Predicting Stock Price Index Movement Using Machine Learning Techniques

Master of Applied Economics

University of California, Los Angeles

Author: Dakan Wei

Advisor: Patrick Convery

Date Completed: Jun 12, 2018

**Table of Contents**

1. **Abstract** ..............................................................................................1
2. **Introduction** ......................................................................................2
3. **Section**................................................................................................3
4. **Section**................................................................................................3
5. **Section**................................................................................................3
6. **Conclusion** .........................................................................................12

**Abstract**

For the industry and academia, predicting the movement of the stock market is always a very tempting task. Rather than for obtaining financial benefits, predicting price is more like challenging ourselves. The stock market fluctuates every day and seem unpredictable. Imagine how beautiful it would be if we can find some models in this market to predict the future and beat those experienced business school graduates. So, in this paper, I attempt to forecast the trend (up or down) of an oil company stock in the near future, like the end of the next day, through general machine learning techniques. We are mainly concerned with technical indicators and other potentially related financial data. Experiment result shows that we can only achieve slightly more than 55% accuracy in predicting next-day closing price with traditional classification learning algorithms.

1. **Introduction**

Predicting stock price, or any index, is an art, so is machine learning. Historically, the use of seemingly chaotic market data to predict stock price movements has been an engaging topic attracting for investors and researchers all over the world. Among these popular methods that have been adopted, machine learning technology is trendy in these years, because "machine learning" can identify and analyze trends, including stock trends from a significant amount of historical data and can capture the natural dynamic trends behind stock prices. In this project, I will use supervised learning methods to forecast rise and fall in the stock price before the actual event of an increase or decrease in the stock price occurs.

According to the research on the theory of market efficiency, we learned in asset pricing class; the US stock market now is a semi-strong form efficient market. Efficient market theory tells that the current price of stocks already reflects all available public information, and investors cannot obtain short-term (one day or one week) excess return whether by using a fundamental analysis method or a technical analysis method. In fact, the accuracy of our initial forecast for stock prices for the second day is not high, slightly above 50%. The principal object of this paper was to study and digest the implementation of machine learning knowledge as much as possible in a specific field – Stock Market.

1. **Feature Space**
   1. **Data Collection**

The time-series stock information (opening price, closing price, trading volume, etc.) is pulled from the Quandl database and other financial index data used in the study are mainly from the Investing.com database. I picked energy company Phillips 66's stock data (PSX) from 2013-01-01 to 2017-12-31. The goal of this paper is to predict stock’s trend using various features including technical indicators derived from its time series plus related index for augmentation. The Phillips 66 Company is a multinational oil company headquartered in Houston, Texas. The Philips brothers, founder of the company, started opening the service station in 1927 in Bartlesville, OK, Wichita, Salina, and Topeka, KS. Until 2018, Phillips 66 have been serving Midwest communities more than 90 years and now own over 2000 gas station across the country (**‘About us’, 2018 https://www.phillips66gas.com/about-us**). It went public independently in stock market after ConocoPhillips selling its downstream and midstream assets. Financial index features we added here includes commodity trading prices, currency information, mainstream stock index, dollar index and 10-year Bond Index as follows:

* Index:
  + S&P 500 Historical Data (SP500)
  + NASDAQ Composite Historical Data (NSDQ)
  + Dow Jones Industrial Average Historical Data (DJI)
  + FSTE 100 (FSTE100)
  + DAX Historical Data (DAX)
  + FSTE Singapore Straits Times Price Index (F TSE)
  + Hang Seng Historical Data (HKI)
* Commodity:
  + Gold Historical Price (GLD)
  + United States Oil Fund LP (USO)
  + Brent Oil Futures Historical Price (BOIL)
  + Crude Oil Futures Historical Price (COIL)
* Currency:
  + U.S Dollar Index Futures Historical Price (USDI)
  + Japanese Yen - U.S Dollar Historical Price (JPYUSD)
  + European Euro - U.S Dollar Historical Price (EURUSD)
  + British Pound - U.S Dollar Historical Price (GBPUSD)
* Bond:
  + iShares Core 10Y US Bond (BOND)



These data include daily information (e.g., open, high, low, close, volumes) from the start of January 2013 to the end of December 2017, a total of 1004 data points (544 days up and 460 days down). Besides price information, we also created technical indicators underlying PSX stock price as follows:

