU_RAT TSC_FOR_ACCU_YUA TSCRB_FOR_ACCU_RAT
## CHIEPUINDXM                0.05163391      0.346139303         0.27571420
## CHNTOT                     0.02775444     -0.491856896        -0.22185554
## MANU_CI                   -0.22382353      0.400640576        -0.09650647
## EXCH_MANUCPI              -0.11625207      0.066533607        -0.11826352
## TSALE_AMOUNT              -0.08927420      0.999843671         0.01449301
## TSALE_RATE                 0.97038513     -0.005962518         0.99843554
## TSC_COM_ACCU_YUA          -0.08279582      0.995418778         0.01698767
## TSCRB_COM_ACCU_RAT         1.00000000     -0.090667214         0.95540163
## TSC_FOR_ACCU_YUA          -0.09066721      1.000000000         0.01391525
## TSCRB_FOR_ACCU_RAT         0.95540163      0.013915246         1.00000000
```

We checked the correlation among ten attributes with same lag days, and found Policy Uncertainty Index and Import Price Index are highly correlated; Total Sale of Buildings Sold, Total Sale of Buildings Sold (complete apartment) and Total Sale of Buildings Sold (forward delivery housing) are highly correlated; Growth Rate of Buildings Sold, Growth Rate of Buildings Sold (complete apartment) and Growth Rate of Buildings Sold (forward delivery housing) are highly correlated.

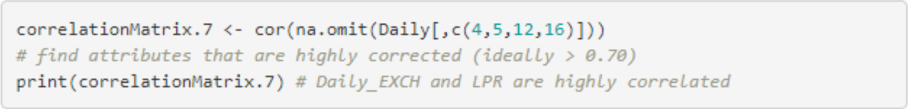


```
correlationMatrix.6 <- cor(na.omit(Monthly[,c(39:47)]))
# there are so many CPI related indicators, we will change the ideally number
# find attributes that are highly corrected (ideally > 0.60)
print(correlationMatrix.6)
```

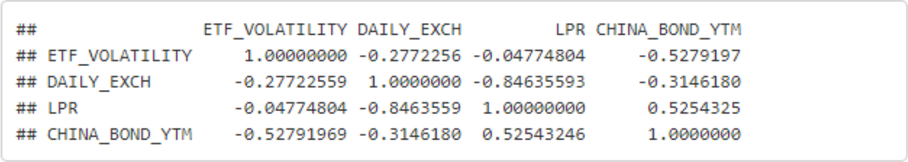


```
##                               CPI    CPI_FOOD CPI_CLOTHING CPI_RESIDENCE
## CPI                   1.000000000  0.61162354   0.05889334    0.34921793
## CPI_FOOD              0.611623536  1.00000000   0.41293095    0.24184933
## CPI_CLOTHING          0.058893335  0.41293095   1.00000000    0.10132851
## CPI_RESIDENCE         0.349217932  0.24184933   0.10132851    1.00000000
## CPI_HOUSEHOLD         0.552448812  0.28580961   0.08798484    0.57089657
## CPI_TRANSPORTATION  0.006326226 -0.06511687   0.31828324    0.29939751
## CPI_EDUCATION        -0.281770011 -0.47792493  -0.17746116   -0.17794068
## CPI_HEALTH           -0.302440846 -0.62587603  -0.67440099   -0.02480941
## CPI_MISCELLANOUS      0.082139195 -0.53304551  -0.69293823    0.34296392
##                      CPI_HOUSEHOLD CPI_TRANSPORTATION CPI_EDUCATION
## CPI                     0.55244881        0.006326226   -0.28177001
## CPI_FOOD                0.28580961       -0.065116870   -0.47792493
## CPI_CLOTHING            0.08798484        0.318283240   -0.17746116
## CPI_RESIDENCE           0.57089657        0.299397512   -0.17794068
## CPI_HOUSEHOLD           1.00000000        0.280404132   -0.05211996
## CPI_TRANSPORTATION      0.28040413        1.000000000    0.18769489
## CPI_EDUCATION          -0.05211996        0.187694887    1.00000000
## CPI_HEALTH             -0.32867182       -0.449960466    0.21922728
## CPI_MISCELLANOUS        0.23334568        0.168727578    0.32735202
##                       CPI_HEALTH CPI_MISCELLANOUS
## CPI                  -0.30244085        0.0821392
## CPI_FOOD             -0.62587603       -0.5330455
## CPI_CLOTHING         -0.67440099       -0.6929382
## CPI_RESIDENCE        -0.02480941        0.3429639
## CPI_HOUSEHOLD        -0.32867182        0.2333457
## CPI_TRANSPORTATION -0.44996047        0.1687276
## CPI_EDUCATION         0.21922728        0.3273520
## CPI_HEALTH            1.00000000        0.6041368
```

We checked the correlation among all the CPI related indicators, since there are many indicators and we decided to change the ideally correlation to 0.6 to reduce more redundant attributes. We found CPI and CPI of Food are highly correlated; CPI of Clothing, CPI of Health and CPI of Miscellaneous are highly correlated.

