Master Thesis

A quantitative intraday trading strategy based on regression algorithms

Yuxia Yan
S233163

Supervised by
Cagliero Luca
Paolo Garza

Final Project Report for the
Master in Computer Engineering

Politecnico di Torino
Italy, Turin
July, 2018

ACKNOWLEDGEMENTS

I would like to express my deep sense of gratitude to my supervisors Cagliero Luca and Paolo Garza, for their inspiring & invaluable suggestions. I am deeply indebted to them for giving me a constant guidance throughout this work.

I acknowledge with thanks, the assistance provided by my supervisors. Finally, I would like to thank my classmates and friends directly or indirectly helped me for the same.

Date:02/07/2018
Place:Turin

Yan Yuxia

Contents

1 Introduction

Online trading is basically the act of buying and selling financial products through an online trading platform. These platforms are usually provided by Internet-based brokers and apply to everyone who wants to make money from the market. Its transactions mainly rely on virtual currency to complete the purchase of various physical goods, information services and virtual products. The financial online trading process includes buying and selling bonds, stocks and other investments.

The main strategies used to generate trading signals for the stock market are buying and holding, intraday trading, scalping trading, weekly trading and so on. These trading strategies can be addressed by using technical analysis or fundamental analysis. Fundamental analysis attempts to use data such as revenue, expenses, growth prospects and competitive landscape to calculate the intrinsic value of stocks, while technical analysis uses past market activities and stock price trends to predict future activities. The fundamental uses a long-term approach of market analysis. Therefore, long-term investors often use fundamental analysis because it helps them to choose assets that increase in value over time. Technical analysis uses a relatively short-term method to analyze the market and use it for weeks, days or even minutes. So it is more commonly used by daily traders because it aims to select assets that can be sold to others at higher prices in the short term.

Market data can be automatically crawled, collected, and analyzed. Market price data (current and historical data) is the best source of information for performing price movement analysis to identify patterns in current market trends and correlate them with past trends to predict price changes and trend signals in order to make profitable trades by buying or selling orders. Data-driven methods are used because more data is added every other trading day, resulting in more patterns that can be discovered in the future. The main strategies are quantitative trading strategy using intraday historical price data and end of day historical price data. The quantitative trading strategy using the end of day historical price data is to derive the price model and analyze past historical price data to check whether a similar price model has occurred in the past. If so, it predicts the future price and provides a buying or selling position for forecast price based on the price trend. The quantitative intraday trading strategy is to square-off the trade on the same day. Squaring off the trade means that you must do the buy and sell or buy and sell transaction on the same day before the market close.

The thesis work focuses on investigating the use of regression algorithms to generate trading signals for intraday stock trading. The proposed strategy separately analyzes the historical prices of one stock at a time and discovers patterns relevant for predict the future price of the stock in the next day. The proposed strategy automatically detects the right direction of investment (long- or short-selling) and decides whether to open an intraday trade or not on a given stock based on the expected profit. Therefore, potentially one trading signal per day is generated in case the potential profit of the recommended

trade is sufficiently high.

In this thesis work I conducted a large campaign of experiments on data acquired from the main Italian stock market index (FTSE MIB). To get significant results, we tested our trading strategies in different years (2011, 2013, and 2015), which correspond to different market conditions. We tested a variety of different algorithms in order to look for profitable trend forecasting models by analyzing the prediction performance and financial performance of the proposed models based on different algorithms, such as support vector machines, RepTree, Linear regression, NeuralNet and Random Forest. We tuned the configuration setting of each algorithm in order to achieve better prediction results. We tested also two baseline strategies: random signal generation and follow the trend of the last market day. For random strategy, I generate in Excel a new column where a random value (up, down, or no-action) is stored. Then, I compute the performance according to these values, not close price. The random value is the target that I evaluated. Baseline strategy is an simple strategy for regression that we developed. Specifically, the simple strategy is used to forecast the same direction as the one happened in the last day. The strategy is operated by constructing a sliding window, then the window will advance one day every time. For each sliding window, if you want to predict the trend (up, down or no action) of every stock on specific time point, you just need to evaluate the trend in the previous time point, then your prediction will be the same to the trend of the previous time point.

By running experiments in different models and strategies, we found that, it is successful for day ahead forecasting of daily stock price movements by using these techniques. In addition, the prediction performance and the financial performance of the proposed models were verified and compared. The results show that The RepTree-based prediction model has higher prediction accuracy and financial performance than other models. We can conclude that prediction methods based on regression models in intraday trading can produce more efficient prediction systems than naive approaches.

As a future job, developing long-term forecasting will be interesting through these regression algorithms. Increasing the forecasting range, which is not just limited to daily forecasting.

The paper is organized as follows: in Section 2 the State-of-the-art analysis is introduced. Section 3 describes the proposed methodologies (stock data acquisition and collection, dataset preparation, the prediction algorithms, trading signal generation). Section 4 presents and discusses the results of experiments, I analyze these results by describing the analyzed indices, scenarios and results of different techniques. In section 5, I concludes the paper.

2 Related work

A lot of articles devoted to the financial market analysis are based on the stock movements prediction. As the previous study [14], an intelligent hybrid trading system that uses rough sets and genetic algorithms to discover trading rules

in the future market. The system proposes a new rule discovery mechanism to handle the discretization of data and reduces it through genetic algorithms. And the number of decision sets was noticed and the training period size was analyzed for improving trading performance. The results show that the proposed model is considered as a risk-adjusted measure from average return. An adaptive stock index trading decision support system [5], which is designed to adapt both the inputs and the prediction model based on the final output. Forecasting stock exchange movements using neural networks: Empirical evidence from Kuwait [17], which use two neural network architectures, multilayer perceptron (MLP) neural networks and generalized regression neural networks were used to predict theclosing price changes of KSE. The results of this study show that the neural network computing model is an effective tool for predicting stock trading in emerging markets. A Tensor-based information framework for predicting the stock market [16], which is used to study the influence of the internal relationship of multiple informantion resources on the prediction of stock trends. Financial time series forecasting using artificial neural networks [2], which introduces some financial time series analysis concepts and theories related to the stock market. And these theoryes are based on neural networks and hybrid technologies, which are used to solve several prediction problems involving capital, financial and stock markets. And the study also implements a multi-layer forward neural network for financial time series forecasting system.

There are also a lot of artificial intelligence methods have been proposed and applied to predict stock market indices, for example, Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange [13]. This study attempts to develop two effective models and compare their performance to predict the movement direction of the daily Iatanbul Stock Exchange (ISE) National 100 index. There are ten technical indicators were selected as input for the proposed model. An Ensemble of Neural Networks for Stock Trading Decision Making [3], which uses the intelligent piecewise linear representation method, this method can generate a large number of signals stored from the history database, then the integrated neural network system will be used for training patterns and retrieve similar stock price patterns from historical data for training. Using Volume Weighted Support Vector Machines with walk forward testing and feature selection for the purpose of creating stock trading strategy [26], the result shows that the combination of example weighting and feature selection can significantly increase the overall return on sample trading strategy results. Stock market trend prediction using dynamical Bayesian factor graph [23], the result shows the relationship between each other and the evaluation of these relationships within a certain period of time.

According to its good performance in noisy environments, many methods have been used in various predictions. Forecasting stock indices with back propagation neural network [22], which proposed a new method using noise data to directly predict the stock price or index. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction [4]. Often, these technical analysis and developed intelligence methods

are used together to predict the trend of stocks. A Hybrid Neurogenetic Approach for Stock Forecasting [15], which proposes a hybrid neurogenesis system for stock trading. It designs a context-based neural network integration method that dynamically changes according to the context of the test day. Tests have shown that the average value of the method in terms of purchase and holding strategies has increased significantly, and scenario-based collection has further improved the results. By comparing the predictions of different companies, it is also possible to observe that some companies are more predictable than others, which means that the proposed neurogenic hybrids can be used for financial investment portfolio construction. Predicting stock returns by classifier ensembles [21], which studies the use of a classifier set method to analyze the predictive performance of stock returns. It compares the performance of using two classifier sets with the performance of using a single baseline classifier (e.g, decision trees, neural networks, and logistic regression), and it also examines the average prediction accuracy and return on investment of these models. Their results show that multiple classifiers are superior to single classifiers in terms of prediction accuracy and return on investment.

Another study is prediction of stock price movement based on daily high prices [9], the purpose of this study was to use machine learning techniques to try to use less volatile daily high prices, but the research focused only on a specific non-statistic machine learning approach for a few specific securities. The results show that incorporating statistical classifiers in daily high-price motion forecasts into some simple portfolio management techniques can significantly improve their performance. Random Walks in Stock Market Prices [6], this study briefly describes the theory of random walks and some of the important questions it raises about the work of market analysts. Discuss two common methods of predicting stock prices - chart (or technology) theory and basic (or intrinsic) value theory. The challenge of each analysis is to prove that their method produces more returns than random sampled securities.

Most of the previous researches attempt to assume that every feature makes some same contribution to the classification, but this kind of assumption is not always correct in real world, because the relative importance of every feature is not considered. In my thesis, to achieve accurate prediction of stock price in short term (one day), I use different regression algorithms to create models, to simulate with stock indexes, the results show that the traders can generate higher return and accuarcy.

3 Proposed methodology

In this section, I will describe the methodology used to generate intraday trading signals on stock markets. Trading signals are the driving force for action, and we can buy or sell by analyzing the generated securities or other assets. This kind of analysis can be produced using technical indicators. It can also be generated using mathematical algorithms based on market behavior, or it can be combined with other market factors, such as economic indicators.

The quantified trading signals can be generated based on different types of strategies. Some people want to sell at a higher price when buying at a higher price, while others try to create a huge risk or return rate by buying at low price and selling at high prices or reversing in price action. In general, there are four different types of trading signals. They are momentum signals, breakout signals, buying oversold dips and trend following signals. The momentum signal is based on purchasing power. The momentum trader waits for the strong trend of the stock and buys it in a short period of time. Momentum traders usually trade in a short period of time. These jobs are mainly in the bull market. The breakout signal is based on buying historical highs or 52-week highs, trying to buy highs and sell highers. The breakout is to catch up with the parabolic trend. The stock price may double or even triple in weeks and months. When the index breaks through record highs, these jobs mainly work in a strong bull market. The basis for buying oversold dips is to buy long-term price support or oversold oscillation indicators. This signal attempts to create a huge risk or reward rate by purchasing the difference of the historical price range. Trend following signals attempt to move toward long-term trends by using long-term moving averages. This work applies to the trend of highs or lows. If you want to know more detailed content, you can see [1].

In this thesis, the process of this experiment is divided into 3 steps. The first step is the preprocessing of data. In this process, we will use a sliding window to convert the data that needs to be predicted into a data set one by one for subsequent process. After that, what we need to do is modeling. Support vector machines, RepTree, Linear regression, NeuralNet, Random Forest and baseline strategies are used to product predictive models for direction of movements of financial time series by selecting stock market indices. After these models are created, we can adjust these configuration parameters to achieve better prediction results. Then we collect these prediction results when these models are executed. The last step is to analyze these predictions. We analyze the performance and financial performance from two aspects. The entire process is implemented through java and data mining tools.

The following subsections provide detailed concepts of these methods. By comparing and analyzing the results, we can get the impact of different models on stock prediction.

3.1 Stock data acquisition and collection

The price of a commodity or asset produces a so-called time series. In past several years, different types of financial time series have been recorded and studied. Nowadays, all transactions in the financial market are recorded, resulting in a large amount of available data. Financial time series analysis is very meaning ful to practitioners and theorists for inference and forecastion. Time series data has the characteristics of large data volume, high dimensionality and continuous updating. Moreover, time series data is always considered as a whole rather than a single numeric field. In fact, a large amount of time series data comes from the stock market. The stock time series has its own characteristics

compared with other time series. Recently, the increasing use of time series data has prompted various research and development efforts in the field of data and knowledge management. Therefore, the observation and sample is an important component of the prediction system.

3.1.1 Data collection

Data collection is the process of collecting and measuring information about target variables in a systematic way. People can then answer relevant questions and evaluate the results. Data collection is an important part of many research areas. Although the method varies according to rules, the focus of ensuring accurate and honest collection is consistent. The goal of all data collection is to obtain high-quality evidence, which can lead to convincing and credible answers. Regardless of the field of research, accurate data collection is critical to maintaining the integrity of the research. Choosing the right data can reduce the chance of errors. The formal data collection process is necessary because it ensures that the data collected is accurate and guarantees subsequent development.