* Simple moving average (SMA) and exponential moving average (EMA): indicate the real-time direction of a stock with a lag, and it helps smooth price action and eliminates the short-term shock. In technical analysis, a moving average can act as support or resistance. If the price keeps moving above a moving average level, the forecasting trend of price is increasing. However, if the price keeps moving below a moving average level, the trend is decreasing. The difference between simple moving average and an exponential average is that the latter moves faster and can recognize the change in trend more quickly because it poses more weight on recent price.
  + Formula:
* Commodity channel index (CCI): It is a relatively unique technical analysis indicator. The CCI indicator is a measure of whether the stock price exceeds the normal distribution range and is one of the overbought and oversold categories of indicators. However, it is unique in comparison with other overbought and oversold indicators. Most overbought and oversold indicators have upper and lower bounds. Therefore, they are more suitable for measuring normal market conditions and are likely to fall into passivation when an extreme situation occurs due to the limit of a boundary. However, the CCI does not have such bounds and can fluctuate from positive infinity to negative infinity, so it will pre-caution investors if prices begin to soar or plunge. Formula:

* Ease of Movement (EMV): The purpose of this indicator is to see how easily the stock price changes, that is, how much volume it needs to reflect changes in prices. The price changes require a large volume, it means that the price changes are more difficult, and vice versa. Usually when the stock price peaks and lows, the price changes are more difficult; in the middle of the stock price movement, the price changes are relatively easy.
  + Formula: Distance Moved =

Box Ratio = () /

1. Period EMV =

14- Period EMV = 14-Period SMV of 1-Period EMV

* Rate of Change (ROC): it measures the percentage change in the value of security changes (similar to momentum concept of physics) over a specific of time. A stock with consistent high ROC indicates the stock is likely to outperform the market, so ROC is also a good indicator of a short-term market bubble.
  + Formula: ROC =
* Bollinger Band (BB): Created by Mr. John Bryn, it uses statistical principles to determine the standard deviation of the stock price and its confidence interval to determine the fluctuation range and future trend of the stock price. The band is used to indicate the security level of the stock price. The range of upper and lower limits is not fixed and changes with the rolling of a stock price. The Bollinger scale indicator is a path indicator. The stock price fluctuation is within the upper and lower limit ranges. The width of the strip area changes with the magnitude of the stock price fluctuation. When stock price increases or decreases sharply. The area becomes wider and when the rate of change narrows, the band narrows too (‘Bollinger bands’, n.d.).
  + Formula: Middle Band = 20-day SMV

Upper band = 20-day SMV + 2 \* 20-day standard deviation

Lower band = 20-day SMV - 2 \* 20-day standard deviation

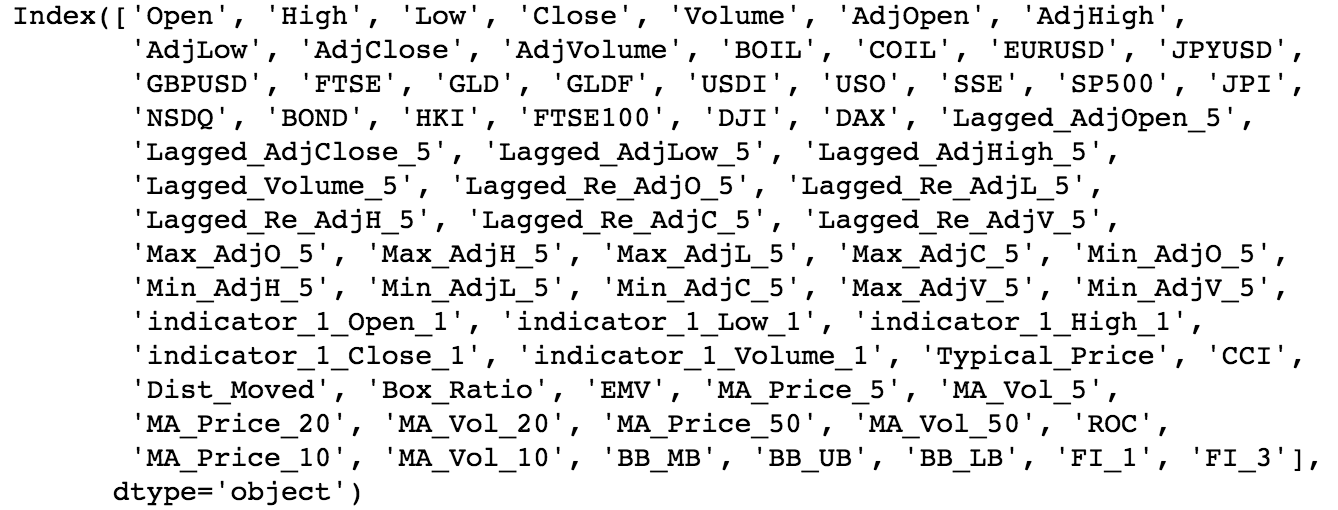
* Force Index (FI): It uses price and trading volume information to assess the bull or bear power behind market price movement. Force Index takes three elements of a stock's price change (direction, extent, and volume) into account to form an oscillator that fluctuates in the positive and negative regions as a signal of potential and price correction (‘Elder’s Force Index’, n.d.).
  + Formula: Force Index (1) = (Close now– Close prior) \* Volume