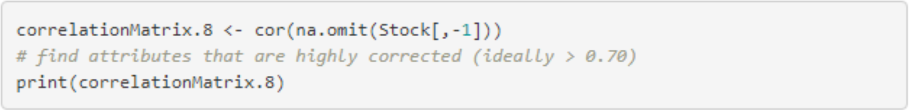


```
correlationMatrix.7 <- cor(na.omit(Daily[,c(4,5,12,16)]))
# find attributes that are highly corrected (ideally > 0.70)
print(correlationMatrix.7) # Daily_EXCH and LPR are highly correlated
```

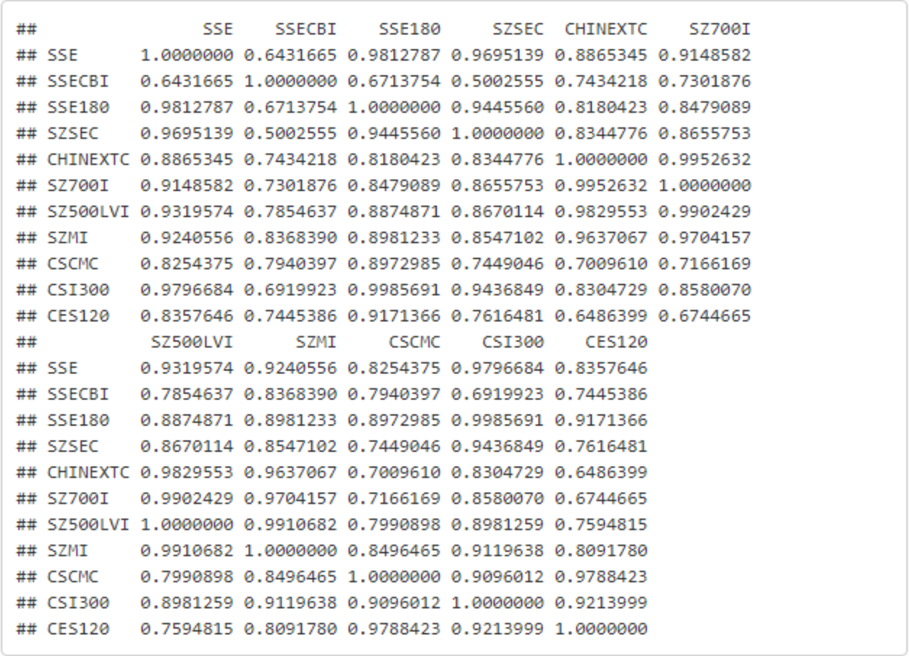


```
##                 ETF_VOLATILITY DAILY_EXCH         LPR CHINA_BOND_YTM
## ETF_VOLATILITY      1.00000000 -0.2772256 -0.04774804     -0.5279197
## DAILY_EXCH         -0.27722559  1.0000000 -0.84635593     -0.3146180
## LPR                -0.04774804 -0.8463559  1.00000000      0.5254325
## CHINA_BOND_YTM     -0.52791969 -0.3146180  0.52543246      1.0000000
```

For daily indicators, Daily Exchange Rate and Loan Prime Rate are highly correlated.