In the experimental section, we collect the datasets FTSE MIB 2011, FTSE MIB 2013, and FTSE MIB 2015. The FTSE MIB [25] (Milano indice di Borsa) is the benchmark stock market index for the Borsa Italiana, which is the Italian national stock exchange. The index superseded the MIB-30 in September 2004. This index consists of the 40 most traded stock classes on the exchange. These data is split in single stock. Daily open, high, low, close and volume time series are considered for each stock, where these prices have been adjusted for dividends and splits. These market indices were used to train these models, then using these models to test these stock attribute value in order to generate the forecasting for the stock price trend. In this thesis work, the trend prediction is based on the daily close price. These data samples were formed by three years, respectively. From 1 December 2010 to 30 December 2011, from 1 December 2012 to 31 December 2013, from 1 December 2014 to 31 December 2015, a total of about 830 trading days for each stock. The used length of sample is different between calendar coverage and total sample lengths because of the vacation and missing data and incomplete data. For the incomplete data, e.g.if we need to analyze every row data (open, low, high, close, volume) to predict the trend, but one or more of these data are missing, in this condition, we solve this problem is to ignore the line of data.

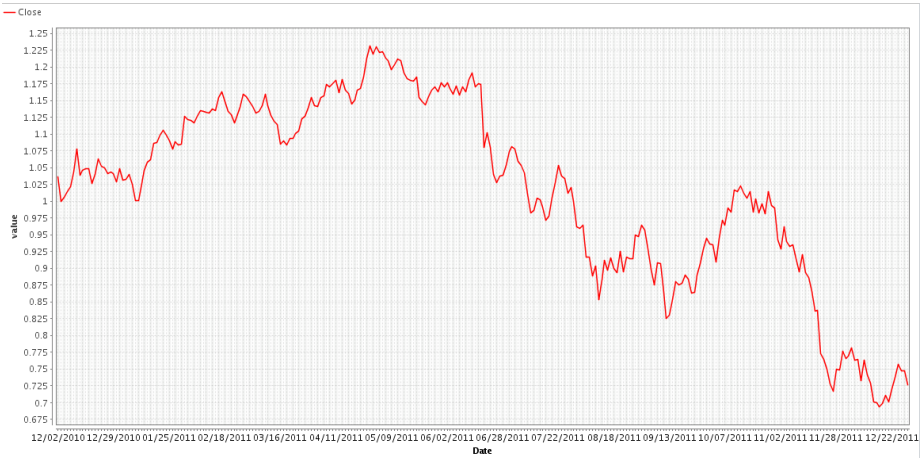


Figure 1: Daily close price of A2A dtock with one year

The visualization of data sets (drawings) illustrate the trends, missing values, noise and outliers. Figure 1 illustrates the close values of A2A stock with one year.

3.1.2 Technical indicators for technical analysis

In contrast to subjective methods, indicators represent statistical methods for technical analysis. By viewing currency flows, trends, volatility and momentum, they provide an aid to actual price movements and help traders to confirm the quality of chart patterns or to generate their own buy or sell signals. There are two main types of indicators: leading indicators and lag indicators. Leading indicators precede price movements and try to predict the future. These indicators are most useful during horizontal or non-trend price movements as they can help identify breakouts or crashes. Lag indicators follow price changes and serve as confirmation tools. In the trend period, these indicators are very useful and can be used to confirm whether the trend still exists or whether it has weakened. We will explain some indicators used for technical analysis in the following.

1.The Acceleration or Deceleration (AC) measures the acceleration or deceleration of current market drivers. The operationg principle of the AC indicator is that the momentum of change shoule be reduced before the price change direction changes. If we wnat to compute the AC, we need to calculate the AO first. The Accelerator Oscillator (AO) show overbought and oversold levels and also be used to find price differences. The accelerating oscillator comes from the awesome oscillator (AO), which is another indicator of Bill Williams. Awesome Oscillator compares the 5-cycle time frame with the 34-cycle time frame to gain an insight into market dynamics. Specifically, AO is a 5-period simple moving average (SMA) that subtracts the median price from the median price

of 34-period SMAs. AC is defined by the following equation:

median price = (High + Low)/2

AO = SMA(median price, 5)-SMA(median price, 34)

AC = AO - SMA5(AO,5)

2.ADX (average direction index) is the main indicator of a technical trading system consisting of five technical indicators. It was developed by J. Welles Wilder, Jr. and calculated using other indicators that make up the trading system. ADX is mainly used as a momentum indicator or trend strength indicator, but the total ADX system is also used as a directional indicator. There are three lines for the computation of this indicator, and the indicator must be calculated based on the real range and the average real range. Assuming the period is 50 trading days, the computation equation is as following:

True range=abs(Current High-Current Low)

+DM1=abs(Current High- previous Close)

-DM1=abs(Current Low- previous Close)

TR50=SUM(True range)

+DM50=SUM(+DM1)

-DM50=SUM(-DM1)

+DI50=(+DM50)*100/TR50

-DI50=(-DM50)*100/TR50

DX=abs((+DI50-(-DI50)))*100/(+DI50+(-DI50))

ADX50=average(DX)

ADX=((ADX50*49)+current DX)/50

3.CCI (Commodity Channel Index) is a multipurpose indicator that can be use to identify new trends or warn of extreme conditions. CCI can be computed with close prices and the previous low and high values for a specified number of trading days. Specifically, CCI measures the current price level based on the average price level on a given period of time. If the price is much higher than the average, the CCI values are relatively high. If the price is far below the average, the CCI price is relatively low. The following equation is based on the

20-period commodity channel index (CCI) calculations.

TP (Typical Price)=(Close+High+Low)/3

Mean Deviation=SUM(abs(TP-average(TP)))/20

CCI=(TP- 20-period SMA of TP)/(0.015 * Mean Deviation)

4.EMA (Exponential moving average) is a method to compute average based on daily close price. This is often used to delay under very important situations, such as real-time financial time series analysis. In this average, the weights decrease exponentially. The value of each sample is smaller than the next most recent sample. With this constraint, you can calculate effectively the moving average. The following is the computation equation:

EMAp=the previous period exponential moving average

EMA=EMAp+ $K * (currentcloseprice - EMAp)$

5.CHO (Chaikin Oscillator) is another method to compute the different between two EMA, its calculations need to use daily close price. For example, we need to compute the difference between EMA3 and EMA10. The equation is the following:

CHO=EMA3-EMA10

6.MACD (Moving Average Convergence or Divergence) is the trading indicator used in technical analysis of stock prices. The MACD indicator is a collection of three time series calculated from historical price data,usually the closing price. The three series are: MACD series itself, "signal" or "average" series. MACD is a trend tracking momentum indicator, comparing two moving averages of prices ("slow" and "fast") history. In general, the "fast" line is the difference between the 26-day and 12-day index financial time series moving average, and the "slow" line is the 9-day exponential moving average of the previous line. If the "fast" line falls below the "slow" line, you can perform sell operation. If the "fast" line rises above the "slow" line, which represent a buy signal.

7.MOM (Momentum index) is a leading indicator of the rate of change in financial time series. It compares the current close price with the previos price before given periods. The following is the computation equation:

MOM=current closing price-OCP, the OCP is the previous closing price for a given period (e.g. 5 trading days).

8.PO (Price Oscillator) is also known as the "percentage price oscillator", which aims to highlight the overall price trend. The calculation of price oscillator

is similar to the moving average indicator and is calculated by averaging two moving averages. In my experiments, I compute it based on 5-period and 10-period.

9.RSI (Relative Strength Index) is a momentum oscillator that measures the speed and change in price movements. RSI changes between 0 and 100. If the RSI-value is below 20 and rises above 20 again, it will generate buy signals. If the RSI value is above 80 and drops below 80 again, it will generate sell signals. The following is the computation equation:

$$\text{RSI=100-100/}(1 + (SUM(positive_changes)/SUM(negative_changes)))$$

10.ROC (Daily Price Rate of Change) is the ratio of the current price to the previous closing price of the scheduled period.

POC=(current closing price/the previous closing price)*100

11.WAD (Williams Accumulation or Distribution) is a volume-based indicator that measure the cumulative flow of funds into and out of securities. By identifying discrepancies between stock prices and flows, we try to measure supply and demand by determining whether investors purchase or sell certain stocks. The accumulation/distribution is calculated by first calculating the money flow multiplier and then the money flow multiplier is then multiplied by the trading volume during the period to calculate accumulation or distribution.

There are more indicators,but we only rely in these indicators in these project.

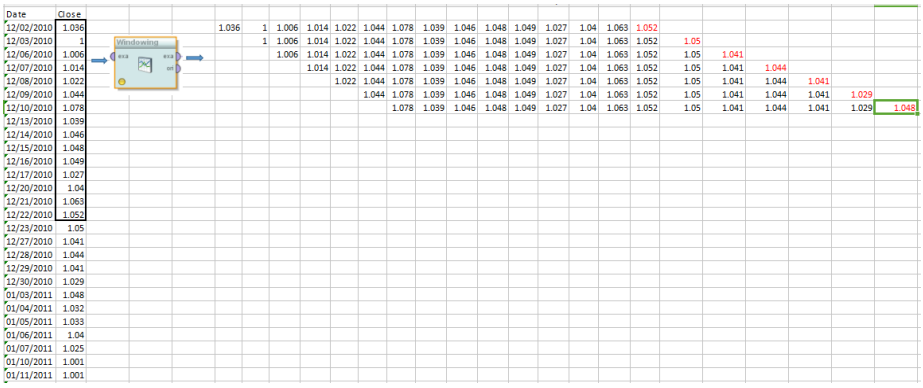
The following is an example of a stock in FTSEMIB in Figure 2.

Date	Open	High	Low	Close	Volume	AC	ADX	CCI	CHO	EMA50	MACD	MOM	PO	RSI	ROC	WAD	WPR
01/04/2016	1.254	1.25	1.22	1.22	12496412	0.00112	8.423388	-88.5476	-0.01292	1.257348	-0.02811	-0.002	-0.0027	40.14933	-2.39234	5.02E+08	-90.5063
01/05/2016	1.234	1.24	1.22	1.24	9773848	0.00265	8.373143	-80.5249	-0.01024	1.256589	-0.02694	0.014	0.0034	43.78245	1.143791	5.08E+08	-81.6456
01/06/2016	1.232	1.25	1.23	1.23	12137840	0.0017	8.304898	-71.0278	-0.01076	1.255546	-0.02511	-0.017	0.0042	42.20582	-0.6462	5.01E+08	-86.7089
01/07/2016	1.215	1.25	1.2	1.24	17878384	-0.0028	8.321221	-76.9971	-0.00586	1.255054	-0.02191	-0.011	0.0017	45.63201	1.056911	5.16E+08	-66.6667
01/08/2016	1.244	1.26	1.23	1.24	14424671	-0.0033	8.302051	-45.7686	-0.00587	1.254268	-0.01918	-0.019	-0.0035	43.907	-0.6436	5.04E+08	-57.8947
01/11/2016	1.24	1.25	1.22	1.24	16012506	-0.0006	8.307858	-46.2084	-0.00375	1.253708	-0.01585	0.016	-0.0017	45.29882	0.404858	5.05E+08	-52.1277
01/12/2016	1.24	1.24	1.17	1.18	38278597	-0.0023	8.462647	-148.059	-0.02258	1.2507	-0.01779	-0.061	-0.0092	33.8888	-5.08065	4.79E+08	-88.9908
01/13/2016	1.2	1.21	1.15	1.16	38161268	-0.0073	8.644575	-178.241	-0.03237	1.2473	-0.02065	-0.066	-0.0141	32.09245	-1.1045	4.57E+08	-88.8889
01/14/2016	1.151	1.16	1.13	1.13	25217827	-0.0142	8.868191	-239.967	-0.0433	1.242818	-0.02562	-0.11	-0.024	28.2473	-2.66323	4.33E+08	-99.2187
01/15/2016	1.126	1.14	1.1	1.12	20817266	-0.0268	9.160322	-253.723	-0.04765	1.23804	-0.03095	-0.114	-0.0335	26.90353	-1.05914	4.38E+08	-85.8025
01/18/2016	1.124	1.13	1.07	1.09	17990465	-0.036	9.495196	-223.71	-0.05655	1.232039	-0.03816	-0.155	-0.0506	23.3195	-3.21142	4.28E+08	-94.086
01/19/2016	1.098	1.1	1.06	1.1	22502234	-0.0382	9.858521	-186.153	-0.05123	1.226743	-0.04291	-0.08	-0.0525	26.81914	1.105991	4.44E+08	-79.902
01/20/2016	1.076	1.08	1.05	1.06	29796865	-0.0362	10.22998	-174.756	-0.05617	1.220204	-0.04925	-0.104	-0.0563	23.28968	-3.37284	4.34E+08	-94.2584
01/21/2016	1.059	1.1	1.04	1.07	22968792	-0.0268	10.55621	-138.55	-0.04959	1.214353	-0.05336	-0.062	-0.0515	26.39116	1.037736	4.35E+08	-87.3239
01/22/2016	1.085	1.09	1.06	1.08	23199550	-0.015	10.87593	-113.076	-0.04048	1.208967	-0.05546	-0.044	-0.0445	28.09879	0.560224	4.39E+08	-84.507
01/25/2016	1.077	1.08	1.03	1.06	29554822	-0.0071	11.24248	-113.38	-0.03912	1.203046	-0.05778	-0.027	-0.0317	26.03877	-1.76416	4.39E+08	-88.8393
01/26/2016	1.05	1.09	1.04	1.09	24279473	0.00163	11.57717	-87.8548	-0.02641	1.198417	-0.05642	-0.012	-0.0249	33.49985	2.551985	4.58E+08	-76.7857
01/27/2016	1.089	1.1	1.07	1.1	24092591	0.01166	11.85992	-62.6711	-0.01309	1.194675	-0.05222	0.043	-0.0102	37.99089	1.658986	4.82E+08	-68.75
01/28/2016	1.103	1.11	1.06	1.07	20418854	0.01703	12.13702	-67.423	-0.0179	1.189668	-0.04961	-0.004	-0.0044	33.16654	-3.26383	4.69E+08	-84.6154
01/29/2016	1.089	1.11	1.08	1.1	19932484	0.02011	12.40857	-47.4552	-0.00742	1.186191	-0.04346	0.024	0.0024	40.81122	3.186504	4.82E+08	-67.3077

Figure 2: the example of indicators based on A2A stock for one month

3.2 Dataset preparation

After acquiring and collecting these stock data, we need to convert these stock data in a packaged structure (dataset) so that different algorithms can be called to train these datasets. So the method used by RapidMiner is the sliding window, which establish fundamentally different financial time series predictions than standard techniques. Windowing converts financial time series data to a generic data set, the last column of the window within a financial time series to a tag or target variable. The concept of typical time series data and its transform structure (after windowing) is as following Figure 3.



