Machine Learning Applications in Finance

Irmak Akkuzuluoglu
irmakakkuzulu@gmail.com

Mehmet Cicek
cmehmet@gmail.com

Baris Duru
durubaris7@gmail.com

Ismail Hakki Karaman
ihkaraman0@gmail.com

***Abstract*— Artificial Intelligence and Machine Learning applications have started to be used in many areas including finance. Use of the information-based methods along with new financial instruments have the potential to revolutionize decision making process. This study is going to investigate use of some Machine Learning algorithms to give insights about the possible price movement of stocks of some major energy and technology companies with some commodities in the future.**

Keywords—machine learning, finance, forex, trading

I. Introduction

Finance has always been one of the first business practices integrating new developments of academia. The rise of Artificial Intelligence and Machine Learning applications has not been missed by the finance world [1]. Use of the information-based methods along with new financial instruments starts a new era for investment decisions. Modern financial technics require computing power since 1950's. Modern portfolio theory of Harry Markowitz, published in 1952 by the Journal of Finance (also earned a Nobel prize to its author), requires calculation of correlation of all possible assets using historical data to have the optimum portfolio with the highest return and low level of risk [2]. This optimization algorithm was one of the first applications which needs computing power. In more modern world of finance, applications of the information-based methods are not limited to optimization of available data but also generating data and find best ways to make decision and also executing in real-time and do it very fast.

Fundamental analysis and technical analysis are the major ways to investigate historical data to give decisions. While fundamental analysis deals more with "balance sheet" of a company and compares it with other companies and industry trends, technical analysis only takes price and volume data into account. As a result, technical analysis has a wider area of application such as forex, futures, cryptocurrencies and commodities like oil and corn markets.

Technical analysis uses indicators, which are calculated moving rates, and which are argued to be able to signal future price movements. The problem with the indicators is that there are many indicators and each of them can work very well in some cases however, there would be times different indicators suggest opposite positions to take. Moreover, there are many parameters in each of the indicators to use which needs some tuning. Which indicator combination to use with which parameters in changing financial position will be investigated in this study and only technical analysis will be used.

This study investigates use of three classification algorithms (Decision Tree, Random Forest, Support Vector Machine), as well as Long Short Term Memory algorithm for financial time series to give insights about possible price movements of 4 technology companies (Amazon, Apple, Google, Microsoft), 4 energy companies (BP, ConocoPhillips, Chevron, Exxon) and 3 commodities (Gold, Silver, Copper). 6 different technical indicators (MACD, RSI, Stochastic Oscillator, William Rate Percentage, PPO, OBV) are used for each time series. Hypothetical annual compounded profit for each time series and algorithm pair is calculated and performance of different algorithms are compared. 3 time series (Amazon, Exxon, BP) are also back-tested for last 10 years and results are given compared to Buy&Hold strategy.

II. Literature Review

While many of the machine learning (ML) algorithms have been in the literature for a while, it is data-centric business models and cheap/widely-available fast computers which makes finance world to adapt ML techniques into trading. This triggers a new literature to flourish focusing ML applications in trading.

In this study we go over a literature about ML algorithms for FOREX, stock exchange markets to cover a wide range of interest to observe best practices and some flaws present in current practices.

Authors of [3] uses kNN to analyze stock market. First, they use the eigenvalues and eigenvectors of the financial correlation matrix to analyze the structure of the network. They have found that the degree is related to the average correlation coefficient, and that it has a relationship between the components of the eigenvector corresponding to the maximum eigenvalue. Existing research to confirms that the community structure of the kNN network can be used to cluster financial time series. They construct a correlation coefficient matrix to study the correlation of time series. They use Pearson's correlation coefficient, which is the most commonly used coefficient in financial analysis. Moreover, they mention empirical studies based on financial markets, which in three countries showed that there is a high correlation between the community structure and dimensions. Therefore, this study shows that the structure of the financial kNN network is related to the properties of the correlation matrix, and it extracts a meaningful correlation structure. After applying modularity as a variable to characterize the community structure, they found that there is a high negative correlation between the modularity series and the dimension series.

Conventional trading mostly bases upon technical analysis. Technical analysis in basic terms uses calculated time series which accompany the real time series of price movements. Authors of [4], uses technical indicator set of MA15, MACD26, K14, D3, RSI14 and WR14 and used these indicators, as well as, market closing price to feed extreme learning machine (ELM) algorithm. Output of the algorithm is considered a trading signal in form of buy, sell or hold. Another study using Neural Networks to predict market movement is [5]. In this study, authors use two large datasets of 424 S&P 500 index component stocks (SPICS) and 185 CSI 300 index component stocks (CSICS) between 2010 to 2017. They consider both with and without transaction cost and they compare daily trading performance of ML and DNN algorithms focusing on the availability of transaction cost.

According to results obtained traditional ML algorithms show better performance in most of the directional evaluation indicators with no transaction costs. When transactions costs are taken into consideration, DNN models have better performance and ML algorithms are sensitive to the changes in transaction costs.

Authors of [6] presents a stock trading method by combining the filter rule and the decision tree technique. The filter rule, having been widely used by investors, is used to generate candidate trading points. The idea of the filter rule is to buy when the stock price rises k% above its past local low and sell when it falls k% from its past local high. This research aimed to improve the filter rule by considering both the past and the future information in clustering the trading points. These points found by the filter rule are subsequently clustered and screened by the application of a decision tree algorithm C4.5. The criteria for the clustering involve four variables, three of which are associated with the past information. The remaining variable is associated with the future information. Compared to other literature that applied such a combination technique, this research is distinct in incorporating the future information into the criteria for clustering the trading points. They also use various return values, H, rather than a single one in the application of the decision tree unlike others. Taiwan and NASDAQ stock markets are used to justify the proposed method. Empirical tests reveal that the filter rule performs the best at (n, k) = (10, 10%) in both markets, n is being the window of time for moving average and k is being the percent change from local maxima or minima.

Besides [4], [7] and [8] also uses technical indicators. [7] uses Exponential Moving Average, Moving Average Convergence Divergence, Relative Strength Index, Money Flow Index, and parabolic Stop and Reverse. The proposed method uses these indicators as features and also it uses stock prices and volume as input variable. Volume is a vital feature for stock analysis and using it would be helpful. The target variable is hold, buy and sell decision. ‘buy’ represents if the next period’s price is expected to be higher than the recent one, if it’s lower than ‘sell’ decision should be made and if it is predicted as same ‘hold’ decision is made. In this study argues that using multiple indicators offers 20% of increase in decision making efficiency compared to using single indicator. In [8], stock prices of three technology company is used as the data set, which are Microsoft, Intel, and IBM. This study used a two-layer bias decision tree. The methodology used in this study differs from that used in other studies in two respects. First, this study modified the decision model into the bias decision model to reduce the classification error. Second, this study used the two-layer bias decision tree to improve purchasing accuracy. In the two-layer bias decision tree, according to the market movement direction that the trading days are classified into upward and downward classes in the first layer, and upward class data is categorized into buy and not- to-buy classes in the second layer. In the learning stage, learning data is classified and features (mean, variance, priori probability, and degree of freedom) are obtained for each class.