```
correlationMatrix.8 <- cor(na.omit(Stock[,-1]))
# find attributes that are highly corrected (ideally > 0.70)
print(correlationMatrix.8)
```



```
##                SSE    SSECBI    SSE180     SZSEC  CHINEXTC    SZ700I
## SSE      1.0000000 0.6431665 0.9812787 0.9695139 0.8865345 0.9148582
## SSECBI   0.6431665 1.0000000 0.6713754 0.5002555 0.7434218 0.7301876
## SSE180   0.9812787 0.6713754 1.0000000 0.9445560 0.8180423 0.8479089
## SZSEC    0.9695139 0.5002555 0.9445560 1.0000000 0.8344776 0.8655753
## CHINEXTC 0.8865345 0.7434218 0.8180423 0.8344776 1.0000000 0.9952632
## SZ700I   0.9148582 0.7301876 0.8479089 0.8655753 0.9952632 1.0000000
## SZ500LVI 0.9319574 0.7854637 0.8874871 0.8670114 0.9829553 0.9902429
## SZMI     0.9240556 0.8368390 0.8981233 0.8547102 0.9637067 0.9704157
## CSCMC    0.8254375 0.7940397 0.8972985 0.7449046 0.7009610 0.7166169
## CSI300   0.9796684 0.6919923 0.9985691 0.9436849 0.8304729 0.8580070
## CES120   0.8357646 0.7445386 0.9171366 0.7616481 0.6486399 0.6744665
##           SZ500LVI      SZMI     CSCMC    CSI300    CES120
## SSE      0.9319574 0.9240556 0.8254375 0.9796684 0.8357646
## SSECBI   0.7854637 0.8368390 0.7940397 0.6919923 0.7445386
## SSE180   0.8874871 0.8981233 0.8972985 0.9985691 0.9171366
## SZSEC    0.8670114 0.8547102 0.7449046 0.9436849 0.7616481
## CHINEXTC 0.9829553 0.9637067 0.7009610 0.8304729 0.6486399
## SZ700I   0.9902429 0.9704157 0.7166169 0.8580070 0.6744665
## SZ500LVI 1.0000000 0.9910682 0.7990898 0.8981259 0.7594815
## SZMI     0.9910682 1.0000000 0.8496465 0.9119638 0.8091780
## CSCMC    0.7990898 0.8496465 1.0000000 0.9096012 0.9788423
## CSI300   0.8981259 0.9119638 0.9096012 1.0000000 0.9213999
## CES120   0.7594815 0.8091780 0.9788423 0.9213999 1.0000000
```

For those stock indices, only SSE and SSECBI are not highly correlated. After we choose the features and combined them to our final dataset, we will do correlation check again to make sure everything is good to go.

Appendix D – Figures and Tables

Table of Tables

Table of Figures

References

Absolute Return. (2018, Feb). Retrieved Feburary 2018, from https://www.investopedia.com/terms/a/absolutereturn.asp

Amadeo, K. (2018, February 19). *Why the Dollar Is the Global Currency?* Retrieved from https://www.thebalance.com/world-currency-3305931

Barranco, R. (May 2017). Using Machine Learning Gradient Boosting to model commercial. *European Commission - Joint Research Centre (JRC).*

China's Stock Markets vs U.S. Stock Markets. (2015, September 24). Retrieved from https://www.investopedia.com/articles/investing/092415/chinas-stock-markets-vs-us-stock-markets.asp

DAMA UK Working Group. (October 2013). *Defining Data Quality Dimensions.*

Detecting Multicollinearity Using Variance Inflation Factors. (2018). Retrieved from https://onlinecourses.science.psu.edu/stat501/node/347

Dorard, L. (2013, October 16). *When Machine Learning Fails*. Retrieved from http://www.louisdorard.com/blog/when-machine-learning-fails

eHowTech. (2012, May 14). *How to Model Triangular Distribution in Excel*. Retrieved from https://www.youtube.com/watch?v=LOl6GeM5xN4

Focus Economics. (2018, March 20). *China Economic Outlook*. Retrieved from Economic Forecasts from the World's Leading Economists: https://www.focus-economics.com/countries/china

Goh, J., Jiang, F., & Tu, J. (2011). Can US Economic Variables Prediction Chinese Stock Market? *SSRN Electronic Journal*, doi:10.2139.

Goval, A., & Welch, I. (2004). *A Comprehensive Look at the Empirial Performance of Equity Premium Prediction.*

Ho, T. (2012, October 26). *5 Technical Indicators ETF Investors Should Look At*. Retrieved from https://www.investors.com/etfs-and-funds/etfs/best-technical-indicators-for-etf-investors/

Huang, W., Lai, P.-C., & Bessler, D. (February 2018). On the changing structure among Chinese equity markets: Hong Kong, Shanghai, and Shenzhen. *European Journal of Operational Research, 264*(3), 1020-1032.