Date	Close																								
12/02/2010	1.036			1.036	1	1.006	1.014	1.022	1.044	1.078	1.039	1.046	1.048	1.049	1.027	1.04	1.063	1.052							
12/03/2010	1				1	1.006	1.014	1.022	1.044	1.078	1.039	1.046	1.048	1.049	1.027	1.04	1.063	1.052	1.05						
12/06/2010	1.006					1.006	1.014	1.022	1.044	1.078	1.039	1.046	1.048	1.049	1.027	1.04	1.063	1.052	1.05	1.041					
12/07/2010	1.014						1.014	1.022	1.044	1.078	1.039	1.046	1.048	1.049	1.027	1.04	1.063	1.052	1.05	1.041	1.044				
12/08/2010	1.022							1.022	1.044	1.078	1.039	1.046	1.048	1.049	1.027	1.04	1.063	1.052	1.05	1.041	1.044	1.041			
12/09/2010	1.044								1.044	1.078	1.039	1.046	1.048	1.049	1.027	1.04	1.063	1.052	1.05	1.041	1.044	1.041	1.029		
12/10/2010	1.078									1.078	1.039	1.046	1.048	1.049	1.027	1.04	1.063	1.052	1.05	1.041	1.044	1.041	1.029	1.048	
12/13/2010	1.039																								
12/14/2010	1.046																								
12/15/2010	1.048																								
12/16/2010	1.049																								
12/17/2010	1.027																								
12/20/2010	1.04																								
12/21/2010	1.063																								
12/22/2010	1.052																								
12/23/2010	1.05																								
12/27/2010	1.041																								
12/28/2010	1.044																								
12/29/2010	1.041																								
12/30/2010	1.029																								
01/03/2011	1.048																								
01/04/2011	1.032																								
01/05/2011	1.033																								
01/06/2011	1.04																								
01/07/2011	1.025																								
01/10/2011	1.001																								
01/11/2011	1.001																								

Figure 3: Example of Sliding Window Method

The parameters of the windpw operator allow you to change the size of the window (shown on the left as a colored vertical box), the overlap between windows is step size and the horizon is used to predict. In experiments, the parameters of the window are very important for gaining the successful results. I set the window size is 15, which determines how many "attributes" are created for the cross sectional data. step size is 1, which means that the sliding window advanced one trading day. Therefore, a series of data is now converted to a common data set that can be handled by any available RapidMiner operator.

The next major process required for financial time series analysis is to create some models based on different algorithms.In this study,I using machine learning to create the model by using RapidMiner tools.The process consist of two steps, first, we need to train these model by calling several different algorithm.I will specifically introduce their configuration and use for these algorithms below. Then we load the unlabeled data samples into these models to Evaluate the forecasts. The process is as the following figure 4.

Figure 4.Sliding window method based on one-year testing samples.

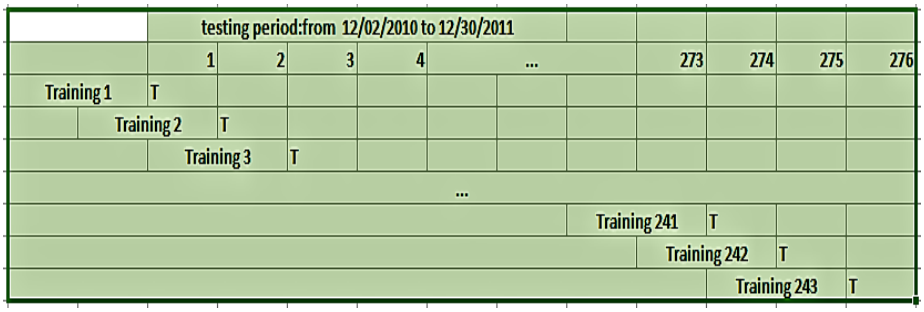


Figure 4: Sliding Window Method

3.3 The prediction algorithms

Machine learning often uses statistical techniques to enable computers to learn from data without explicit programming. Many inductive learning methods, classifier systems and regression systems are included in machine learning. All of these methods need to use a set of data samples to generate an approximation of the subordinate functions. Unlike traditional prediction systems, machine learning techniques can track linear and nonlinear models. It can be said that the disadvantage is that these methods require more calculation time and their performance based on a large number of parameters. In this chapter we will present financial time series forecasting applications using different techniques below.

3.3.1 SVM model

SVM (support vector machines) is a supervised learning model with associated learning algorithms that analyze data using classification and regression analysis. If there are a set of training examples, each marked as belonging to one or the other of two categories. An SVM training algorithm creates a model that distributes new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, which is mapped so that the examples of the separate categories are separated by a clear gap that is as wide as possible. New examples are them mapped into that same space and predicted to belong to a category based on which side of gap they fall. The detailed concept is explained in [10].

More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for regression, classification or other tasks like outliers detection. In practice, real data is messy and can not be separated perfectly with a hyperplane. The limits of the edges of the lines that separates the categories must be relaxed. This is commonly referred to as the soft margin classifier. This change allows certain points in the training data to violate the separating line. An extra set of co-

efficients are introduced to provide the margin wiggle space in each dimension. These coefficients are sometimes referred to as slack variables. This increases the complexity of the model because there are more parameters for the model to fit to the data to provide this complexity. You can learn more from [19].

Classifier model based on SVM was used as the benchmark machine learning method. It is based on the internal Java implementation of mySVM by Stefan Rueping. This learning method can be used for regression or classification and provides a fast algorithm and good results for many learning tasks. Support Vector Machine have various kernel types including dot, radial, polynomial, neural, anova, epachnenikov, gaussian combination and multiquadric.The kernel function for the SVM model is selected is Gaussian Radial basis function.In this paper, we study the financial time series forecasting based on the support vector machine by choosing the dot kernel type. Although the speed of prediction process is slowly, it can increase the prediction accuracy of the financial time series. The experimental results show the prediction accuracy of this approach based on the support vector machine.

A tuning parameter C is introduced that defines the magnitude of the wiggle allowed across all dimensions. The C parameters defines the amount of allowed boundary violations. A C=0 has no violation and we return to the described inflexible Maximal-Margin Classifier. The larger the value of C the more violations of the hyperplane are permitted. The smaller the value of C, the more sensitive the algorithm is to the training data (higher variance and lower bias). The larger the value of C, the less sensitive the algorithm is to the training data (lower variance and higher bias).

In our experiment,the SVM with dot-product kernel function was selected as basis for generated SVM models. The dot-product is called the kernel and can be re-written as:

K(x, xi) = sum(x * xi)

The kernel defines the similarity or a distance measure between new data and the support vectors. The dot product is the similarity measure used for linear SVM or a linear kernel because the distance is a linear combination of the inputs. The following figure 5 illustrates the results obtained by using different inputs and parameters described above.

Figure 5: Daily stock price prediction of A2A stock between single variable and multiple variables for one year

3.3.2 Neural Net model

Neural networks based machine learning systems that are inspired by and modeled loosely off of the idea of the actual brain.The actual brain being this thing with like neurons and axons that connect other neurons.The Neural nets for short is that we have nodes that have some connections between them,this is similar to the neurons in your brain and the synapses they form to get a neuron there to do something.We trigger a node with some input and that node in turn triggers the nodes it is connected to but this alone is not very useful so we usually organize the Neural nets in a way that makes it easy to produce good results.We want our connections to have different values that is some connections should be more important than others.So the connection value is called weights are represented.

Neural networks are a set of algorithms that are designed to recognize patterns.They analyze sensory data through a kind of machine perception ,labeling or clustering raw input.The patterns contained in vectors that they recognize are numerical,must be translated. Neural networks help us cluster and classify.You can think of them as a clustering and classification layer on top of the data you store and manage.They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled data set to train on. The detailed concept can be obtained from [24].

The neural network is one of the so-called "black box" methods because they have almost no responsibility for economic structures. The Neural networks is used to train in order to approach the thinking and behavior of certain stock market traders. Different indicators are able to use as inputs for neural networks, and stock indexs are used to monitor the training process.

Most supervised neural networks are used for financial forecasting because of

inputs (in the past values and a number of technologies or basic indicators) and outputs (target values) are know, and their goal is to discover the relationship between this two kind of variables, that is, a function that approximates the value of the stock market. It is highly recommended to use the basic knowledge of neural networks, because in this article we will deal with various terms, components and methods related to neural networks. A comprehensive explanation for neural networks is given in [11]. Seeing [7] for a more detailed description. Several practical applications and papers can be found on the Internet. RapidMiner programming environment offers the possibility to implement in an elegant way based on Neural Networks.

The most popular supervised Neural Networks involved in financial forecasting are multi-tiered perceptron based on its architecture that belongs to feedforward network. In a feed-forward network, information flows through the network from input to output. Of course, other types of radial basis function networks and competitive networks can also be used. To develop a successful financial forecasting system, which is recommended to follow some suggestions and avoid major pitfalls and common mistakes [18]. In [18] it is pointed out that although Neural Networks can hide many pitfalls, they are more efficient in technical analysis.

Many people have tried to use financial markets to simulate and predict financial markets. Computational tools that can be used to study time series and comples systems, such as linear autoregressive models, principal component analysis, artificial neural networks, genetics algorithms and others. The first Neural Network to forecast market stock trend was implemented by White in 1988. He evaluate fluctuations in the price of common stocks with previously undetected asset price fluctuations.All classification tasks depend upon labeled datasets,that is,humans must transfer their knowledge to the dataset in order for a neural to lean the correlation between labels and data. Clustering or grouping is the detection of similarties.Learning without labels is called unsupervised learning.The more data an algorithm can train on,the more accurate it will be.

In this paper, the neural network is used to gain the relationship between these historical stock prices and the future forecasting. Different models and configurations were experimented by training,validating and testing data sets. The applied NeuralNet-based prediction models were classifiers using different inputs generated based on these training samples. In order to get the hightest accuracy or lowest classification error for the model based on neural network algorithm, some parameters need to be selected. Hidden layers describe the size and the name of all hidden layers, we can use this parameter to define the structure of the Neural Network. The hidden layer size value is set to -1 in the experiment because I want to calculate the layer size based on the number of properties of the input samples. Training cycles are used to specify the number of iterations for neural newwork training. In back propagation, the output values are compared with the correct answer to calculate the value of a predefined error function. In this experiment, I set this process to repeat 500 times in order to make this error smaller. The learning rate (0.3) and the momentum (0.2) for the neural network were very small as well. The following figure 6 illustrates

the results obtained by using different inputs and parameters described above.

Figure 6: Daily stock price prediction of A2A stock between single variable and multiple variables for one year

3.3.3 Linear regression model

Linear regression is a linear method to research the relationship between dependent variables and independent variables. The amount of independent variables can be one or more. If the independent variables is one, that is called simple linear regression. If the independent variables are more, we can call this process multiple linear regression. For dependent variables, the case of more scalar response is called multivariate linear regression, there is a big difference from multiple linear regression. Multivariate linear regression need to predict multiple related dependent variables, rather than a single dependent variable.

In linear regression, we need to create linear models for researching the relationships from the data based on linear predictor functions and analyzing the effective from different parameters. Like other regression analysis, linear regression pay attention to the conditional probability distribution of the prediction value of dependent variables, rather than on the joint probability distribution of all of these variables. Linear regression is widely used in practical applications. There are two reasons, the first is that the model of depending linearly on their unknown parameters are easier to fit than models of depending non-linearly on their parameters. The second is that the statistical properties of the resulting estimators for the model of depending linearly on their unknown parameters are easier to analyze.The more information can be obtained in [8].