Force Index (13) = 13 – Period EMA of Force Index (1)

In general, the above technical indicators cover a different type of features of stock:

* Price Indicator: ROC
* Volume Indicator: FI, EMV
* Trend Indicator: SMV
* Momentum Indicator: CCI
* Volatility Indicators: BB
* Noise Elimination: EMV

Those indicators are computed against different periods from 5-day, 10-day, 20-day, 50-day and each of them is treated as individual feature space such as: 'MA\_Price\_5', 'MA\_Price\_10' means the relative price to 3 days ago and to 6 days ago. The features set used for this project is defined as follows:



We use 73 metrics/features in our learning theory. Since we intend to predict “up or down” of stock price, we mark the label as follows: If the adjusted closing price of the stock is higher than that of the previous day, it is marked as "1", otherwise it is marked as "-1" (there is no same closing price of two consecutive days). For example, if a stock's closing price on November 12, 2015, is higher than November 11, 2015, then the label of row “2013-11-11” is +1. All indicators and financial data such as the 5-day simple moving average (SMA), 5-day exponential moving average (EMA), rate of change (ROC) on November 10, 2013, ..., the S&P 500 index are combined to be our feature space: (X1, X2..., X73), so the data of the A stock on November 10, 2013, is (X, Y), where X = (X1, X2..., X73), Y = (+1).