Authors of [9] uses ARIMA as well as other nine attributes in the ML algorithms. The paper focuses on usefulness of low complex binary classification to generate consistent profit. 6 machine learning models, which are OneR, C4.5, Jrip, LMT, Kstar and NaiveBayes, are compared with performance. Though the main performance metric of machine learning classification is accuracy, paper emphasizes that high accuracy does not mean high profitable trading. Therefore, cumulative return parameter combined with accuracy is used to evaluate the performance. Retraining period, retraining set size, the number and type of attributes are also investigated to optimize the best performance for each machine learning algorithm.

Comparison of single classifier models with respect to ensemble methods also give interesting results as in [10]. In this study, logistic regression, neural networks, kNN, Support Vector Machine (SVM), Random Forest (RF), AdaBoost and Kernel Factory algorithms are used for the same dataset. Models are cross-validated and compared for Receiver Operating Characteristics (ROC) curve and Area Under Curve (AUC) value. RF outperforms all other methods including SVM. However, this study, likewise many others, does not mention any proposed trading efficiency level increase, thus the readers are left empty for their expectations and are lack of any mean to compare this study with other studies in the literature or in the market. However, one superiority of this study is incorporating fundamental financial information such as, number of employees, non-current liabilities, depreciation, ROI, etc. Yet, many of the input parameters are highly correlated with each other.

Still another study [11] uses Decision Tree based algorithms of Random Forest (RF) and gradient boosted decision tree (GBDT) via XGBoost to predict the price movement before n days. This paper also takes care of the problem as a classification problem not a regression one. Daily market closing prices of 10 stocks and technical indicators calculated from them are used. Time period to predict the price movement Is selected in a wide range beginning 3 till 90 days and interestingly as the forecast period increases the accuracy also increases. This paper also gets rid of noise first using exponential smoothing with a smoothing factor of 0.0095. One problem with the data set is using continuous numerical values in all attributes and the authors explicitly indicated those are not converted to categorical or ordinary values which may cause an overfitting problem. However, Accuracy, Recall, Precision and AUC values are reported to be all very high, i.e. AUC being between 94% and 99% for 90 days-in-advanced prediction. Four of those stocks (APPL, AMZN, FB, MSFT, NKE, TATA, TWTR) are in the upward trend throughout the time period of focus and one of them (TYO) is in the downward trend. These cases are easy ones, however, AMS and SNE stock prices are horizontal and their performance metrics are not different than others. The authors explains the reason behind the better performance of 90-days-in-advanced prediction than 3-days as follows, as the value of “t” increases, more economic indicators are able to capture more information regarding the price movement, however this is not the case because both of the predictions (90-days and 3-days) are done simultaneously with the same available data.

Rather than market closing price of stocks or FOREX parities, voice of crowd may also show the direction in the market. [12] aims to binary option trading as a sentiment detection in order to forecast price movements considering fundamental analysis technique. First step is to determine specific market sentiments which enable to understand a private market emotion. Then, two major emotions can be selected to express the market moods. However, moods should not be considered as positive or negative because same price impacts/movements are not observed in different markets.

III. Methodology

A. Algorithms

According to literature review and applied projects done so far, we investigated several algorithms for classification to apply on financial time series. Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Long Short Time Memory (LSTM) algorithms on financial time series, were applied accordingly. Financial time series are selected from company stocks in 2 different industries, and commodities to show a wide range of applicability of the algorithms.

Decision Tree Algorithm is a well-known classification algorithm and applicable for financial data. It creates a training model to predict classes by inferring decision rules from training data. Decision Tree is especially suitable for the purpose and dataset of this study compared to many other algorithms. According to Mitchell [13], decision tree learning is best suited to problems with such characteristics:

- Instances are represented by attribute-value pairs (as the binary-split attributes),
- The target function has discrete output values (outputs being 1 or 0),
- The training data may contain errors and may have missing attribute values (Technical indicators in some markets do not even generate any signal because all the "interesting" things happen in another market, yet the effect reaches to the market in question by the arbitrage effect.).

One advantage of using Decision Tree is having a descriptive layout to explain the findings in sentences. These could also be used to persuade possible investors.

Second algorithm that was used, Random forest, is an ensemble algorithm and it consists of a great number of individual decision trees. Each individual tree in the random forest has a class prediction and the class that has the most votes according to all trees becomes the model's prediction. Although Random Forest is more complex version of Decision Tree, it is faster to train than decision trees and also parallelizable to split the process into multiple machines to run. So, it provides convenience to work with hundreds of features. Prediction speed of Random Forest is significantly faster than training speed because it generates forests and saves for future uses. Also, it is robust for unbalanced datasets. Averaging property of Random Forest provides usefulness for some aspect. Firstly, it is prone to minimize the overall error. Unlike the Decision Tree which has high variance, Random Forest averages all the trees and also is averaging the variance as well so that low bias and low variance model could be obtained.

Another algorithm used in this study is Support Vector Machine (SVM). Objective of the SVM is to form a hyperplane to distinctly classify the data points in number of feature dimensional space. SVM's are very useful when there is no pre-cognition about the dataset. If there is a clear margin that separates the data, SVM guarantees to find it. In our study, in order to maximize the profit, the negative and positive signals should be separated clearly. Another advantage SVM provides, it has good generalization capabilities which we want to generalize the rules to have healthy buy/sell decisions.

Lastly, Long Short Term Memory Network (LSTM) is a type of Recurrent Neural Networks which is constructed based on the idea where the output from previous steps are fed as input to the next steps and they have memory to learn long-term dependencies. Generally, an LSTM Network tries to remember past data that is fed to the network so far and to forget the irrelevant data which has no effect on prediction. LSTM is used a lot for different kind of problems, especially for sequence prediction problems. Having memory for long time periods makes LSTM very useful for prediction of financial time series. A network with 3 layers with 128 nodes followed by one layer of 32 node and then one layer with 2 nodes is used.

B. Datasets

Financial data comes generally in time series format which includes opening (O) and closing (C) prices for the timestamp, the highest (H) and the lowest (L) prices in that timestamp as well as volume (V) of trade within the timestamp. Technical indicators are generated time series which are hypothesized to signal future movement of price. A detailed over-look of technical indicators used in this study is also given in another section in this study.

Those daily OHLCV values of 4 technology companies (Amazon, Apple, Google, Microsoft), 4 energy companies (BP, ConocoPhillips, Chevron, Exxon) and 3 commodities (Gold, Silver, Copper) are taken from NASDAQ database for the last 10 years.

6 different technical indicators (MACD, RSI, Stochastic Oscillator, William Rate Percentage, PPO, OBV) are calculated for each time series. These generated time series are converted to categorical (with the method mentioned in the next section). Categorical technical indicators fed 3 classification algorithms. On the other hand, for LSTM, closing price and volume of stocks in each industry (energy and technology) and commodity group used in batches to predict price movement in each time series in each of those groups.