Linear regression is used widely in many practical applications, but it is mainly used into two aspects. The first category is for classification, which can be used for forecasting, or prediction, or error reduction. In this case, Linear regression can be used to fit a predictive model based on the values of dependent

variables(response) observed and independent variables. If without collecting an response value, just only collecting the additional values of the independent value, we can use fitted model to predict the response. The another category is to explain variation in the response variable according to the variation in the independent variables. The strength of the relationship between the response and the independent variables can be researched by linear regression analysis. Particularly, it can be used to distinguished if some independent variables may have no linear relationship with the dependent variables at all, or to know which subsets of independent variables may contain redundant information about the dependent variables(response). E.g.Given a data set $\{y_i, x_i l, ..., x_i p\}^n, i = 1$, we assume that the relationship between the dependent variable y and the p-vector of regressors x is linear, then we create a linear regression model to model the relationship by a disturbance term or error variable e-an unobserved random variable.So the model takes the form:

$yi = B_0 * 1 + B_1 * x_i l + ... + B_p * x_i p + e_i, i = 1, ...n$, where

$y = y_1, y_2, ...y_n$,

$e = e_1, e_2, ...e_n$,

In this experiment, the parameter of feature selection is set to M5 prime, which use the so-called AIC (Akaike information Criterion) to select the feature of the linear regression. In short, AIC allows trade-offs between increasing the number of model parameters and using information entropy reduction errors. The reason to choose M5 prime is that the M5 prime feature selection selects the attribute with the smallest normalization coefficient, removes it and performs another regression in each iteration. If there is attribute to improve AIC, the attribute is discarded. This is repeated until no attributes are discarded. The following figure 7 illustrates the results obtained by using different inputs and parameters described above.

Figure 7: Daily stock price prediction of A2A stock between single variable and multiple variables for one year

3.3.4 RepTree model

RepTree(Reduced Error Pruning Tree) is fast decision tree learner, which is used to build a decision or regression tree using information gain or variance and prunes it using reduced-error pruning(with backfitting). This algorithm is first proposed in [20]. The algorithm sorts values for numeric attributes once. Missing values are deal with by splitting the corresponding instances into pieces(i.e.as in C4.5) [12]. RepTree is one of the data mining models for Weka extension of RapidMiner. Weka(Waikato Environment for Knowledge Analysis) is a suite of machine learning software written in java. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a data set or called from your own Java code. More specifically,Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. It is also well-suited for developing new machine learning schemes.

RepTree is a good machine learning tool analyze different data. RepTree uses the logic regression tree and creates multiple tree in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In order to prune the tree, the mean square error is used to measure on the predictions made by the tree.

In my experiments, some parameters of regression model based on RepTree algorithm were set. The M is 2.0, V is 0.001, N is 3.0 and S is 1.0. The RepTree model with these parameters resulting in the lowest prediction error. The walk-forward processing was repeated for each data sets in this procedure. The following figure 8 illustrates the results obtained by using different inputs and parameters described above.

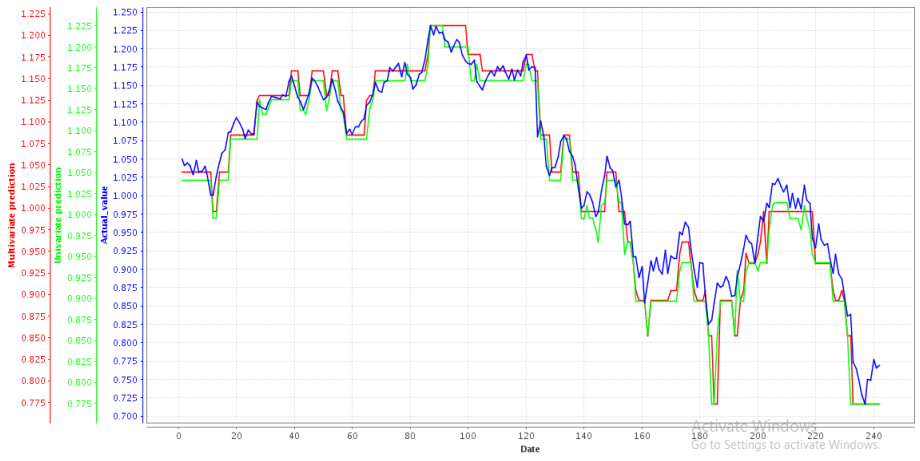


Figure 8: Daily stock price prediction of A2A stock between single variable and multiple variables for one year

3.3.5 Random Forest model

Random forests is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes(classification) or mean prediction(regression) of the individual trees. Random Forest is flexible and easy to use machine learning algorithm, even without hyper-parameter tuning, but a great result need most of the time. It is also one of the most used algorithms,because its simplicity and the fact that it can be used for both classification and regression tasks. Random forests are a way of averaging multiple deep decision trees, trained on different parts of same training set, with the goal of reducing the variance.

Random Forest algorithm is to build multiple decision tree and merge them together in order to get a more accurate and stable prediction. The RepTree algorithm and Random Forest algorithm are related to the decision tree, but their algorithms are very different. Random Forest is not trimmed, except for simple trimming, it will stop at the specified depth. However, RepTree can reduce error clipping. Random Forest considers a set of K randomly selected attributes to split on each node, but RepTree considers all attributes. In my experiment, I set the number of trees to establish is 10. With these parameters, I can get the results by using different inputs and parameters as following Figure 9.

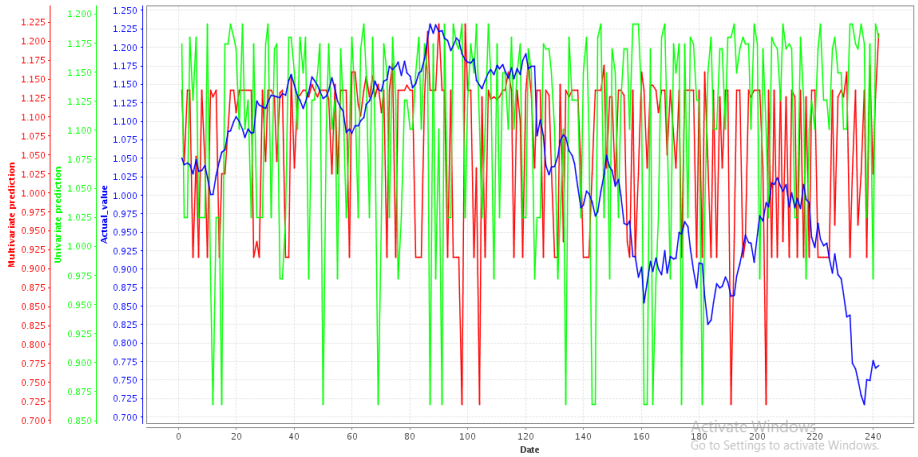


Figure 9: Daily stock price prediction of A2A stock between single variable and multiple variables for one year

3.3.6 Baseline strategy

Baseline strategy is an simple strategy for classification.This strategy does not exist in existing algorithms of RapidMiner,I achieved it by java code.Specifically,the

simple strategy is used to forecast the same direction as the one happened in the last day.The strategy is operated by constructing a sliding window,then the window will advance one day every time.For each sliding window,if you want to predict the trend(up/down/no action)of every stock on specific time point,you just need to evaluate the trend in the previous time point,then your prediction will be the same to the trend of the previous time point.E.g.We suppose that the window size is set to 3,the values in this window is x1,x2,x3.If you want to predict the trend of time point x3,you just need to distinguish the daily variation between the time point x1 and the time point x2,if the difference is positive,then you can predict the trend of time point x3 is up,although this actual direction that the difference between the time point x2 and the time point x3 may be declining.If the variation is negative,then the prediction for time point x3 trend will be down,if the variation is the same with the previous day,the prediction trend for time point x3 is no action.

This strategies were generated and tested using the walk-forward method with a sliding time window.The method allows to perform a series of experiments using train and non-overlapping test samples of constant lengths.For each experiment,SVM,Neural network,RepTree and regression algorithms were called in order to create models that were used in the full training sample and predefined parameter sets.Then these models were applied to test the data sample.For each subsequent experiment,all samples were moved forward by one day until the end of available data was reached.

3.4 Trading signal generation

There are many attempts to find effective trading rules in current financial time series. So many theoretical researchers pay attention to the predictability of financial time series events, mainly about the dynamic changes of the stock market. It is important to know if the time series is meaningfull, because their attributes will indicate whether it is worth making a prediction. There are two important assumptions, which is the characteristic process prediction: (RWH) random walk hypothesis and (EMH) effective market hypothesis. RWH (random walk hypothesis) point out that the market price lingers in a purely random and unpredictable way. EMH (effective market hypothesis) states that the market fully reflects all available information, and once the new information is available, the price will be fully adjusted immediately. In the actual market, some people react to information immediately after receiving the information, while others wait for confirmation of the information. Financial time series forcasting have been used by economists for decades. This involves statistical analysis, the so-called technology and basic analysis, both are considered security analysis.

Trading decisions must be made without knowing the future price, so it is impossible to determine whether the current price is low or high. In order to make the trading rules practical, which means that trading decisions can only use past value but not future price information. For the reliability of some technical trading rules that use only past information, even if we are allowed to use future information, the development of optimal strategies is still

a problem that can not be ignored. We will introduce the method of trading signal generation.

After modeling, we get the predicted values about the target variable. The next process is to analyze these results in order to get the trading signal by using effective trading rules. In my experiments, I calculate the actual direction and forecasted direction of the stock price based on the result. The actual direction is calculated by subtracting the last closing price of the window from the next actual closing price of the stock. If the value is larger than zero, the actual direction is "UP", if the value is smaller than zero, which represent the actual direction is "DOWN", if the value is equal to zero, which means the actual direction is stable. In addition, the forecasted direction is calculated by subtracting the last closing price of the window from the forecasted value. The way to get this direction is the same as the actual method above.

For each walk-forward procedure, if prediction direction is "UP", which means that you can perform "sell" operation, the trading signal is set to 0. If the prediction direction are "DOWN", which represent that you can perform "buy" operation, the trading signal is set to 1. The trading signals for a stock are calculated in Table 1.

2011					
Time Series	Actural Value	Prodiction Value	Trend	Trading point	Trading signal
1	1.05	1.052782	up	sell	0
2	1.041	1.050341	up	sell	0
3	1.044	1.040494	down	buy	1
4	1.029	1.040907	up	sell	1
5	1.048	1.032553	down	buy	1
6	1.032	1.039286	up	sell	0
7	1.033	1.03237	down	buy	1
8	1.04	1.035205	up	sell	0
9	1.025	1.037839	up	sell	0
10	1.001	1.025077	down	buy	1
11	1.027	0.992739	down	buy	1
12	1.046	1.029434	down	buy	1
13	1.058	1.048558	down	buy	1
14	1.062	1.060122	down	buy	1
15	1.086	1.064865	down	buy	1

Table 1: The example of trading signals based on A2A in 2011

4 Experimental results

Several models with a subset of attributes as an input were built, which captures the movement based on daily close price. Different studies select their different

indicators as their input in order to create different models,The important idea to get successful stock market prediction is achieving the best results using the minimum required input data and the least complex stock market model. In our study, we just consider only influence of single indicator and multiple fundamental indicators on fitting rate and financial performance.

4.1 Description of the analyzed indices and scenarios

In the experimental section, we collect the datasets FTSE MIB 2011, FTSE MIB 2013, and FTSE MIB 2015. The FTSE MIB (Milano indice di Borsa) is the benchmark stock market index for the Borsa Italiana, which is the Italian national stock exchange. The index superseded the MIB-30 in September 2004. This index consists of the 40 most traded stock classes on the exchange.

Next, we will analyze the market situation for FTSE MIB 2011, FTSE MIB 2013, and FTSE MIB 2015. "Market conditions" refers to overbought (OB) or oversold (OS). Buying and selling is considered as alternative up and down stages that cause the price index to make a tortuous progress. First, buyers have overwhelmed sellers in the tide of price increases. Then the buying pressure has been exhausted and the market is described as overbought. Next, profit and selling begin as the sellers take control, and prices drop until the sales reach a point of exhaustion and the market is oversold. When these alternating pressures are roughly equal, the market moves sideways. Otherwise, the rising or falling trend tells us whether the buyer or the seller is dominant. By using certain indicators, the market's OB / OS status can be evaluated to understand when price reversal may occur.

Volume is a very good technical indicator for verifying price changes. The "real" price movements, in contrast to the temporary rebound movements, almost always occur with increasing trading volume. One of the keys that analysts use to distinguish between retracement and market reversal is the change in the overall signal trend. The fact is that reversal is usually accompanied by a large number of transactions, while retracement usually occurs at low volume. As a result, traders and analysts often track trading volume indicators, such as the Balance Volume (OBV) indicator or the Volume Price Trend indicator.