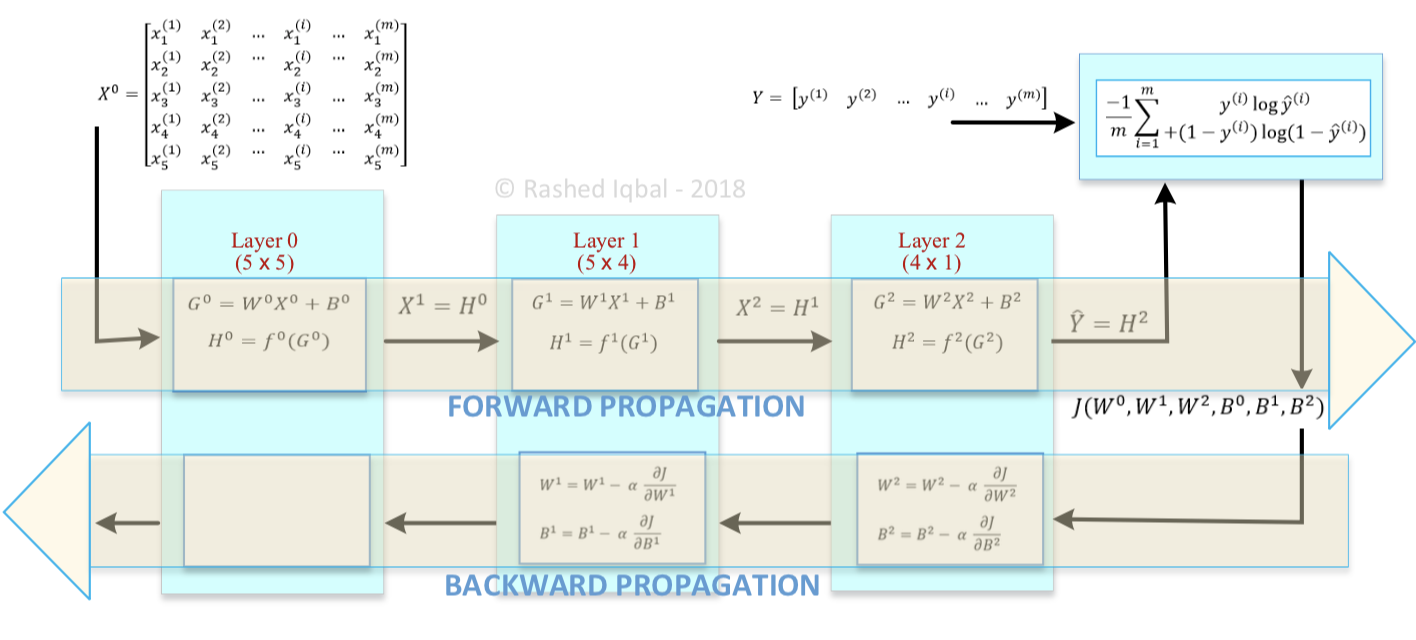
* 1. **Model Selection**

We have learned linear statistical time series models such as ARIMA model to regress the movement of indices. However, stock price and other asset prices is the inherent volatility makes linear techniques suboptimal in predicting. Otherwise, our goal is predicting the trend of movement, so we need to solve classification problem rather than regression problem. Recently, researchers turned to focus on applying machine learning and deep learning techniques to analyze the time series trend. These algorithms take advantage of computer’s computing power to extend the boundary of mathematics and statistics.

Machine learning techniques are generally divided into two parts: supervised learning and unsupervised learning. The major difference between these two techniques is that, in supervised learning, both input and output examples are given. But in unsupervised learning, only input is provided, we need to classify/cluster examples into distinct groups or find other trends from the input based on all information already known.

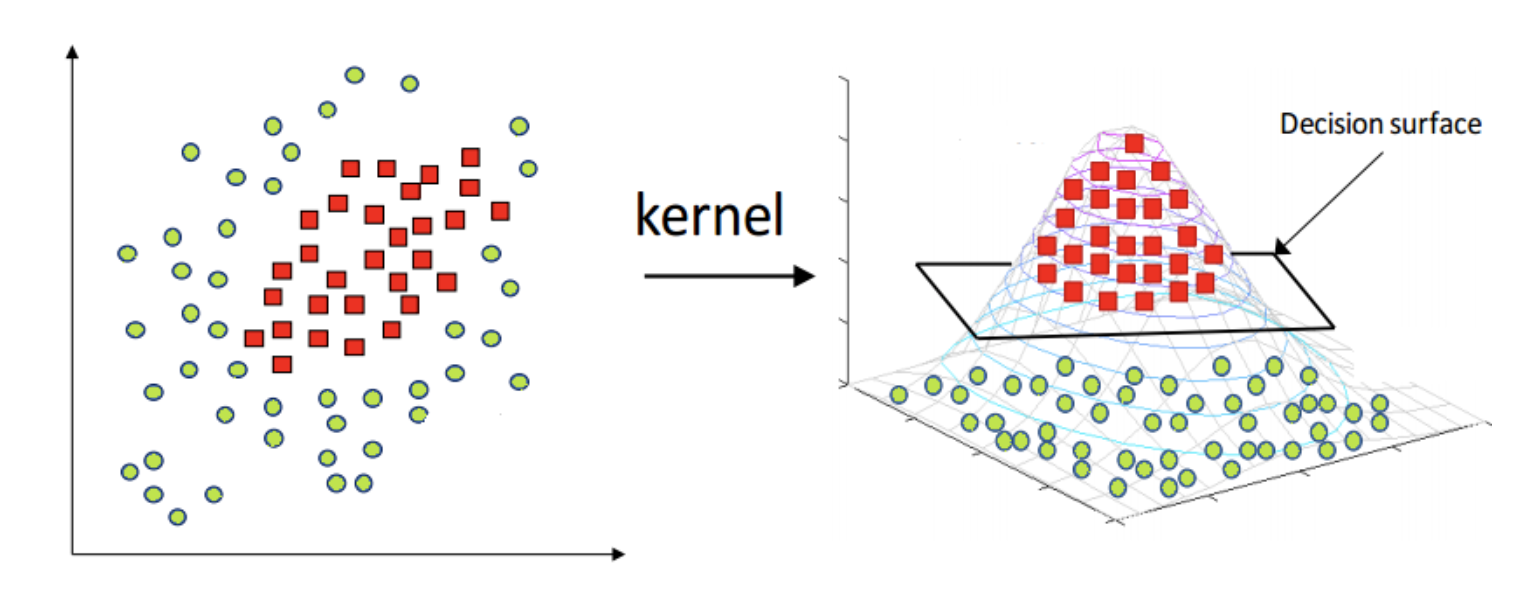
In supervised learning, we are trying to solve two types of problems: classification and regression problems. The result of the regression problem prediction is a continuous value, and the prediction result of the classification problem is discrete. In solving both classification problem and regression problem we must find a real-valued function g(x) according to the training sample. The requirement of the regression problem is: infer a corresponding output according to the training set. That is, use y=g(x) to infer the output value corresponding to an input x. The classification problem is: infer a corresponding category based on the training set (e.g., +1, -1). That is, use y=sign(g(x)) to infer the corresponding class for any input x. In short, the problem of regression is the same as the nature of classification. The only difference is that the. In the classification problem, the output is only allowed to take concrete values; but in the regression problem, the output can take any real number. In this project, we treat the problem of stock price forecasting as a binary classification problem since we only care about up or down of the price trend. The stock’s next day price in the future will be labeled (+1) or (−1) according to the current day’s price.