C. Labeling through N-K optimization

The most important step in the whole process is labeling the data. Although there are some other choices as well, we labeled data with N percent increase in K timestamps (day, generally) as 1; and 0 if there is no price movement more than or equal to N percent in K days. This N percent and K days not only label the data but also help to balance the dataset. Although there are other methods to balance the dataset, such as oversampling, undersampling, using different costs or having an active learning setup; our method worked very well in the scope of this study.

D. Technical Indicators, Parameter Optimization and Categorizing

Second step in the process is calculating technical indicators. Technical indicators used in this study have their own parameters and those parameters are generally selected based on heuristics. We optimized each parameter of all technical indicators for each time series to have better signals.

Training the models with continuous numerical technical indicator values causes overfitting. To overcome this situation, continuous values are first converted to ordinal values. However, this method did not work well. Then they are converted to binary-categorical according to some thresholds, which are also optimized in the next step. This binary-categorical conversion is worked very well. It both increased the learning speed and performance metrics. As mentioned before, the threshold for each indicator for each time series are also optimized. In traditional use, RSI below 20 (or 30 in some contexts) indicates an oversold state and signals prices to increase in near future. Our threshold optimization yields betters result with a threshold at 35 for technology companies with each algorithm, for example.

Another use of technical indicators is to observe the direction of them. Directions are determined comparing to moving averages. Directions are used alone or in combination with continuous, ordinal, or categorical values. But none of those settings increased performance metrics, if not decreased them significantly.

Having a binary-categorical dataset also eases the necessity to scale technical indicator values, because all the inputs are either 1 or 0.

E. Training/Validation/Test Set Optimization

Datasets were divided between training and test sets for classification algorithms with different ratios and with shuffled and unshuffled settings. No significant difference is observed in this grid search application. Thus 7:3 ratio, shuffled is used for classification algorithms.

In similar manner, training, validation and test sets size along with feeder sequence length, future period to predict and epoch numbers are optimized. 65:5:30 ratio for training, validation and test sets is used. Epoch number of 10 is selected because higher values caused overfitting and lesser values caused underfitting.

F. Performance Metrics

Precision and recall are used as the performance metrics. When talking about financial time-series, precision and recall would have extra meanings. Recall shows how much percentage of trading opportunities we can catch and precision shows how much of those caught opportunities we can really make profit. If average profit per price increase is observed 1% and if the stop-loss order is set at, again, 1%, as long as precision is above 50% we can make profit no matter what recall value we have. (Surely, higher recall levels increase the profit coefficient near-linearly if we roll over profit with simple interest rather than compound-interest-like fashion.)

Sentiment analysis over Twitter users is also intended to be done and feed each of the mentioned machine learning algorithms. To be able to have a query for old tweets, Twitter requires Developer API access. Unfortunately, none of our applications returned with acceptance. Starting real-time tweet mining will not be convenient for the historical datasets used in this study, either. For those reasons, sentiment analysis could not be practiced and included in this study.

A Short Overview of Technical Indicators

1. *RSI*

The RSI was developed by J. Welles Wilder, Jr. and presented in his 1978 book, New Concepts in Technical Trading Systems [15]. It is used to overcome two main problems that occur in market. The first one is that it enables to smooth the momentum movements. In a technical analysis, longer time periods such as weekly or daily graphs are most commonly used to forecast the future prices on grounds that analysts tend to see the big picture instead of unexpected changes in a short time period. However, when a sharp price change occurs in a very short time period, huge difference directly affects the price momentum which does not show the correct value of the momentum on price graphs. Therefore, RSI aims to shape general market momentum by minimizing the huge effects of short time period movements which constructs a distorted momentum line. The second one is that RSI gives a value between 0 and 100 to a momentum line in order to scale the market movements. According to the calculated value, market extreme conditions and market characteristics can be easily seen by traders.

RSI actual formula is calculated as follows: [15]

$$RSI = 100 - \frac{100}{1 + RS}$$

$$RS = \frac{Average\ of\ x\ days'up\ closes}{Average\ of\ x\ days'down\ closes}$$

If the value of RSI is over 70, the price line takes place in overbought case. On the contrary, if it is below 30, the market is in oversold situation. Moreover, as time period of the oscillator is descending in other words used samples are decreasing, the value of RSI is more sensitive to the movements that cause wider amplitude and more fluctuations. Therefore, depending on the trade process, selecting a proper time period is a critical issue. To combine following RSI strategy with experimented time period might bring to a successful conclusion.

Uncorrelated price and RSI movements cause failure swings which show the trend might be changed in following pattern. Furthermore, if momentum line crosses 70 from top to bottom or 30 from down to up, the price leaves extreme cases that generate a sell or buy signal respectively.

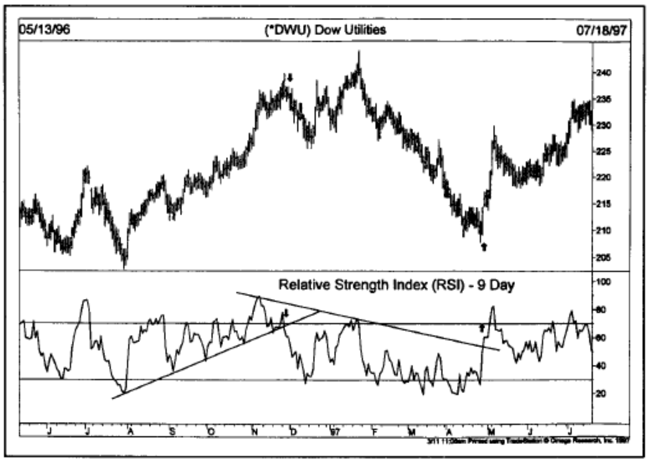


Figure 1: Uptrend Failure Swing and momentum line crosses 70 from top to bottom [15]

According to the Figure 1, the second (2) RSI peak is lower than the first (1) while price line reaches its maximum value. Therefore, a failure swing occurs in RSI oscillator which signals that price value might be decreased. Also, after the second (2) RSI peak, the oscillator crosses 70 from top to bottom which generates a sell signal too. Hence, after obtaining two sell signals, the price line is dramatically decreased.

To sum up, RSI is one of the most powerful oscillators in order to get information from market momentum without affecting a sharp change in prices. Applying a swing failure and extreme condition situations as a strategy might result a beneficial price profit.

2. *Stochastic Oscillator*

The Stochastic oscillator was popularized by George Lane (president of Investment Educators, Inc., Watseka, IL). [15] Stochastic oscillator is based on the logic that closing prices are tend to be closer to trend-oriented price ranges. So, it compares the recent price with the price range for a chosen time period. Stochastic process consists of two main lines which are the %K line and %D line.

Stochastic actual formula is calculated as follows: [15]

$$\%K = 100 * \frac{C - L}{H - L}$$

C: Latest Close

L: Lowest Low

H: Highest High

%K line determines the position of the current price as a scale of 0 to 100 due to the time period price range. %D line is averaged value of %K lines which makes %D line is more stable than %K line. Actual signal is generated when faster %K line crosses the %D line from top to bottom or from bottom to top as a sell alert or buy alert respectively. Very high reading (over 80) and very low reading (under 20) are considered as extreme regions.