Momentum indicators such as the Average Directional Index (ADX) and Moving Average Convergence Departure (MACD) are also used to define and confirm bull and bear markets. Although indicators, like market prices, move up and down in the process of continuing trends, in the long run, the trend shows that the bull market continues to rise. In the long run, the trend shows that the continuous downward trend is a bear market.

The following figures are examples of displaying market conditions in different years based on price movement. Figure 10 shows that it is downtrend (bearish). Figure11 shows that it has an upward trend in 2013. And Figure 12 shows that there is also an oscillatory trend in 2015.

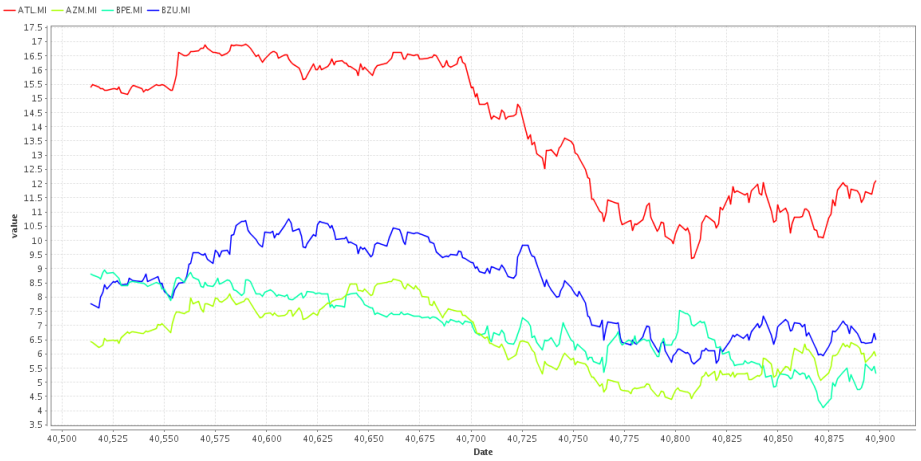


Figure 10: Market condition of stocks in 2011

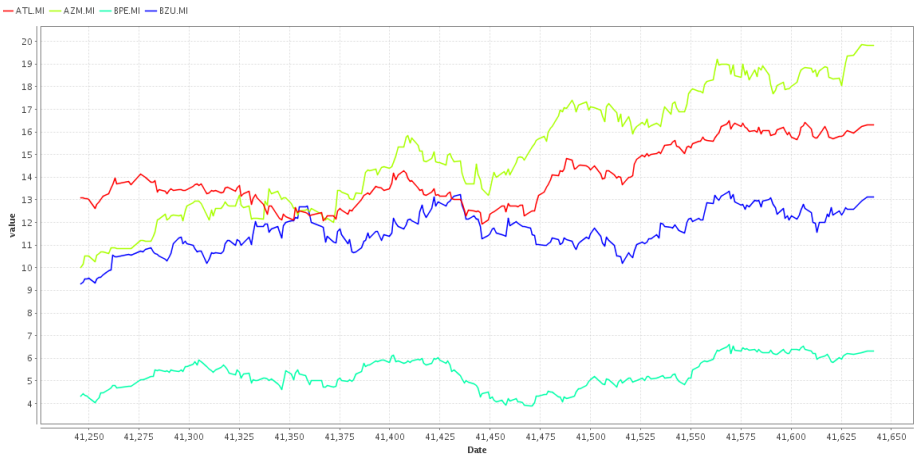


Figure 11: Market condition of stocks in 2013

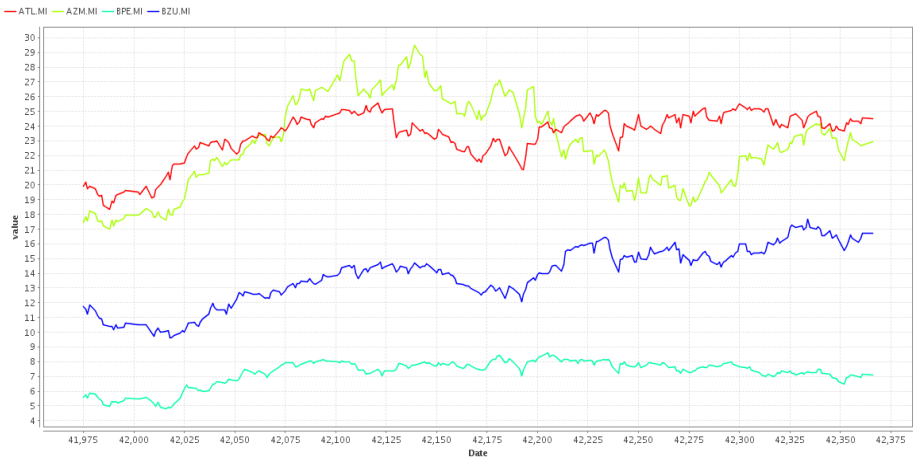


Figure 12: Market condition of stocks in 2015

4.2 Results of different techniques

In order to test the prediction performance of these different methods, we analyze and evaluate this experimental results from different points. First of all, we compute the prediction accuracy. In other words, we are only going to pay attention to the hit rate. This is just the percentage of achieving accurate prediction. After this actual direction and prediction direction are obtained, we use these values to evaluate the accuarcy of this model with different algorithms. The accuracy is calculated based on the following method. We attempted to allocate the actual direction and the forcasting direction into three kinds of situations: "up", "down" or "no action". If the actual direction and the prodiction direction are the same, the prediction is true. If the actual direction and the forecasted direction are different, the prodiction is false. The following table shows a clear calculation process.

direction	Forcasting direction		
actual	up	down	no action
up	true	false	false
down	false	true	false
no action	false	false	true

Table 2: showing the process of calculating accuracy

After having understood this process of correct prediction, I have collected the number of times that our strategy based on different algorithms can correctly predict the actual direction independently of the exact value of increase or decrease in the following table3 and table 4. Table 3 represents the results based on different algorithms with single variable input, Table 4 represents the results based on different algorithms with multiple variable input.

Model	2011		2013		2015	
	True	False	True	False	True	False
SVM	131	111	133	109	138	104
RepTree	155	87	155	87	154	88
Random Forest	122	120	121	121	125	117
NeuralNet	133	109	134	108	137	105
Linear Regression	123	119	126	116	133	109

Table 3: showing the average number of true and the average number of false for all stocks with different mode.

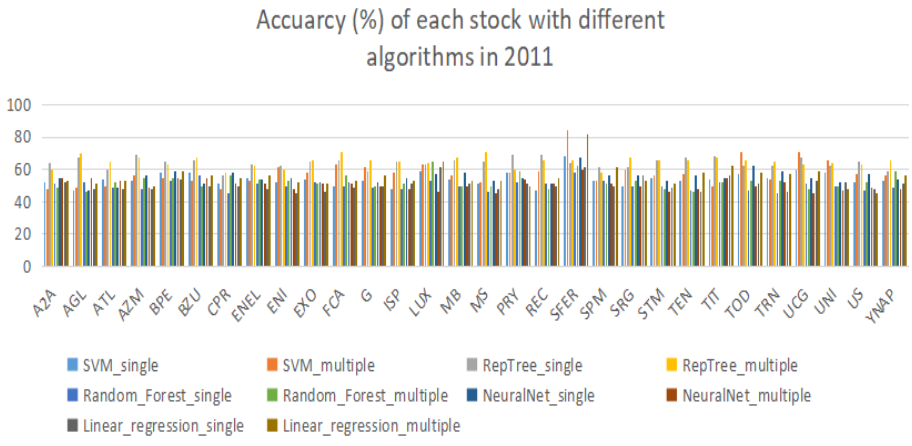
Model	2011		2013		2015	
	True	False	True	False	True	False
SVM	140	102	143	99	151	91
RepTree	157	85	158	84	157	85
Random Forest	127	115	125	117	126	116
NeuralNet	123	119	119	123	125	117
Linear Regression	134	108	133	109	144	98

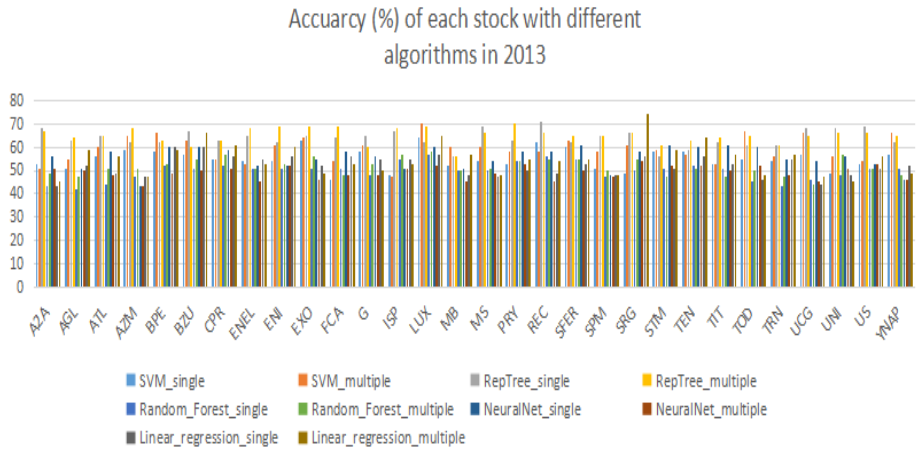
Table 4: showing the average number of true and the average number of false for all stocks with different mode.

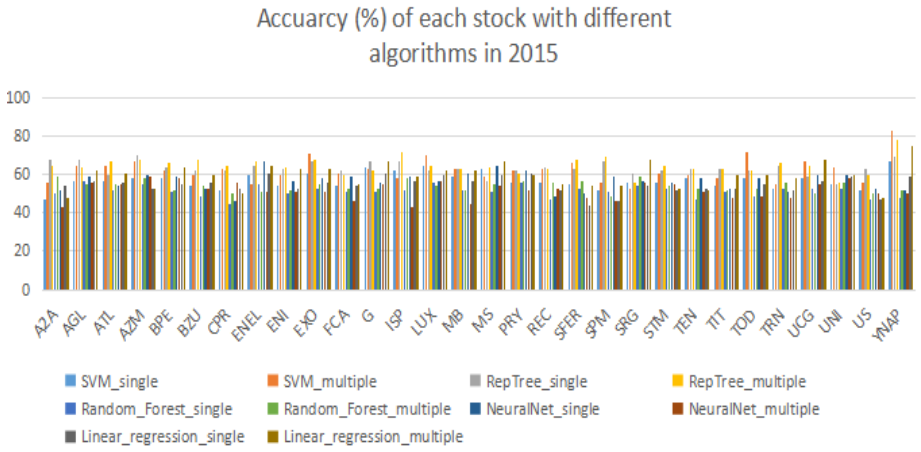
The equation to calculate the accuracy is as folloing:

accuracy=the number of true/ (number of true + number of false) * 100

I collect some results as following figures.







After analyzing the prediction accuracy of different algorithms for each stock, I calculated the average forecasting accuracy for all stocks, and these results are listed in the following table.

Model	2011		2013		2015	
	AAS	AAM	AAS	AAM	AAS	AAM
SVM	54.066	57.866	55.033	59.233	57.033	62.466
RepTree	64.1	64.833	63.933	65.067	63.5	64.767
Random Forest	50.167	52.567	49.834	51.7	51.5	52
NeuralNet	55.033	50.633		49.2	56.667	51.733
Linear Regression	50.633	55.467	52	54.967	54.9	59.533

Table 5: showing the average accuracy for all stocks with different models, AAS is average accuracy of univariate prediction, AAM is average accuracy of multivariate prediction.

For each technique, if the configuration settings are different, the prediction results are different. For the technique based on SVM algorithm, when I tune the paramater C, the prediction accuracy is changed. The larger the value of C the more violations of the hyperplane are permitted. The smaller the value of C, the more sensitive the algorithm is to the training data (higher variance and lower bias). The larger the value of C, the less sensitive the algorithm is to the training data (lower variance and higher bias). For the technique depends on Neural Net algorithm, I set different number of hidden layers. Then I tested the different layers in my experiment, and I found that when I increase the number of layers, it may lead to better accuracy, so I concluded that if we want to improve the prediction performance, we can consider to change the number of layers. I set random forest models with different forest sizes (the number of trees in the forest) and the depth of the trees, I found that when the growth is too large, the forest tends to be overconfigured without additional gain accuracy. So we can choose the appropriate parameters according to the actual situation. For the technique based on RepTree, I tune it with different number of folders. I found that when I increase the number of folders, the prediction accuracy is better. For Linear Regression models, I set the model by using different linear predictor functions. From these results, we can choose appropriate linear predictor functions in order to get better pridiction performance. The following tables show that the results of prediction accuracy with different parameters.