There is a lot of research focusing on predicting the trend of the stock price using powerful deep learning tool Artificial Neural Network (ANN). In the ANN model, neurons receive input signals from a large number of other neurons. These input signals are transferred through links with weights. Each neuron compares input values with thresholds (bias), then process input through "activation function" to produce the output of the neuron. Summarized in one sentence: The neurons process the received information and pass it on to other neurons. After the iteration of forward propagation and back propagation, ANN gradually reaches its optimal outcome by reducing total cost.



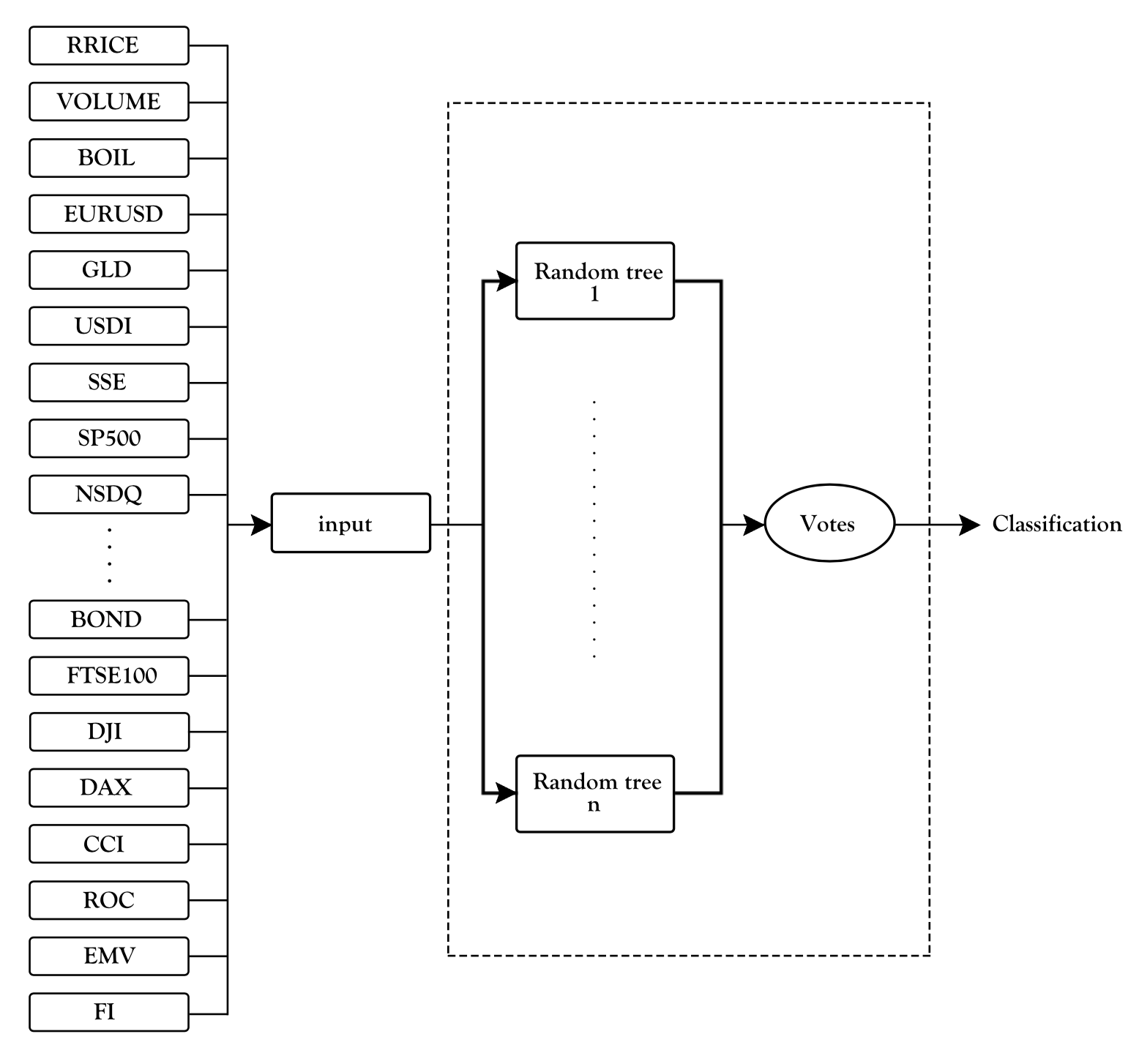
But we will not experiment on ANNs in my project since ANNs and deep learning tools are usually being used to deal with data such as pictures and sound waves that we cannot understand or explain its features. Also, researchers found that due to the volatility and non-linearity of stock price, powerful ANNs can only predict with accuracy slightly greater than 50% (Saahil, 2015). More important, since Phillips 66 went public in the mid of 2012, we can only fetch 5 years’ daily price (1004 days in total). Feeding less than 10,000 data to deep learning system usually does not provide a better result than using general machine learning algorithm.

So, in this project, we will use general classification technique such as Logistic regression, Support Vector Machines (SVM), Random Forest (RF) and XGBoost. Logistic Regression is a machine learning method primarily used to solve the binary classification problem (0 or 1). For example, the likelihood of whether a user buying a product or not, the likelihood of whether a patient suffering from a certain disease or not, and the likelihood of whether a pop-up advertisement being clicked on by a user or not. Note that we use "likelihood" instead of the mathematical "probability" here. The result of the logistic regression is not the probability value in the mathematical definition. It cannot be directly used as a probability value. SVM is another classification approach being researched and tested for stock market prediction very often so far. SVM can be used for both linearly and non-linearly separable datasets. When the data is linearly separable, Support vector machines (SVM) will create a hyperspace to separate examples. When data is non-linearly separable. SVM will classify nonlinear data by mapping data space to a higher dimension and create a new hyperspace (through different dot product, so called kernel function) in this new space to separate examples. In our case, we have used most common Radial Basis Function kernel function in our classifier algorithm.



SVM figure: https://www.hackerearth.com/blog/machine-learning/simple-tutorial-svm-parameter-tuning-python-r/

Another model we are using here is Random Forest (RF). Random Forest is a kind of classifier composed of many decision trees through Bagging. Among them, Bagging is a parallel method that distinguishes Boosting factions (serial methods) from the two main sources of classifier ensemble learning. It is characterized by various weak learners. There is no dependency between them, and they can be fitted in parallel. The Bagging method is the application of Bootstrap random sampling in machine learning. As shown in Figure, we generate N Bootstrap datasets from the original dataset, train a weak classifier for each Bootstrap dataset, and finally use a method such as voting and averaging to form a strong classifier. The Bootstrap random sampling idea is to collect a fixed number of samples from our training set. In other words, the collected samples may be collected more than once after being returned. In the Bagging algorithm, there are *m* sample training sets doing T random sampling, because of the randomness, in general, T sample sets are different. It is because the Bagging algorithm uses different sampling sets to train the model each time, so its generalization ability is very strong and helps to reduce the variance of the model. However, inevitably, the degree of fit to the training set will be worse, which increases the bias of the model.