The major signal takes place if a divergence occurs between %D line and price in extreme territory.

Additionally, in order to obtain more beneficial and safe trading, stochastic can be combined with RSI.

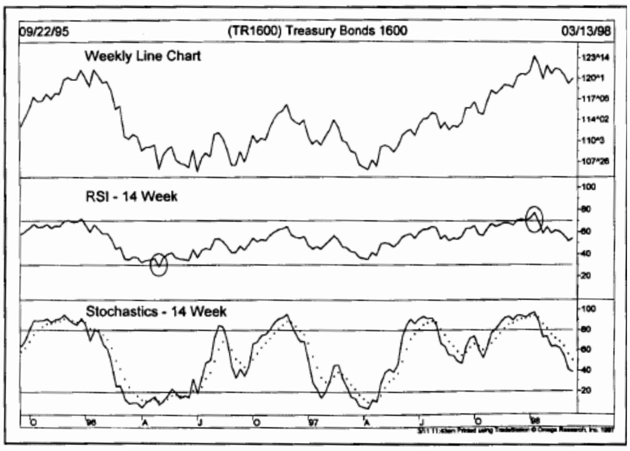


Figure 2: Comparison of RSI and Stochastic [15]

According to the Figure 2, stochastic generates more signal than RSI. However, best signal is obtained when both oscillators are in extreme territory.

3. *Moving Average Convergence/Divergence (MACD)*

The Moving Average Technical Indicator enables to determine the trend either up or down. It averages out the instrument price for a certain period of time. When the actual price is above the moving average, the instrument tends to go on uptrend. On the contrary, if it is below the moving averaged value, it continues its downward trend. According to the calculation, moving averages are categorized into four different types: Simple, Exponential, Smoothed and Weighted. Exponential and weighted types are most commonly used on grounds that both of them assigned more significant coefficient to the latest data.

The Moving Average Convergence/Divergence indicator, or simply MACD, was developed by Gerald Appel. [15] MACD is a trend following dynamic indicator that examine the relationship between two different exponential moving averages closing prices.

MACD actual formula is calculated as follows:

MACD = EMA (CLOSE, LP)-EMA (CLOSE, HP)

SIGNAL = SMA (MACD, T)

LP = Low Period

HP = High Period

T = Period of MACD

EMA = Exponential Moving Average

SMA = Simple Moving Average

SIGNAL = The Signal Line of the indicator

Actual signal is generated when faster line (lower period moving average) crosses the slower line (higher period moving average from top to bottom or from bottom to top as a sell alert or buy alert respectively. Moreover, overbought is occurred when two averages are far away above the zero-reference line. Likewise, oversold condition is presented when two averages are far away however, below the zero-reference line. Two MACD lines are turned into a histogram in order to visualize among relationship. According to the Figure 3, crossing histogram below or above zero represents actual buy or sell signal. Though histogram crossing zero cannot give an actual signal, it provides earlier warning about momentum of the trend.

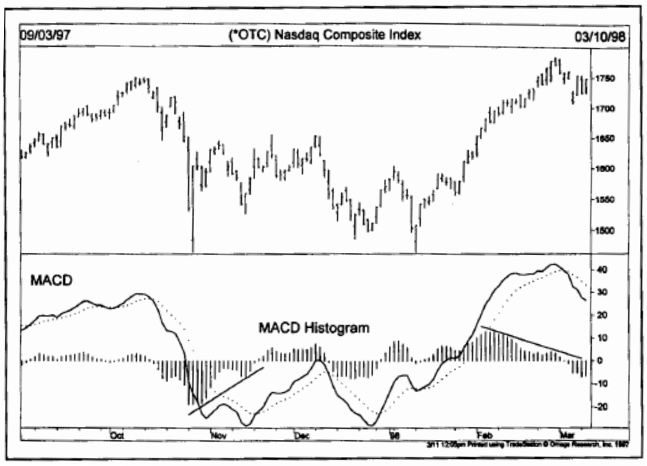


Figure 3: MACD Histogram Plot [15]

4. *On-Balance Volume (OBV)*

On-balance volume is a technical momentum indicator that analyses the volume to price changes. Its working principle is based on the previous price movements. When the current price close is higher than the previous one, the volume of the current bar is added to the on-balance volume indicator. On the other hand, if it is lower than previous bar, the volume of the current bar is subtracted from the on-balance volume indicator. The main purpose is to agree or disagree trend movement by on balance volume. If price and OBV are in same direction, no doubt occurs in a trend which is expected to go on to same direction. However, when uncorrelated movements occur between price and OBV, trend reversal case may be taken place.

OBV actual formula is calculated as follows:

If the current close is higher than previous one;

OBV (t) = OBV (t-1) + VOLUME (t)

If the current close is lower than previous one;

OBV (t) = OBV (t-1) – VOLUME (t)

If the current close is equal to previous one;

OBV (t) = OBV (t-1)

VOLUME (t) = volume of the current bar

OBV (t) = value of the indicator in the current period

OBV (t-1) = value of the indicator in the previous period

IV. RESULTS

N-K values for each time series; parameters and threshold values of each indicator for all time series and classification algorithm pair are optimized. Results for 3 different time series group are given in respective tables (

Table *1*, Table 2, Table 3).

A row in the mentioned tables means that particular algorithm did not work for that particular dataset with indicated N-K values. For example, 1% increase in 1 day (N-K values) is appropriate for BP stock prices to label and balance the dataset. However, Support Vector Machine algorithm does not work for BP in the given settings. More interestingly, prediction of gold price (GCCMX) was impossible for any algorithms with the given settings.

Table 1: N-K and Technical Indicator Optimizations for Technology Stocks

					Indicator Parameters						Indicator Thresholds					
Group	Ratio	N	K	ML Algo	RSI	Stochastic Oscilator	William RP	MACD	PPO	OBV	RSI	Stochastic Oscillator	William RP	MACD	PPO	OBV
Tech	AAPL	1.6	2	DTREE	3	14, 3, 3	5	12, 26, 9	12, 26	NA	35	35	-50	50	50	50
				RF	3	14, 3, 3	5	12, 26, 9	12, 26	NA	35	35	-50	50	50	50
				SVM	3	14, 3, 3	5	12, 26, 9	12, 26	NA	35	35	-50	50	50	50
				Voting	3	14, 3, 3	7	12, 26, 9	12, 26	NA	35	35	-50	20	50	50
	AMZN	1.4	1	DTREE	3	5,3,3	5	12, 26, 9	2, 5	NA	35	10	-50	45	50	50
				RF	3	5,3,3	5	12, 26, 9	2, 5	NA	35	10	-50	45	50	50
				SVM	3	20, 5, 5	5	3, 10, 16	2, 5	NA	35	10	-50	45	50	50
				Voting	3	14, 3, 3	5	8, 17, 9	2, 5	NA	35	35	-50	50	50	50
	GOOG	1.4	2	DTREE	3	10, 3, 3	5	12, 26, 9	2, 5	NA	35	35	-50	50	45	50
				RF	3	10, 3, 3	5	12, 26, 9	2, 5	NA	35	35	-50	50	45	50
				SVM	3	10, 3, 3	5	12, 26, 9	2, 5	NA	35	35	-50	50	45	50
				Voting	3	10, 3, 3	5	12, 26, 9	2, 5	NA	35	35	-50	50	45	50
	MSFT	1.4	2	DTREE	3	14, 3, 3	5	3, 10, 10	12, 26	NA	35	35	-50	50	50	45
				RF	3	14, 3, 3	5	3, 10, 10	12, 26	NA	35	35	-50	50	50	45
				SVM	3	14, 3, 3	5	3, 10, 10	12, 26	NA	35	35	-50	50	50	45
				Voting	3	14, 3, 3	5	3, 10, 10	12, 26	NA	35	35	-50	50	50	45