SVM	2011	2013	2015
C=0	54.066	55.033	57.033
C=0.5	56.233	56.866	58.633
C=0.8	57.233	57.4	57.567

Table 6: showing the prediction accuracy of same technical with different configure settings

NeuralNet	2011	2013	2015
hidden layers=1	52.667	52.133	49.7
hidden layers=3	55.9	55.067	56.667
hidden layers=5	56.033	55.067	56.667
default	55.033	55.5	56.667

Table 7: showing the prediction accuracy of same technical with different configure settings

Random Forest	2011	2013	2015
tree count=5	49.8	50.2	51.333
tree count=10	50.167	49.834	51.5
tree count=15	49.5	50.2	51.4

Table 8: showing the prediction accuracy of same technical with different configure settings

RepTree	2011	2013	2015
number of folders=1	63.567	63.9	64.467
number of folders=3	64.1	63.933	63.5
number of folders=5	64.5	65.367	64.967
number of folders=10	65.6	64.967	64.967

Table 9: showing the prediction accuracy of same technical with different configure settings

Linear Regression	2011	2013	2015
M5 prime	50.633	52	54.9
greedy	50.767	52.4	55.2
T-Test	50.833	52.167	55.867

Table 10: showing the prediction accuracy of same technical with different configure settings

Next, I will calculate the profit and the loss, annual balance and average annual balance for each stock. I calculate and evaluate them because the balance sheet and the profit or loss table are the most objective ways to look at a company's financial status. The profit or loss statement reports the financial performance of the company over a specific period of time and provides a summary of how the business generates revenue and expenses. This is an important document that investors used to determine the profitability of the business during the given period. But in this experiment, we calculate this profit and loss is different from ordinary economic calculations. I calculate them as the

following methods. When the actual direction and the forecasted direction are the same, the result is profit regardless of whether this direction is decreasing or rising. This profit is the percentage between the difference (the next actual closing value and the last closing value of window) and the previous value. If the actual direction and the forecasted direction are different, the result is the loss although this direction is "UP', the method of the loss calculated is the same to the way to get the profit above. The following table show the process of calculating the profit and the loss.

Actural direction	Forecasting direction	Result
UP	UP	profit
	DOWN	loss
	NO ACTION	loss
DOWN	UP	loss
	DOWN	profit
	NO ACTION	loss
NO ACTION	UP	loss
	DOWN	loss
	NO ACTION	profit

Table 11: showing the process of calculating the profit and the loss

The walk-forward procedure was executed based on different models (e.g.SVM, RepTree, Linear regression, NeuralNet, Random Forest and baseline strategy) generated for all combinations of training and test sample lengths defined above. These procedures are tested on the Italian stock time series of different years (e.g.2011,2013,2015). For each stock time series, the resulting prediction and financial performance of the proposed regression models were discussed and compared with two aspects. One point is univariate prediction, evaluation and comparison for different years. One perspective is the multivariate prediction, evaluation and comparison for different years.

In order to get more accurately evaluation of these results, we recommend to use as much data as possible for training and testing. In the case of the daily observations we used about 30 stocks, there were a total of 276 samples for each stock time series. These data are separated in chronological order. Next, we will analyze each stock based on different technologies from the perspective of prediction performance and financial performance.

The results of the prediction based on SVM algorithm is collected. The value column contains the percentage of the correct direction estimated, the percentage of the profit and loss, the percentage of the yearly balance and average yearly balance. The accuracy values in the table only consider the situation where the direction is accurately predicted. In 2011, the average percentage of predicted accuracy is 54.066, in2013, the value is 55.033, in 2015, the value is 57.033. The calculation of these values is only considered univariate forecasting. However, from the point of view of multivariated prediction, the average percentage of the accuracy prediction has increased with the same dataset. The value increases

to 57.866 in 2011, and the value has changed from 55.033 to 59.233 in 2013, however in 2015, the value changed to 62.466. From these changes, we can get that multivariable prediction is more precise than single variable prediction. You can see these values in Table 5.

We have made an assessment of the model based on SVM from the perspective of predictive performance, then we will analyze the financial performance below. According to [7], financial performance is a more relevant indicator for assessing financial market forecasting. Trading simulation was made by using trading signals decribed above. I analyzed the three years historical data from Italian stocks. The prediction based on the SVM model for each walk-forward procedure, I calculated their profits and losses. Then I also calculated their annual balance for each year. According to these results, I calculated the average annual balance of each stock. Table 12 shows the related financial performance statistics after the SVM model prediction.

	2011		2013		2015	
Stock	Sbalance	Mbalance	Sbalance	Mbalance	Sbalance	Mbalance
A2A	0.096	0.248	0.285	0.259	-0.046	0.297
AGL	-0.024	0.128	0.036	0.470	0.438	0.735
ATL	0.021	0.182	0.227	0.493	0.285	0.545
AZM	0.364	0.603	0.321	0.710	0.331	0.857
BPE	0.532	0.645	0.614	1.165	0.550	0.901
BZU	0.283	0.522	0.289	0.615	0.427	0.580
CPR	0.060	0.283	0.074	0.264	0.113	0.629
ENEL	0.258	0.380	0.295	0.198	0.332	0.440
ENI	0.058	0.535	0.171	0.313	0.120	0.435
EXO	0.054	0.910	0.423	0.690	0.291	0.737
FCA	0.176	1.105	-0.195	0.454	0.194	0.590
G	0.110	0.586	0.244	0.515	0.316	0.416
ISP	0.192	1.065	0.058	0.185	0.420	0.609
LUX	0.259	0.443	0.321	0.449	0.540	0.743
MB	0.300	0.401	0.254	0.711	0.310	0.749
MS	0.074	0.148	0.365	0.765	0.554	0.662
PRY	0.273	0.720	0.062	0.441	0.218	0.571
REC	0.081	0.519	0.309	0.351	0.264	0.711
SFER	1.589	2.330	0.376	0.609	0.332	0.855
SPM	0.234	0.475	-0.125	0.471	0.310	0.545
SRG	0.154	0.303	0.035	0.386	0.177	0.309
STM	0.311	0.435	0.377	0.397	0.229	0.742
TEN	0.185	0.557	0.302	0.299	0.526	0.677
TIT	0.117	0.200	0.155	0.156	0.220	0.674
TOD	0.295	0.924	0.107	0.415	0.382	0.790
TRN	0.200	0.433	0.111	0.173	0.112	0.298
UCG	0.445	1.205	0.400	0.643	0.374	0.751
UNI	0.414	0.847	0.195	0.562	10.728	11.207
US	0.514	0.786	0.293	0.294	0.222	0.323
YNAP	0.159	0.652	0.411	0.749	1.083	1.886

Table 12: showing the average annual balance for all stocks with SVM, Sbalance is average annual balance for univariate prediction, Mbalance is average annual balance for multivariate prediction

The results of the prediction ferformance based on RepTree algorithm is shown in Table 5. From this table,we can see that the prediction of this accuracy is relatively high. For univariate predictions, this range fluctuates between 59 and 70. Multivariate prediction accuracy is a little higher than univariate prediction. At the same time, we can get the average of prediction accuracy is 64.1, 63.933, 63.5 in 2011, 2013 and 2015 for univariate prediction. For multivariate predictions, the average of prediction accuracy is 64.833, 65.067 and 64.764. From these results, We can get the accuracy of the prediction

model based on RepTree algorithm is higher than that based on SVM.

With the same calculation method, the average annual return has an increasing trend based on RepTree algorithm than that based on the SVM algorithm. For example, the average annual return is 0.675 based on RepTree algorithm in 2011 for A2Astock, but the average annual return is just 0.096 based on SVM algorithm in 2011. There are a lot of data showing this feature, and I collected these data in Table 13.

	2011		2013		2015	
Stock	Sbalance	Mbalance	Sbalance	Mbalance	Sbalance	Mbalance
A2A	0.675	0.451	0.964	0.832	0.695	0.610
AGL	0.669	0.719	0.677	0.741	0.741	0.583
ATL	0.490	0.775	0.571	0.551	0.409	0.570
AZM	1.020	1.095	0.643	0.859	0.896	0.847
BPE	1.165	1.004	0.879	0.913	0.928	0.979
BZU	1.040	1.109	0.751	0.550	0.765	1.046
CPR	0.332	0.351	0.468	0.509	0.585	0.641
ENEL	0.685	0.641	0.665	0.777	0.677	0.725
ENI	0.644	0.560	0.445	0.621	0.680	0.660
EXO	0.944	1.062	0.600	0.702	0.621	0.596
FCA	1.253	1.360	0.763	1.085	0.638	0.746
G	0.423	0.817	0.577	0.494	0.547	0.427
ISP	1.407	1.358	0.818	0.894	0.904	1.023
LUX	0.544	0.588	0.392	0.599	0.720	0.797
MB	0.797	0.837	0.648	0.525	0.831	0.798
MS	0.671	1.068	1.271	1.182	0.655	0.801
PRY	1.125	0.757	0.507	0.677	0.588	0.621
REC	0.779	0.693	0.596	0.530	0.637	0.639
SFER	1.485	1.496	0.610	0.648	0.657	0.986
SPM	0.705	0.685	0.703	0.688	1.102	1.172
SRG	0.435	0.450	0.371	0.428	0.397	0.315
STM	0.991	1.133	0.435	0.600	0.689	0.814
TEN	0.691	0.781	0.402	0.486	0.780	0.804
TIT	0.782	0.734	0.880	1.047	0.756	0.829
TOD	0.725	0.827	0.506	0.509	0.429	0.418
TRN	0.500	0.590	0.306	0.323	0.440	0.394
UCG	1.235	1.215	0.912	0.820	0.544	0.806
UNI	0.735	0.940	1.049	0.918	12.489	12.527
US	1.394	1.038	0.860	0.794	0.586	0.521
YNAP	0.677	1.012	0.666	0.750	1.482	1.843

Table 13: showing the average annual balance for all stocks with RepTree, Sbalance is average annual balance for univariate prediction, Mbalance is average annual balance for multivariate prediction

We can get these resullts of these prediction performance from Table 5. The correct forecast is about half the total forecasting. In 2011, the forecasting accuracy was 50.167 and 52.567, respectively. But in 2013, this prediction accuracy decreased slightly, and this decrease was less than 1. In 2015, the accuracy was 51.5 and 52, respectively, which was similar to that in 2011.

As the prediction accuracy of the model decreases, we get a slight uncertainty in the prediction of the average annual return. So the results of these average annual returns have aome fluctuations for each stock. We can compare and analyze these results from these data in Table 14.

	2011		2013		2015	
Stock	Sbalance	Mbalance	Sbalance	Mbalance	Sbalance	Mbalance
A2A	0.037	-0.066	-0.100	0.255	0.198	0.293
AGL	0.082	0.000	-0.242	-0.122	0.300	0.280
ATL	-0.040	-0.063	0.018	0.132	0.100	0.221
AZM	0.078	0.409	-0.010	0.090	0.305	0.511
BPE	0.123	0.304	0.334	0.332	0.377	0.382
BZU	0.221	0.331	0.081	0.200	0.044	0.236
CPR	0.131	0.226	0.157	0.217	-0.045	0.178
ENEL	0.165	0.090	0.071	0.083	0.231	0.102
ENI	0.032	0.128	0.147	0.170	0.088	0.149
EXO	0.220	-0.002	0.064	0.208	0.184	0.119
FCA	0.083	0.449	0.356	0.024	0.312	0.238
G	0.147	0.039	0.111	0.193	0.133	0.093
ISP	0.182	0.081	0.063	0.277	0.309	0.412
LUX	0.107	0.385	0.245	0.248	0.436	0.454
MB	-0.039	-0.014	0.072	0.097	0.230	0.230
MS	-0.168	0.058	-0.196	0.149	0.250	0.338
PRY	0.201	0.189	0.133	0.150	0.337	0.250
REC	0.000	0.068	0.136	0.146	-0.064	0.296
SFER	0.406	0.745	0.225	0.225	0.158	0.253
SPM	0.140	0.098	0.118	-0.151	0.184	-0.004
SRG	0.110	0.092	0.070	0.139	0.144	0.210
STM	0.236	-0.016	0.208	0.085	0.186	0.220
TEN	0.044	-0.038	0.142	0.109	0.026	0.270
TIT	0.173	0.107	0.125	-0.045	0.297	0.179
TOD	0.165	0.267	0.053	0.105	0.121	0.198
TRN	0.140	0.112	-0.005	0.037	0.125	0.266
UCG	-0.125	-0.172	0.082	-0.046	0.191	0.034
UNI	-0.134	-0.105	0.217	0.463	1.858	-3.027
US	-0.447	-0.145	0.168	0.125	0.057	0.124
YNAP	-0.017	0.271	-0.167	-0.261	0.176	0.481

Table 14: showing the average annual balance for all stocks with Random Forest, Sbalance is average annual balance for univariate prediction, Mbalance is average annual balance for multivariate prediction

The results of the prediction based on NeuralNet algorithm show that the prediction is less accurate. The percentage of the accuracy is about 55, just half of the total forecasting. Although its prediction accuracy is not as high as that of the RepTree model, it is higher than that of the Random Forest model, and it is similar to the prediction accuracy of the model based on SVM. Another advantage is that the Neural net outputs these results very fast by using the sliding validation method, this is a trade off between time consuming and precision. The detailed data collected in Table 15.