There are several advantages of random forest. One of the main advantages of RF is its lower need for the quality of data. It does not need any data scaling or normalization before training it since it performs very well on noisy data containing outliers. For another one, they are easily paralleled, which means that they can be used for large-scale machine learning (‘T.Manojlovic’, 2015).

XGBoost is an efficient implementation of the Gradient Boosting method and an improvement and enhancement of the GBDT algorithm. Compared to the traditional GBDT algorithm, XGBoost has been improved in terms of loss function, regularization, point finding, and parallelization design, making it more than five times faster than common toolkits. For example, the GBDT algorithm needs to use the residual of the n−1 tree when training the nth tree, which makes it difficult for the algorithm to implement parallelism; while XGBoost does the second order Taylor expansion for the objective function, making the final objective function only Depends on the first derivative of the loss function at each data point with the second derivative, it is easier to achieve parallelism.

* 1. **Feature Selection and Evaluation**

In this project, we choose to evaluate results by classification accuracy and F-measure. Classification accuracy is the ratio of correctly classified results to total test instances. In other terms, the formula for calculating the classification accuracy is below:

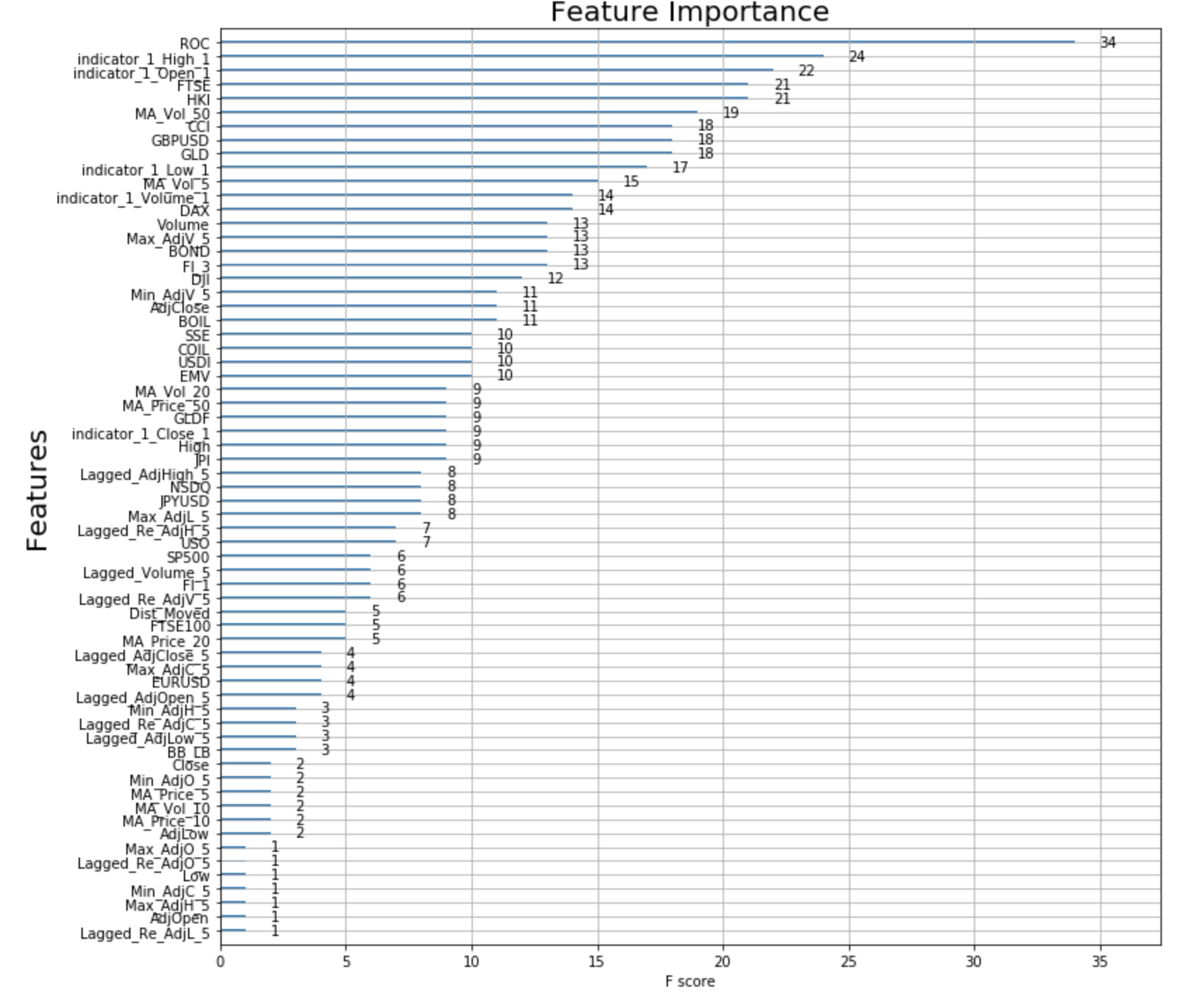
,

For calculating the F-measure, also called F-score, it is necessary to calculate precision and recall:

Our models were evaluated using stratified 5- fold cross-validation. We partitioned our stock price time series data into 5 approximately equal subsets. Each of the subsets remains its continuity in time. For example, we use the stock price of 2013 to 2014 as training data to forecast movement trend of 2015 and make a comparison with the reality. Then use data of 2014 to 2015 to forecast movement trend of 2016. We repeat the process 5 times and calculate the average accuracy of these validations as our final accuracy. Results are shown in table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| Logistic Regression | 53.78% | 0.528 | 0.547 | 0.535 |
| SVM | 52.59% | 0.523 | 0.543 | 0.536 |
| Random Forest | 58.57% | 0.576 | 0.594 | 0.582 |
| XGBoost | 55.38% | 0.532 | 0.575 | 0.556 |

We then try to get the feature importance score of each feature. We will use an algorithm that does feature selection by default – XGBoost. XGBoost uses gradient boosting to optimize creation of decision trees in the ensemble.  Each tree contains nodes, and each node is a single feature. The number of instances of a feature used in XGBoost decision tree’s nodes is proportional to its effect on the overall performance of the model.