Table 2: N-K and Technical Indicator Optimizations for Energy Stocks

					Indicator Parameters						Indicator Thresholds					
Group	Ratio	N	K	ML Algo	RSI	Stochastic Oscilator	William RP	MACD	PPO	OBV	RSI	Stochastic Oscillator	William RP	MACD	PPO	OBV
Energy	BP	1	1	DTREE	5	14, 3, 3	3	12, 26, 9	2, 5	NA	5	35	-50	45	50	50
				RF	5	14, 3, 3	3	12, 26, 9	2, 5	NA	5	35	-50	45	50	50
				SVM												
				Voting												
	COP	1.6	2	DTREE	3	20, 5, 5	5	12, 26, 9	2, 5	NA	35	35	-50	50	45	50
				RF	3	20, 5, 5	5	12, 26, 9	2, 5	NA	35	35	-50	50	45	50
				SVM	3	20, 5, 5	5	12, 26, 9	2, 5	NA	35	35	-50	45	45	50
				Voting	3	20, 5, 5	5	12, 26, 9	2, 5	NA	35	35	-50	50	45	50
	CVX	1	1	DTREE	3	20, 5, 5	14	3, 10, 16	10, 21	NA	10	5	-90	20	25	50
				RF	3	14, 3, 3	14	12, 26, 9	10, 21	NA	10	35	-90	50	50	50
				SVM	3	14, 3, 3	3	12, 26, 9	10, 21	NA	20	35	-50	50	50	50
				Voting	3	5, 3, 3	9	12, 26, 9	10, 21	NA	15	30	-70	50	50	50
	XOM	1.4	4	DTREE	11	10, 3, 3	3	12, 26, 9	12, 26	NA	15	35	-50	20	50	50
				RF	11	10, 3, 3	3	3, 10, 16	12, 26	NA	15	35	-50	50	50	50
				SVM	3	10, 3, 3	3	12, 26, 9	12, 26	NA	35	35	-50	50	50	50
				Voting	11	14, 3, 3	3	12, 26, 9	12, 26	NA	15	35	-50	20	50	50

Table 3: N-K and Technical Indicator Optimizations for Commodities

					Indicator Parameters						Indicator Thresholds					
Group	Ratio	N	K	ML Algo	RSI	Stochastic Oscilator	William RP	MACD	PPO	OBV	RSI	Stochastic Oscillator	William RP	MACD	PPO	OBV
Commodity	GCCMX	1	2	DTREE												
				RF												
				SVM												
				Voting												
	HGCMX	1.4	2	DTREE	3	10, 3, 3	3	12, 26, 9	2, 5	NA	35	5	-90	20	20	40
				RF	3	10, 3, 3	3	12, 26, 9	2, 5	NA	35	5	-90	20	20	40
				SVM	3	5, 3, 3	3	12, 26, 9	12, 26	NA	35	35	-90	50	50	35
				Voting	3	10, 3, 3	3	12, 26, 9	2, 5	NA	35	5	-50	50	50	45
	SICMX	1.4	1	DTREE	3	20, 5, 5	3	12, 26, 9	12, 26	NA	35	35	-50	50	50	35
				RF	3	20, 5, 5	3	12, 26, 9	12, 26	NA	35	35	-50	50	50	35
				SVM	3	20, 5, 5	3	12, 26, 9	12, 26	NA	35	35	-50	50	50	35
				Voting	11	5, 3, 3	9	12, 26, 9	10, 21	NA	15	5	-70	50	50	35

Table 4: Performance Metrics and Annual Hypothetical Compounded Profit Values for Technology Stocks

Ratio	N	K	ML Algo	Precision	Recall	stop-loss	opportunity per day	trade per day	**profit per trade**	compound profit over 1 day	compound profit over 1 mo	compound profit over 1 yr
AAPL	1.6	2	DTREE	0.55	0.54	1.00%	0.50	0.27	0.43%	1.00	1.04	1.52
AAPL	1.6	2	RF	0.55	0.55	1.00%	0.50	0.28	0.43%	1.00	1.04	1.53
AAPL	1.6	2	SVC	0.56	0.56	1.00%	0.50	0.28	0.46%	1.00	1.04	1.58
AAPL	1.6	2	Voting	0.55	0.54	1.00%	0.50	0.27	0.43%	1.00	1.04	1.52
AMZN	1.4	1	DTREE	0.54	0.54	1.00%	1.00	0.54	0.30%	1.00	1.05	1.78
AMZN	1.4	1	RF	0.54	0.54	1.00%	1.00	0.54	0.30%	1.00	1.05	1.78
AMZN	1.4	1	SVC	0.55	0.55	1.00%	1.00	0.55	0.32%	1.00	1.05	1.88
AMZN	1.4	1	Voting	0.56	0.55	1.00%	1.00	0.55	0.34%	1.00	1.06	1.97
GOOG	1.4	2	DTREE	0.54	0.53	1.00%	0.50	0.27	0.30%	1.00	1.02	1.33
GOOG	1.4	2	RF	0.53	0.53	1.00%	0.50	0.27	0.27%	1.00	1.02	1.30
GOOG	1.4	2	SVC	0.56	0.56	1.00%	0.50	0.28	0.34%	1.00	1.03	1.41
GOOG	1.4	2	Voting	0.55	0.55	1.00%	0.50	0.28	0.32%	1.00	1.03	1.37
MSFT	1.4	2	DTREE	0.59	0.59	1.00%	0.50	0.30	0.42%	1.00	1.04	1.55
MSFT	1.4	2	RF	0.59	0.59	1.00%	0.50	0.30	0.42%	1.00	1.04	1.55
MSFT	1.4	2	SVC	0.61	0.61	1.00%	0.50	0.31	0.46%	1.00	1.04	1.66
MSFT	1.4	2	Voting	0.60	0.60	1.00%	0.50	0.30	0.44%	1.00	1.04	1.61

Table 5: Performance Metrics and Annual Hypothetical Compounded Profit Values for Energy Stocks