	2011		2013		2015	
Stock	Sbalance	Mbalance	Sbalance	Mbalance	Sbalance	Mbalance
A2A	0.277	0.143	0.410	0.177	0.254	-0.087
AGL	0.021	0.142	0.293	-0.050	0.492	0.368
ATL	0.122	0.012	0.397	0.084	0.234	0.230
AZM	0.281	0.097	-0.184	-0.196	0.738	0.434
BPE	0.879	0.190	0.572	-0.020	0.784	0.417
BZU	0.464	0.237	0.637	-0.038	0.262	0.091
CPR	0.479	0.030	0.296	0.118	0.009	0.277
ENEL	0.411	0.051	0.030	-0.061	0.708	0.134
ENI	0.348	0.067	0.221	0.093	0.376	0.147
EXO	0.405	-0.061	0.323	0.026	0.383	0.079
FCA	0.305	0.158	0.646	-0.176	0.626	-0.180
G	0.188	0.101	0.361	-0.010	0.281	0.170
ISP	0.692	0.188	0.184	0.010	0.604	-0.123
LUX	0.448	-0.083	0.374	0.016	0.503	0.319
MB	0.559	0.043	0.205	-0.070	0.637	-0.135
MS	0.178	-0.191	0.673	-0.016	0.779	0.257
PRY	0.392	0.170	0.449	0.173	0.604	0.220
REC	0.372	0.071	0.237	0.000	-0.165	0.255
SFER	1.741	0.570	0.453	0.014	0.296	-0.030
SPM	0.348	0.175	0.385	0.008	0.774	-0.089
SRG	0.266	0.104	0.217	0.083	0.360	0.202
STM	0.276	0.062	0.655	0.005	0.447	0.160
TEN	0.286	0.116	0.406	0.103	0.612	0.011
TIT	0.369	0.204	0.770	0.115	0.370	0.003
TOD	0.825	0.148	0.308	0.002	0.385	-0.065
TRN	0.350	0.107	0.222	0.023	0.134	0.083
UCG	0.442	-0.239	0.161	0.038	0.783	0.202
UNI	0.358	-0.297	0.243	0.148	12.733	14.022
US	0.780	-0.032	0.396	0.208	0.224	0.125
YNAP	0.517	-0.030	-0.344	-0.227	0.475	-0.293

Table 15: showing the average annual balance for all stocks with NeuralNet, Sbalance is average annual balance for univariate prediction, Mbalance is average annual balance for multivariate prediction

The results show that the proposed approach greatly increased prediction accuracy and the number of undecided states helps reduce the number of mispredictions. It is also worth noting that the increase of parameters in models reduces the undecided state and improves the number of correct predictions. Although we do not immediately benefit from the improvement in forecast quality, it should be questioned whether the simple portfolio management strategy is suitable for future forecasting. We can get the accuracy is range from about 50 percentage to 70 percentage with single variable for each stock. With the

increase of the parameters, this prediction is more accurate. From table 10, we can get the average accuracy is range from about 50 percentage to 60 percentage. In addition, we can analyze and evaluate financial performance according to Table 16.

	2011		2013		2015	
Stock	Sbalance	Mbalance	Sbalance	Mbalance	Sbalance	Mbalance
A2A	-0.039	-0.021	-0.118	-0.036	0.235	0.204
AGL	0.021	0.163	-0.001	0.549	0.448	0.705
ATL	-0.133	-0.012	0.025	0.248	0.332	0.553
AZM	-0.149	0.317	0.052	0.039	0.206	0.391
BPE	0.262	0.663	0.711	0.661	0.581	0.687
BZU	0.129	0.276	0.494	0.870	0.396	0.549
CPR	0.040	0.323	0.236	0.405	0.240	0.107
ENEL	-0.102	0.348	0.169	0.327	0.460	0.841
ENI	-0.063	0.236	0.317	0.404	0.220	0.593
EXO	-0.119	0.138	0.110	0.146	0.277	0.468
FCA	-0.088	0.218	0.442	0.320	0.254	0.287
G	-0.048	0.351	0.355	0.117	0.399	0.625
ISP	0.044	0.409	0.294	0.016	0.388	0.569
LUX	0.338	0.708	0.209	0.420	0.611	0.682
MB	0.096	0.187	-0.065	0.647	0.471	0.399
MS	-0.173	0.238	-0.105	-0.154	0.620	0.988
PRY	-0.163	0.154	0.203	0.325	0.443	0.358
REC	0.183	0.378	0.013	0.137	0.175	0.286
SFER	1.513	2.836	0.201	0.322	0.008	0.327
SPM	0.263	0.671	0.084	0.516	0.057	0.328
SRG	0.237	0.231	0.253	0.604	0.209	0.614
STM	-0.096	0.034	0.291	0.523	0.194	0.321
TEN	-0.066	0.410	0.254	0.522	0.418	0.341
TIT	0.228	0.561	0.296	0.601	0.365	0.567
TOD	0.120	0.458	-0.023	0.027	0.337	0.613
TRN	0.007	0.403	0.177	0.210	0.190	0.254
UCG	0.067	0.713	-0.029	0.201	0.515	0.982
UNI	-0.077	0.017	0.020	0.010	14.034	14.041
US	-0.282	-0.074	0.152	0.342	0.161	0.191
YNAP	0.085	0.345	0.101	0.035	1.065	1.956

Table 16: showing the average annual balance for all stocks with Linear Regression, Sbalance is average annual balance for univariate prediction, Mbalance is average annual balance for multivariate prediction

The strategy is operated by building a sliding window, then the window will advance one day each time. For each sliding window, if you want to predict the trend of each stock at a specific point, you only need to evaluate the trend at

the previous time point,then your forecast will be same trend with the previous time point. But the prediction accuracy made by this strategy is relatively low, which is less than 50 percentage.

Though comparing the prediction results of different models above, the prediction performance of this RepTree model is better than other models. In addition, input variables as an important factor also have a great effect for the prediction. We input different variables to predict the dependent variables. We find that if you provide more input variables, then the model can predict the dependent variable depend on more information. So the accuracy of this prediction is higher. In addition, we can analyze and evaluate financial performance according to Table 17.

Stock	2011 balance	2013 balance	2015 balance
A2A	-0.111	0.105	0.012
AGL	-0.055	0.099	-0.31
ATL	-0.028	-0.100	-0.105
AZM	0.099	-0.371	-0.038
BPE	-0.042	-0.041	0.044
BZU	-0.263	0.109	-0.034
CPR	-0.14	-0.092	-0.084
ENEL	-0.139	-0.130	-0.155
ENI	0.083	-0.094	-0.064
EXO	-0.009	-0.024	-0.238
FCA	0.032	0	0.152
G	-0.151	-0.166	0.064
ISP	-0.111	-0.126	-0.183
LUX	-0.225	-0.152	-0.055
MB	-0.065	0.137	-0.005
MS	-0.161	0.009	-0.012
PRY	-0.159	0.008	-0.192
REC	-0.079	-0.105	-0.016
SFER	-0.131	-0.024	-0.271
SPM	0.027	-0.107	-0.109
SRG	-0.114	0.051	0.154
STM	0.023	-0.008	0.182
TEN	-0.076	-0.001	-0.189
TIT	0.021	-0.085	-0.099
TOD	-0.271	-0.13	-0.085
TRN	-0.184	-0.011	-0.246
UCG	0.012	0.099	-4.792
UNI	0.002	-0.166	0.073
US	0.020	0.039	-0.033
YNAP	-0.048	-0.04	-0.228

Table 17: showing the average annual balance for all stocks with baseline strategy

After we compute the gross average balance over all the stocks, we need to calculate the net balance over all the stocks. The method of the net balance is that the gross average balance subtract 26 percentage of taxes and 0.12 percentage of tranction fees. The following table show these results.

Model	2011 balance	2011 netbalance	2013 balance	2013 netbalance	2015 balance	2015 netbalance
SVM	0.207	0.173	0.226	0.190	0.678	0.569
RepTree	0.834	0.699	0.665	0.557	1.096	0.919
Random Forest	0.075	0.063	0.089	0.075	0.242	0.203
NeuralNet	0.446	0.374	0.333	0.279	0.857	0.719
Linear Regression	0.068	0.057	0.171	0.143	0.810	0.680
baseline	-0.066	-0.055	-0.047	-0.039	-0.225	-0.189

Table 18: showing the gross average balance and net nalance for all stocks with different models with single variable input.

Model	2011 balance	2011 netbalance	2013 balance	2013 netbalance	2015 balance	2015 netbalance
SVM	0.619	0.519	0.473	0.397	1.008	0.846
RepTree	0.871	0.731	0.702	0.589	1.12	0.943
Random Forest	0.128	0.107	0.121	0.102	0.133	0.112
NeuralNet	0.075	0.063	0.019	0.016	0.573	0.481
Linear Regression	0.389	0.326	0.312	0.262	0.994	0.834

Table 19: showing the gross average balance and net nalance for all stocks with different models with multiple variable input.

For each stock index, we will display the annual forecast data of the five models and the corresponding actual data in the following chart. Due to the input variables of each model are different, the results produced also change. As can be seen from the Figure 13 below, the price of this stock has some fluctuations at the beginning, and in the later period, the price has a downward trend. It can be seen from Figure 13 and Figure 16 that the trend of this change is the same, but the variation trend of the multivariable prediction is less fluctuation. The Figure 14 and Figure 17 show that the stock change trend is on the upward trend. Figure 15 and Figure 18 also show that the stock price is an upward trend, but this flutuation is very amall in different forecasting methods.

The following figures is just an example, I analyzed the forecasting results for all the stocks with different input parameters according to different algorithms. From these results, I got the accuracy of multivariate prediction is better than the accuracy of univariate prediction. The more perameters you input, the more data is referenced when the model predicts the results, so the predicted results

are more accurate.

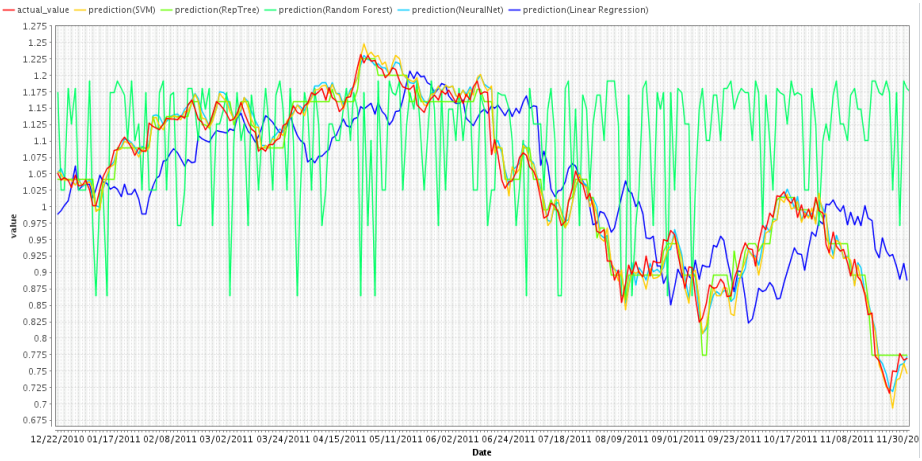


Figure 13: Daily stock price prediction of A2A stock with single variable in 2011

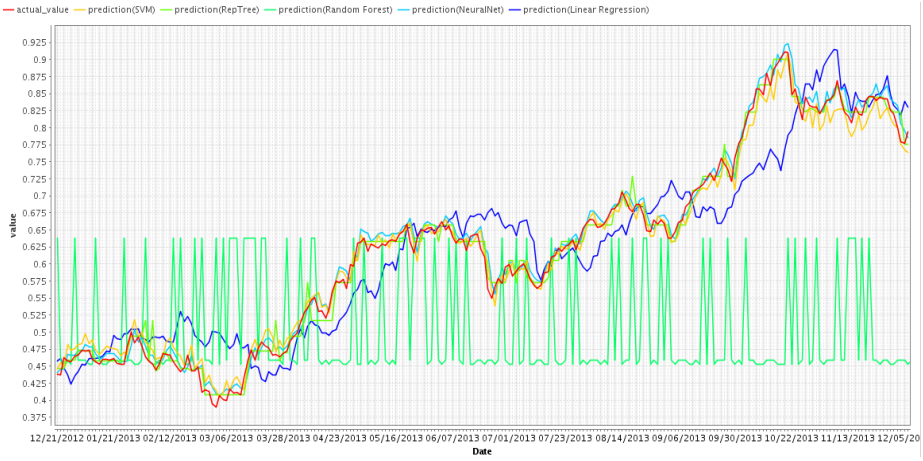


Figure 14: Daily stock price prediction of A2A stock with single variable in 2013