Indian Agricultural Statistics Research Institute. (2012, Apr 23). *Data Preprocessing Techniques for Data Mining*. Retrieved from http://iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf

Interpreting the Variance Inflation Factor. (2018). Retrieved from http://www.statisticshowto.com/variance-inflation-factor/

Investopedia Staff. "Candlestick Charting: What Is It?." Investopedia. 15 Dec. 2003. Web. 17 Apr. 2018. <https://www.investopedia.com/articles/technical/02/121702.asp>

Jareno, F., & Negrut, L. (2015). US Stock Market and Macroeconomic Factors. *Journal of Applied Business Research (JABR)*, 325.

Kuepper, J. (2017, October 18). *What is a developing country?* Retrieved from https://www.thebalance.com/what-is-a-developing-country-1978982

Loh, W.-Y., Eltinge, J., Cho, M.-J., & Li, Y. (2017). Classification and regression trees and forests for incomplete data from sample surveys. In *Statistica Sinica.*

Major Stock Indices Definition. (2018). Retrieved from https://www.nasdaq.com/markets/indices/major-indices.aspx

Multiple Linear Regression. (1998). Retrieved from Statistical Topics, Yale.edu: http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

Mitchell, Cory. "How To Use A Moving Average To Buy Stocks." Investopedia. 20 May 2014. Web. 16 Apr. 2018. <https://www.investopedia.com/articles/active-trading/052014/how-use-moving-average-buy-stocks.asp>

Nguyen, Q., & Sayim, M. (2016). The Impact of Economic Factors on the Foreign Exchange Rates between USA and Four Big Emerging Countries: China, Inida, Brazil and Mexico. In *International Finance and Banking.*

Noble, J. (2015, July 2). *Why are China's stock markets so volatile?* Retrieved from https://www.cnbc.com/2015/07/02/why-are-Chinas-stock-markets-so-volatile.html

Oscar, T.-R. (July 2014). *Cubic Interpolation using Recognition using R.* Princeton, NJ.

Pricipal Global Investors. (July 2017). *Project Summary –Absolute Return Algorithm.* Des Moines, IA: PGI.

Principal Global Invesors. (Feb 2018). *Market Timing Model Ver 4.0.* Des Moines, IA.

Principal Global Investors. (August 2017). *Principal Life Style Fund - China Equity Fund.* Hong Kong, China: PGI.

Principal Global Investors. (May 2016). *Principal Pivot Series Variable Annuity.* PGI.

Psychlopedia. (2018). *Correlation Coeficient Figure*. Retrieved from https://psychlopedia.wikispaces.com/Correlation+Coefficient

Recognition Lag. (2018). Retrieved from https://www.investopedia.com/terms/r/recognition_lag.asp

Regular Press Release Calendar of NBS in 2018. (2017, 12 28). Retrieved from National Bureau of Statistics of China: http://www.stats.gov.cn/english/PressRelease/ReleaseCalendar/201712/t20171228_1567978.html

Shanghai Stock Exchanges. (2018, March). *List of SSE indices*. Retrieved from http://english.sse.com.cn/indices/list/

Srivastava, T. (2016, May 30). *8 Reasons Why Analytics/Machine Learning Models Fail to Get Deployed*. Retrieved from https://www.analyticsvidhya.com/blog/2016/05/8-reasons-analytics-machine-learning-models-fail-deployed/

Synced. (2018). *Tree Boosting With XGBoost — Why Does XGBoost Win "Every" Machine Learning Competition?* Retrieved from https://medium.com/@Synced/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition-ca8034c0b283

The History of Japanese Candlesticks. (2010, Aug 3). Retrieved from http://www.candlestickforum.com/PPF/Parameters/1_279_/candlestick.asp

Valukonis, M. (2014). China Stock Market Trends and Their Determinants Analysis Using Market Indices. *Economics and Management*, 18(4).

Young, P., & McAuley, J. (Apr 1994). *The Portable MBA in Economics.* John Wiley & Sons.