Ratio	N	K	ML Algo	Precision	Recall	stop-loss	opportunity per day	trade per day	**profit per trade**	compound profit over 1 day	compound profit over 1 mo	compound profit over 1 yr
BP	1	1	DTREE	0.50	0.50	1.00%	1.00	0.50	0.00%	1.00	1.00	1.00
BP	1	1	RF	0.50	0.50	1.00%	1.00	0.50	0.00%	1.00	1.00	1.00
BP	1	1	SVC			1.00%	1.00	0.00	-1.00%	1.00	1.00	1.00
BP	1	1	Voting			1.00%	1.00	0.00	-1.00%	1.00	1.00	1.00
COP	1.6	2	DTREE	0.55	0.55	1.00%	0.50	0.28	0.43%	1.00	1.04	1.53
COP	1.6	2	RF	0.55	0.55	1.00%	0.50	0.28	0.43%	1.00	1.04	1.53
COP	1.6	2	SVC	0.55	0.55	1.00%	0.50	0.28	0.43%	1.00	1.04	1.53
COP	1.6	2	Voting	0.55	0.55	1.00%	0.50	0.28	0.43%	1.00	1.04	1.53
CVX	1	1	DTREE	0.52	0.52	1.00%	1.00	0.52	0.04%	1.00	1.01	1.08
CVX	1	1	RF	0.58	0.57	1.00%	1.00	0.57	0.16%	1.00	1.03	1.39
CVX	1	1	SVC	0.59	0.57	1.00%	1.00	0.57	0.18%	1.00	1.03	1.45
CVX	1	1	Voting	0.54	0.54	1.00%	1.00	0.54	0.08%	1.00	1.01	1.17
XOM	1.4	4	DTREE	0.55	0.55	1.00%	0.25	0.14	0.32%	1.00	1.01	1.17
XOM	1.4	4	RF	0.55	0.54	1.00%	0.25	0.14	0.32%	1.00	1.01	1.17
XOM	1.4	4	SVC	0.55	0.55	1.00%	0.25	0.14	0.32%	1.00	1.01	1.17
XOM	1.4	4	Voting	0.55	0.54	1.00%	0.25	0.14	0.32%	1.00	1.01	1.17

Table 6: Performance Metrics and Annual Hypothetical Compounded Profit Values for Commodities

Ratio	N	K	ML Algo	Precision	Recall	stop-loss	opportunity per day	trade per day	**profit per trade**	compound profit over 1 day	compound profit over 1 mo	compound profit over 1 yr
GCCMX	1	2	DTREE			1.00%	0.50	0.00	-1.00%	1.00	1.00	1.00
GCCMX	1	2	RF			1.00%	0.50	0.00	-1.00%	1.00	1.00	1.00
GCCMX	1	2	SVC			1.00%	0.50	0.00	-1.00%	1.00	1.00	1.00
GCCMX	1	2	Voting			1.00%	0.50	0.00	-1.00%	1.00	1.00	1.00
HGCMX	1.4	2	DTREE	0.56	0.56	1.00%	0.50	0.28	0.34%	1.00	1.03	1.41
HGCMX	1.4	2	RF	0.56	0.55	1.00%	0.50	0.28	0.34%	1.00	1.03	1.40
HGCMX	1.4	2	SVC	0.56	0.56	1.00%	0.50	0.28	0.34%	1.00	1.03	1.41
HGCMX	1.4	2	Voting	0.55	0.55	1.00%	0.50	0.28	0.32%	1.00	1.03	1.37
SICMX	1.4	1	DTREE	0.5	0.5	1.00%	1.00	0.50	0.20%	1.00	1.03	1.43
SICMX	1.4	1	RF	0.49	0.49	1.00%	1.00	0.49	0.18%	1.00	1.03	1.36
SICMX	1.4	1	SVC	0.51	0.51	1.00%	1.00	0.51	0.22%	1.00	1.03	1.51
SICMX	1.4	1	Voting	0.54	0.53	1.00%	1.00	0.53	0.30%	1.00	1.05	1.76

Performance metrics and annual hypothetical compounded profit values are presented in respective tables (Table 4, Table 5, Table 6).

Amazon stocks yield higher annual profits. This is due to lower value of K with respective to other stocks, whose K values are generally 2. This means that Amazon prices fluctuates more in the same time period. Exxon, on the other hand, has a K value of 4. This means Exxon stock prices increase 1.4% every 4 days on average. Thus, on average, there would be 90 trading opportunity per annum. Annual profit stays around 1.17 (17% increase with respect to the original capital). However, Amazon stock prices increase 1.4% every other day. (This does not mean a monotonical increase, but a more rapid fluctuation.) Thus, Amazon ends up with around 1.80 (80% increase over the original capital).

There are many N-K pairs to balance the dataset as seen in Figure 4. However smaller pairs yield better profits as in Table 7.

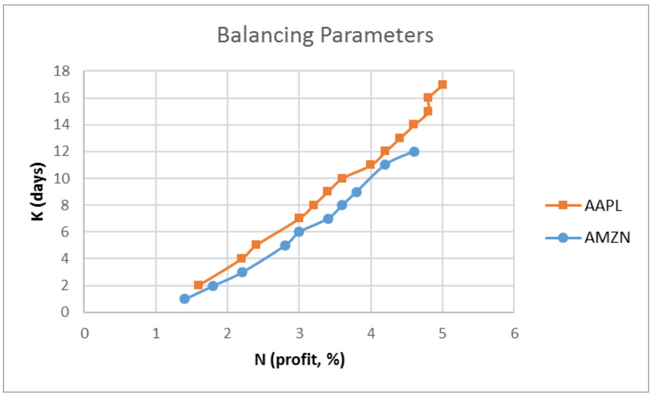


Figure 4: Balancing Parameters

Table 7: Effect of N-K pairs

Ratio	N	K	ML Algo	Precision	Recall	stop-loss	opportunity per day	trade per day	**profit per trade**	compound profit over 1 day	compound profit over 1 mo	compound profit over 1 yr
AAPL	1.6	2	DTREE	0.55	0.54	1.00%	0.50	0.27	0.43%	1.00	1.04	1.52
AAPL	2.2	4	DTREE	0.55	0.54	1.00%	0.25	0.14	0.76%	1.00	1.03	1.44
AAPL	2.4	5	DTREE	0.55	0.54	1.00%	0.20	0.11	0.87%	1.00	1.03	1.40
AAPL	3	7	DTREE	0.55	0.54	1.00%	0.14	0.08	1.20%	1.00	1.03	1.39
AAPL	3.2	8	DTREE	0.55	0.54	1.00%	0.13	0.07	1.31%	1.00	1.03	1.37
AAPL	3.4	9	DTREE	0.55	0.54	1.00%	0.11	0.06	1.42%	1.00	1.03	1.36

For LSTM, two cases are tested: [feeder sequence length: 60 days, K: 3 days, N: 2%] and [feeder sequence length: 20 days, K: 5 days, N: 2%]. Results are very different from each other (Table 8). Google and ConocoPhillips have 51% precision and 96%-93% recall values for different settings.