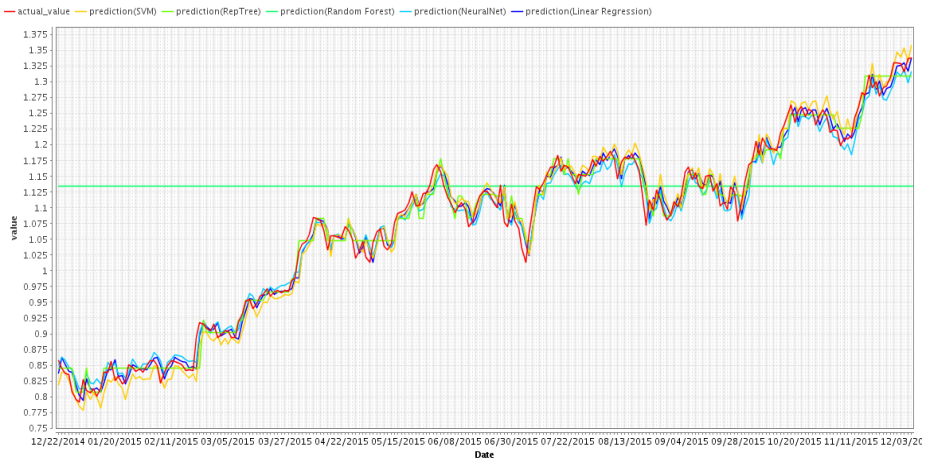


Figure 15: Daily stock price prediction of A2A stock with single variable in 2015

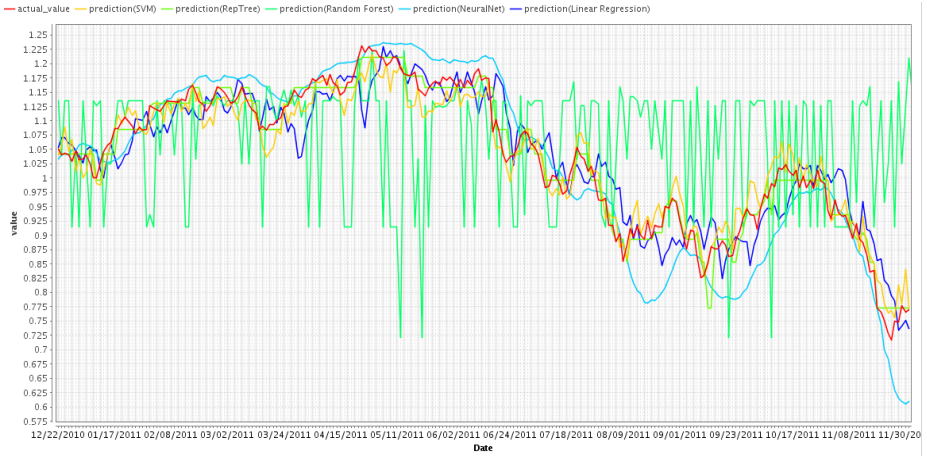


Figure 16: Daily stock price prediction of A2A stock with multiple variables in 2011

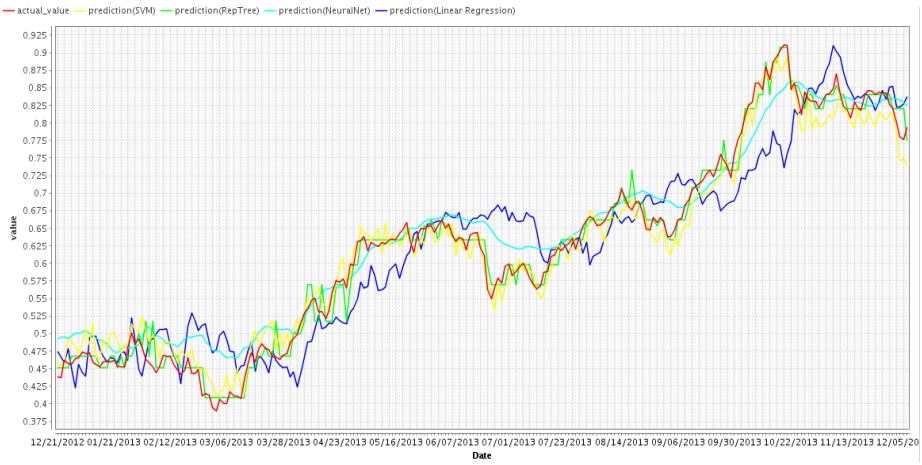


Figure 17: Daily stock price prediction of A2A stock with multiple variables in 2013

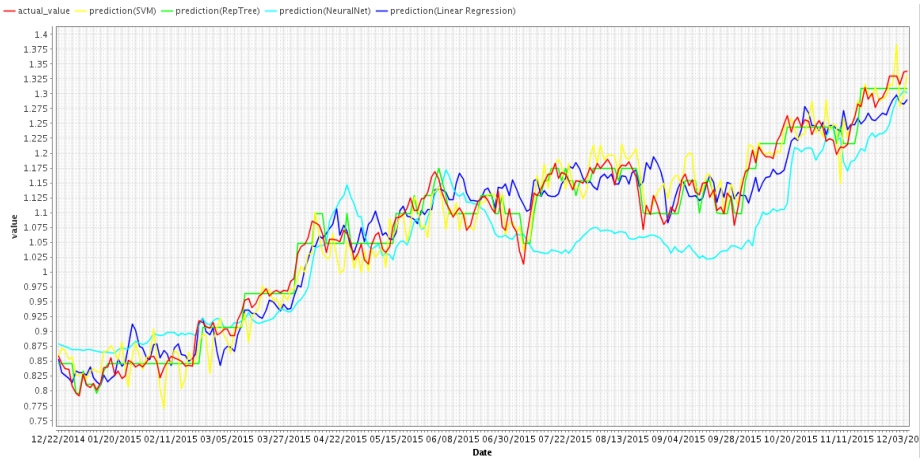


Figure 18: Daily stock price prediction of A2A stock with multiple variables in 2015

5 Conclusions and future works

Financial market is a special system that generate a lot of information, these information is messy (e.g.financial statements records, traffics, flight ways, news, rumours, economic tables, investors records and so on). So it is no doubt that looking for methods able to manage, classify and analyze the available data, infer reliable observations and gain believable results is becoming more and more important and urgent. However, financial market forecasting based on

these data is considered a very difficult task, as they generate large amounts of uncertain information and incomplete information with ongoing evolution of financial markets. Any wrong decision and prediction can have potentially serious consequences for the individuals economic, institutions, financial industry and nations. And financial markets (especially stock markets) are considered as the high return investment fields, but this area is vulnerable to uncertain information and data volatility. So stock market prediction tries to reduce this uncertainty and consequently the risk. Many researchers are interested in financial time series predicting, especially stock price movements. And many methods have been proposed in this area, ranging from the simple to the very sophisticated technology.

A quantitative intraday trading strategy has proved to be sufficiently efficient for the financial time series forecasting system. In this strategy, I use a veriety of regression algorithms used to create models to predict the future results. For each model, we use different stock historical data to test them. An algorithm corresponds to a model. Due to different algorithms, we have different configurations for each model. Their purpose is to get the best prediction results.

The stock market is a complex and noisy environment, which is dominated by uncertainty and high volatility. Classical linear analysis is inefficient to capture the complex dynamics of stock market time series. We have proposed several models based on regression algorithms that can detect the basic mechanism of the stock market. The implemented financial forecasting system has successfully performed some time series testes. The performance of the system depends mainly on the quality of the data. The quality of these data is used together with training purposes in different models. Therefor, data selection, collection and data preprocessing are important components of the forecasting process.

I use different algorithms to create models, to analyze and forecast future data. Each model integrates a variety of different components and parametes, which must be optimally selected and set. The various financial time series involved in the experiment are publicly available. For training and testing purposes, these are isolated in chronological order. In the experiment, I use the sliding window to train these stock data to form a datasets. Then calling different algorithms to train them, finally, loading these data that needs to predict into this model to achieve the purpose of testing. I made daily forecasts, which is easier to predict the future value and get the forecast trend. That is, the direction of the price change. In this way, we get satisfactory results.

Although prediction methods based on regression models can produce more efficient prediction systems. However, several pitfalls must be avoided. Moreover, the result should be explained more carefully. Because the dynamics of the stock market are very complex. Professional forecasters should use the forecasting system as a reminder, and the final decision should also depend on other relevant information.

As a future job, developing long-term forecasting will be interesting through the regression algorithm. Increasing the forecasting range, which is not just limited to daily forecasting. In another important future work, it aims to apply

multiple regression prediction algorithms, which are integrated into a system, in this way it can achieve the possibility of multiple predictions, and prediction accuracy will be higher, then the prediction is more accurate.

References

[1] A Bocharov. Looking for short term signals in stock market data. *WIT Transactions on Information and Communication Technologies*, 41:157–166, 2008.

[2] Lorant Bodis. Financial time series forecasting using artificial neural networks. *Mestrado–Babeş-Bolyai University*, 2004.

[3] Pei-Chann Chang, Chen-Hao Liu, Chin-Yuan Fan, Jun-Lin Lin, and Chih-Ming Lai. An ensemble of neural networks for stock trading decision making. In *International Conference on Intelligent Computing*, pages 1–10. Springer, 2009.

[4] Yingjun Chen and Yongtao Hao. A feature weighted support vector machine and k-nearest neighbor algorithm for stock market indices prediction. *Expert Systems with Applications*, 80:340–355, 2017.

[5] Wen-Chyuan Chiang, David Enke, Tong Wu, and Renzhong Wang. An adaptive stock index trading decision support system. *Expert Systems with Applications*, 59:195–207, 2016.

[6] Eugene F Fama. Random walks in stock market prices. *Financial analysts journal*, 51(1):75–80, 1995.

[7] Emile Fiesler and Russell Beale. *Handbook of neural computation*. CRC Press, 1996.

[8] Farhad Soleimanian Gharehchopogh, Tahmineh Haddadi Bonab, and Seyyed Reza Khaze. A linear regression approach to prediction of stock market trading volume: a case study. *International Journal of Managing Value and Supply Chains*, 4(3):25, 2013.

[9] Marija Gorenc Novak and Dejan Velušček. Prediction of stock price movement based on daily high prices. *Quantitative Finance*, 16(5):793–826, 2016.

[10] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[11] Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.

[12] SK Jayanthi and S Sasikala. Reptree classifier for identifying link spam in web search engines. *IJSC*, 3(2):498–505, 2013.

[13] Yakup Kara, Melek Acar Boyacioglu, and Ömer Kaan Baykan. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert systems with Applications*, 38(5):5311–5319, 2011.

[14] Youngmin Kim, Wonbin Ahn, Kyong Joo Oh, and David Enke. An intelligent hybrid trading system for discovering trading rules for the futures market using rough sets and genetic algorithms. *Applied Soft Computing*, 55:127–140, 2017.

[15] Yung-Keun Kwon and Byung-Ro Moon. A hybrid neurogenetic approach for stock forecasting. *IEEE transactions on neural networks*, 18(3):851–864, 2007.

[16] Qing Li, Yuanzhu Chen, Li Ling Jiang, Ping Li, and Hsinchun Chen. A tensor-based information framework for predicting the stock market. *ACM Transactions on Information Systems (TOIS)*, 34(2):11, 2016.

[17] Mohamed M Mostafa. Forecasting stock exchange movements using neural networks: Empirical evidence from kuwait. *Expert Systems with Applications*, 37(9):6302–6309, 2010.

[18] Randall C O'Reilly and Yuko Munakata. *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT press, 2000.

[19] Ms D Preetha and Mrs K Mythili. Kenerl based svm classification for financial news. *International Journal*, 2(11), 2013.

[20] J. Ross Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

[21] Chih-Fong Tsai, Yuah-Chiao Lin, David C Yen, and Yan-Min Chen. Predicting stock returns by classifier ensembles. *Applied Soft Computing*, 11(2):2452–2459, 2011.

[22] Jian-Zhou Wang, Ju-Jie Wang, Zhe-George Zhang, and Shu-Po Guo. Forecasting stock indices with back propagation neural network. *Expert Systems with Applications*, 38(11):14346–14355, 2011.

[23] Lili Wang, Zitian Wang, Shuai Zhao, and Shaohua Tan. Stock market trend prediction using dynamical bayesian factor graph. *Expert Systems with Applications*, 42(15-16):6267–6275, 2015.

[24] Wikipdia. Artificial neural network.

[25] Wikipdia. Ftse mib.

[26] Kamil Żbikowski. Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. *Expert Systems with Applications*, 42(4):1797–1805, 2015.