Based on the result of Feature Importance Ranking, we selected features whose F score is greater than 11: DJI and above, to build a new dataset. And we apply all machine learning algorithms mentioned to do the experiment again. Results we get are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| Logistic Regression | 54.58% | 0.526 | 0.565 | 0.547 |
| SVM | 54.18% | 0.523 | 0.564 | 0.548 |
| Random Forest | 60.17% | 0.573 | 0.612 | 0.596 |
| XGBoost | 54.98% | 0.536 | 0.556 | 0.549 |

We can see that after feature selection, all of our model improves except for XGBoost. Random Forest performs better than other modes by about 5. It also achieved best performance on Weighted F score: 0.596. But it is too early to apply our model on portfolio management since all experimental result is only slightly above 50%, the chance of flipping a coin. But note here we are predicting the trend of the next-day when may be too soon to reflect or digest all information around the financial market. Because according to the research of Saahil Madge, the accuracy will keep improving until around twenty days ahead (‘Saahil Mad’, 2015).

**Conclusion and Future Study**

In this project, we used the Logistic Regression, Support Vector Machine, Random Forests algorithm and XGBoost models to predict next-day directions of the price of stock PSX. Our evaluation results show that random forests show the best performance of predicting trend among our models, even though all results are very close to each other. We get very similar results to Saahil even though we choose different features. This could be interpreted by the volatility behind the market.

In the future, for making our model more practical, we still have a lot of work to do. Firstly, we can test the result of predicting trend of 1 days ahead until 200 days ahead. From the result we may be able to find the optimal forecasting interval and how long certain technical indicators, or index may affect the underlying stock. Secondly, we can add more features including fundamental information (return of assets, assets turn over, gross profit margin, etc.), text and sentimental information from social network or other publications into our model as features. In addition, we can modify existed technical indicators. Thirdly, we are considering introduce some trading strategy into our model and conduct some back test to see if we can make some real profit in the market.