Table 8: LSTM Results

Group	Ratio	Previous Data	K	N	pportunity	Precision	Recall
Tech	AAPL	60	3	2	0.23	0.83	0.03
	AMZN				0.26	0.51	0.23
	GOOG				0.20	0.51	0.96
	MSFT				0.18	0.00	0.00
Energy	BP				0.17	0.40	0.04
	COP				0.20	0.48	0.65
	CVX				0.15	0.47	0.74
	XOM				0.12	0.50	0.04
Commodit	GCCMX				0.09	0.46	0.81
	HGCMX				0.15	0.40	0.08
	SICMX				0.20	0.00	0.00
Tech	AAPL	20	5	2	0.32	0.42	0.07
	AMZN				0.33	0.59	0.20
	GOOG				0.28	0.83	0.07
	MSFT				0.26	0.47	0.34
Energy	BP				0.24	0.55	0.63
	COP				0.27	0.51	0.93
	CVX				0.23	0.58	0.35
	XOM				0.18	0.91	0.09
Commodit	GCCMX				0.16	0.00	0.00
	HGCMX				0.22	0.44	0.36
	SICMX				0.26	0.67	0.03

Comparison of classification algorithms and LSTM algorithm are given in Table 9. It is important to note high standard deviation values of LSTM. This could mean that we need a better optimization and more values of feeder sequence length, K and N values to have more reliable results.

Table 9: Comparison of classification algorithms and LSTM

	ML				LSTM			
Ratio	Precision	StDev	Recall	StDev	Precision	StDev	Recall	StDev
AAPL	0.55	0.00	0.55	0.01	0.61	0.17	0.10	0.07
AMZN	0.55	0.01	0.55	0.01	0.53	0.05	0.33	0.16
GOOG	0.55	0.01	0.54	0.01	0.60	0.16	0.48	0.37
MSFT	0.60	0.01	0.60	0.01	0.33	0.23	0.37	0.31
BP	0.50	0.00	0.50	0.00	0.49	0.06	0.48	0.31
COP	0.55	0.00	0.55	0.00	0.52	0.04	0.64	0.24
CVX	0.56	0.03	0.55	0.02	0.51	0.05	0.56	0.16
XOM	0.55	0.00	0.55	0.01	0.64	0.19	0.26	0.28
GCCMX					0.33	0.24	0.34	0.34
HGCMX	0.56	0.00	0.56	0.01	0.48	0.08	0.23	0.12
SICMX	0.51	0.02	0.51	0.01	0.38	0.28	0.08	0.09

Effect of precision, recall, target profit (N) and stoploss values are also investigated in Table 10, Table 11.

Table 10: Effect of Precision and Recall Values

average p	1.00%									
stoploss	1.00%									
	recall									
precision	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%
40%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
45%	0.00	**0.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.00
50%	0.00	**0.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.00
55%	108.08	**125.67**	**143.91**	**162.81**	**182.40**	202.71	222.21	243.75	266.04	289.13
60%	268.04	**315.77**	**367.06**	**422.18**	**481.41**	545.06	603.31	674.46	750.69	832.35
65%	481.04	578.18	686.37	806.89	941.12	1090.65	1220.12	1395.48	1589.72	1804.84
70%	764.61	940.28	1143.09	1377.25	1647.60	1959.74	2213.32	2595.46	3032.80	3533.13
75%	1142.01	1439.78	1796.13	2222.56	2732.85	3343.52	3804.66	4581.33	5499.06	6582.91
80%	1644.14	2128.62	2729.53	3474.83	4399.23	5545.76	6342.00	7849.81	9688.22	11928.02
85%	2312.05	3078.25	4063.19	5329.29	6956.82	9048.95	10368.76	13200.86	16761.72	21234.39

Table 11: Effect of Target Profit and Stoploss Values

recall	0.47										
precision	0.57										
	average profit										
stoploss	0.75%	0.80%	0.85%	0.90%	0.95%	1.00%	1.05%	1.10%	1.15%	1.20%	1.25%
0.70%	180.20	236.03	297.48	365.12	439.58	521.53	611.73	711.00	820.27	940.51	1072.85
0.80%	105.49	153.79	206.95	265.48	329.90	400.80	478.85	564.76	659.31	763.37	877.90
0.90%	40.87	82.65	128.64	179.27	235.00	296.35	363.88	438.21	520.02	610.07	709.18
1.00%	0.00	21.10	**60.89**	**104.69**	**152.90**	**205.98**	**264.40**	**328.71**	**399.50**	**477.42**	563.18
1.10%	0.00	0.00	**2.29**	**40.17**	**81.88**	**127.79**	**178.34**	**233.98**	**295.22**	**362.64**	436.84
1.20%	0.00	0.00	**0.00**	**0.00**	**20.44**	**60.16**	**103.88**	**152.01**	**205.00**	**263.33**	327.53
1.30%	0.00	0.00	**0.00**	**0.00**	**0.00**	**1.65**	**39.47**	**81.11**	**126.95**	**177.41**	232.95
1.40%	0.00	0.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**19.78**	**59.43**	**103.08**	151.13
1.50%	0.00	0.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**1.02**	**38.78**	80.35
1.60%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	19.12
1.70%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.80%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.90%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2.00%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

According to Table 10, if target profit (N) and stoploss values are equal to each other, a precision higher than 50% is needed to be able to gain profit. In addition to this observation, recall value is correlated with profit. For average precision and recall values, as in Table 11, if stoploss value is able to be kept lover than target profit (N), we are able to gain. To sum up, if precision is higher than 50% and target profit is higher than stoploss models tested result successful trades; and higher recall values increase the profit increasing number of trades.

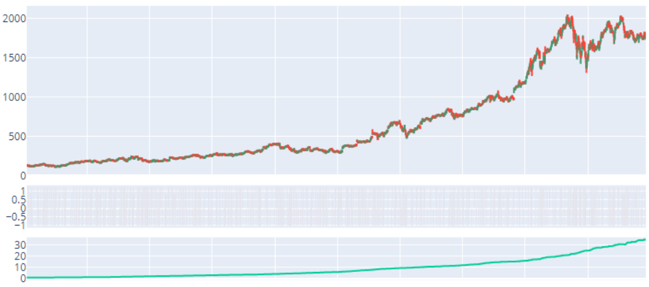
Lastly we simulated performances of three stocks in a backtesting setup. Algorithm used is Support Vector Machine, which is selected for its good results (generally 1-2% higher than Decision Tree and Random Forest) and fast performance.

Table 12: Sample Backtesting Table for Amazon Stock Trading in the Backtesting

Date	Close	Position	Position2	Signal	Buy Price	Sell Price	Trade Profit	Cum Profit	Position3
2010-01-15 0:00:00	127.14	wait to buy	0	0	0.00	0.00	1	1	0
2010-01-19 0:00:00	127.61	wait to buy	0	0	0.00	0.00	1	1	0
2010-01-20 0:00:00	125.78	wait to sell	bought	1	125.78	0.00	1	1	-1
2010-01-21 0:00:00	126.62	wait to buy	sold on target	1	0.00	127.54	1.014	1.014	1.2
2010-01-22 0:00:00	121.43	wait to sell	bought	1	121.43	0.00	1	1.014	-1
2010-01-25 0:00:00	120.31	wait to sell	0	1	121.43	0.00	1	1.014	0
2010-01-26 0:00:00	119.48	wait to sell	0	1	121.43	0.00	1	1.014	0
2010-01-27 0:00:00	122.75	wait to buy	sold on target	1	0.00	123.13	1.014	1.028	1.2
2010-01-28 0:00:00	126.03	wait to buy	0	0	0.00	123.13	1	1.028	0
2010-01-29 0:00:00	125.41	wait to buy	0	0	0.00	123.13	1	1.028	0
2010-02-01 0:00:00	118.87	wait to sell	bought	1	118.87	0.00	1	1.028	-1
2010-02-02 0:00:00	118.12	wait to sell	0	1	118.87	0.00	1	1.028	0
2010-02-03 0:00:00	119.1	wait to sell	0	1	118.87	0.00	1	1.028	0
2010-02-04 0:00:00	115.94	wait to sell	0	1	118.87	0.00	1	1.028	0
2010-02-05 0:00:00	117.39	wait to buy	sold for stoploss	1	0.00	118.04	0.993	1.021	0.8
2010-02-08 0:00:00	116.83	wait to sell	bought	1	116.83	0.00	1	1.021	-1
2010-02-09 0:00:00	118.03	wait to buy	sold on target	0	0.00	118.47	1.014	1.035	1.2
2010-02-10 0:00:00	117.36	wait to sell	bought	1	117.36	0.00	1	1.035	-1
2010-02-11 0:00:00	120.09	wait to buy	sold on target	0	0.00	119.00	1.014	1.050	1.2
2010-02-12 0:00:00	119.66	wait to buy	0	0	0.00	119.00	1	1.050	0
2010-02-16 0:00:00	117.53	wait to sell	bought	1	117.53	0.00	1	1.050	-1
2010-02-17 0:00:00	116.31	wait to sell	0	1	117.53	0.00	1	1.050	0
2010-02-18 0:00:00	118.08	wait to sell	0	0	117.53	0.00	1	1.050	0
2010-02-19 0:00:00	117.52	wait to sell	0	1	117.53	0.00	1	1.050	0
2010-02-22 0:00:00	118.01	wait to sell	0	0	117.53	0.00	1	1.050	0
2010-02-23 0:00:00	117.24	wait to buy	sold on target	0	0.00	119.18	1.014	1.064	1.2

Amazon stocks yielded 3,459% profit (almost 35 times of the original capital) in 10 years. A standard passive buy&hold strategy would bring 1,369%. (Table 13)

Table 13: Price (OHLCV) and Compounded Profit for Amazon Stock



Exxon stocks bring 2,077% profit and buy&hold would bring 98%, which means a loss of 2% of the original capital. (Table 14)

Table 14: Price (OHLCV) and Compounded Profit for Exxon Stock

Table 13 and Table 14 shows daily price in the first graph. Second graph shows buy, sell with target profit and sell at stoploss flags. The third graph shows compounded profit.

It is important to note that trading with ML algorithms allows us to catch small fluctuations (N% in K days), hence we do not get effected from the general trend of stock prices.

V. Conclusion

In this study, we aim to investigate possible use cases of Decision Tree, Random Forest, Support Vector Machine and LSTM algorithms on financial time series. In general, we catch near 55% of trading opportunities (recall) and near 55% success in each trade (precision). We observed that higher frequency brings higher gain however this requires faster computers and data stream.

Quants are increasingly taking major roles of the whole investing world and especially in hedge funds [14]. Today many hedge funds operate with automated trading bots doing very high-frequency trading. As noted in a Forbes Magazine article, "the firm (Renaissance Technologies Corporation) only profits on barely more than 50 percent of its trades, a sign of how challenging it is to try to beat the market." [16] (Renaissance Technologies Corporation is believed to be the most successful hedge funds in the world right now.)

Decisions sterilized from human emotions, even exploiting them, are taken the main investor roles today. We hope this study opens the door of financial machine learning and bring two area of expertise together with usable insights.

References

[1] K. C. Cheung, "10 Applications of Machine Learning in Finance," *Algorithm-X Lab*, 2019. [Online]. Available: https://algorithmxlab.com/blog/applications-machine-learning-finance/. [Accessed: 02-Nov-2019].

[2] H. Markowitz, "Portfolio Selection," *J. Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[3] C. X. Nie and F. T. Song, "Analyzing the stock market based on the structure of kNN network," *Chaos, Solitons and Fractals*, vol. 113, pp. 148–159, 2018.

[4] R. Dash and P. K. Dash, "A hybrid stock trading framework integrating technical analysis with machine learning techniques," *J. Financ. Data Sci.*, vol. 2, no. 1, pp. 42–57, 2016.

[5] D. Lv, S. Yuan, M. Li, and Y. Xiang, "An Empirical Study of Machine Learning Algorithms for Stock Daily Trading Strategy," *Math. Probl. Eng.*, vol. 2019, no. 2, 2019.

[6] M. C. Wu, S. Y. Lin, and C. H. Lin, "An effective application of decision tree to stock trading," *Expert Syst. Appl.*, vol. 31, no. 2, pp. 270–274, 2006.

[7] F. X. S. D. Nugroho, T. B. Adji, and S. Fauziati, "Decision support system for stock trading using multiple indicators decision tree," *2014 1st Int. Conf. Inf. Technol. Comput. Electr. Eng. Green Technol. Its Appl. a Better Futur. ICITACEE 2014 - Proc.*, pp. 291–296, 2015.

[8] J. L. Wang and S. H. Chan, "Stock market trading rule discovery using two-layer bias decision tree," *Expert Syst. Appl.*, vol. 30, no. 4, pp. 605–611, 2006.

[9] E. A. Gerlein, M. McGinnity, A. Belatreche, and S. Coleman, "Evaluating machine learning classification for financial trading: An empirical approach," *Expert Syst. Appl.*, vol. 54, pp. 193–207, 2016.

[10] M. Ballings, D. Van Den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7046–7056, 2015.

[11] S. Basak, S. Kar, S. Saha, L. Khaidem, and S. R. Dey, "Predicting the direction of stock market prices using tree-based classifiers," *North Am. J. Econ. Financ.*, vol. 47, no. June 2018, pp. 552–567, 2019.

[12] A. Cofnas, "Sentiment Analysis: New Predictive Tools," *Trading Bin. Options*, pp. 27–42, 2016.

[13] T. M. Mitchell, *Machine Learning*. 1997.

[14] Financial Times, "Quantitative Investing," *Financial Times*, 2007. [Online]. Available: https://www.ft.com/content/a8e30716-4cae-11dc-a51d-0000779fd2ac. [Accessed: 02-Nov-2019].

[15] Wilder, J. (1999). *Technical analysis of the financial markets*. Paramus: New York Institute of Finance.

[16] Forbes. "How Billionaire Jim Simons Learned To Beat The Market—And Began Wall Street's Quant Revolution", *Forbes*, 2019. [Online]. Available: https://www.forbes.com/sites/forbesdigitalcovers/2019/11/08/jim-simons-the-man-who-solved-the-market-gregory-zuckerman-book-excerpt/#85e0a4213b6a. [Accessed: 19-Dec-2019].