Application of machine learning techniques for stock price direction forecasting

Juliana Negrini de Araujo
Faculty of Engineering, Environment and Computing, Coventry University
MSc Data Science and Computational Intelligence (ECT104) Stage 1
Coventry, United Kingdom
negrinij@uni.coventry.ac.uk

***Abstract*—Machine learning methods are applied to estimate next day Facebook and Apple stocks closing direction. The dataset is composed by previous day closing, open, high and low prices. Technical Indicators and Global Indexes can also provide valuable information regarding stock market trends and are included in the dataset. Sliding window training routine is applied and three machine learning techniques are compared.**

Keywords—stock price movement, SVM, logistic regression, random forest

I. INTRODUCTION

Time series modelling for the prediction of stocks prices is a challenging task. Political events, market expectations and economic factors are just a few known factors that can impact financial market behaviour [1]. The financial market is a complex, noisy, evolutionary and chaotic field of study that attracts many enthusiasts and researches — the first, usually driven by the economic benefit of it, the latter, inspired by the challenge of handling such complex data.

The use of machine learning techniques in time series forecasting started over thirty years ago, with the successful application of Artificial Neural Networks (ANN's) for linear and non-linear time series modelling [2]. After demonstrating similar or better results than ANN in several benchmarks, Support Vector Machine (SVM) also gained popularity in the late '90s [3]. In comparison to ANN, inherent properties of SVM may present some advantages for financial market analysis. SVM models are more likely to find a global optimum solution while ANN tends to find a local optimum solution [4]. Besides, most of ANN's use empirical risk minimisation principle, where the goal is to find a solution that reduces the empirical risk function and therefore reduce training set error. SVM algorithm applies structural risk minimisation (SRM), where the selected model is the one that provides the best trade-off between model complexity and empirical error [5]-[6]. Currently, SVM and Neural Networks (NN) are widely used for stock index prediction [7] although other classification models such as tree-based classifiers and logistic regression have also demonstrated impressive results [8]-[10].

[5] analysed the direction of change in the daily Korea stock price index (KOSPI). The accuracy of SVM results was 3% higher than Back Propagation Neural Networks (BPNN). [11] applied a hybrid solution of Genetic Algorithm and SVM (GASVM) to predict S&P 500 stock prices index. The hybrid model has reached a hit-ratio of 84.6%. In this study, the SVM model accuracy has also surpassed BPNN by 8%. SVM also outperform ANN in work from [3]-[5]. In [12] SVM and random forest models are compared for S&P 500 stock prices index prediction. SVM presented higher accuracy than random forest on this study. [1] and [7] have found that ensemble methods such as Random Forest provided better results than ANN and SVM models for financial market data analysis. [7] also report the importance of transforming technical indicators continuous data into trend deterministic data as a valid method to improve model accuracy. [8] applied logistic regression, ANN, Random Forest (RF), and Gradient Boosted Trees (XGBoost) algorithms to predict the direction of S&P 500 stocks. Their extensive study on 463 stocks showed that logistic regression with lasso penalisation achieved slightly better results than the other models analysed. [10] focused on the implementation of Random Forest and Gradient Boosted Trees to predict the stock price direction (up, down) of ten companies, including Apple and Facebook. On their study, RF and XGBoost models outperform SVM, Logistic Regression and ANN.

SVM and ANN are the leading machine learning techniques applied in the price market prediction, although there are studies with appealing results where different Machine Learning algorithms can provide equal or even better results. This paper aims to predict Apple (AAPL) and Facebook (FB) next day stock price direction with machine learning algorithms. Technical indicators and global market indexes are used, and their influence on the forecast accuracy is analysed. The machine learning models applied to stock movement forecasting are SVM, random forest and logistic regression. Final results are compared with existing literature for Apple and Facebook stocks.

II. METHODOLOGY

A. Research Data

For each stock, daily values were retrieved (volume, open, close, low and high prices) from Yahoo! Finance website [13]. For Facebook (FB), May 2012 was the earliest data available while for Apple (AAPL) stock data was available from July 2005.

The closing price of current day $C_{(t)}$ and closing price from the previous day $C_{(t-1)}$ are compared to build the initial dataset. The objective is to define if the price trend is going up or down by analysing these two values. For each instance, a comparison was made and recorded. If the price is going up, $C_{(t)} > C_{(t-1)}$, class "1" is assigned to the dependent variable y_t^i, as show Eq. (1). Class "0" is assigned for the opposite case, $C_{(t)} \leq C_{(t-1)}$. The examples of both classes can be considered as equally distributed in the dataset. A summary can be found in Table I.

$$y_t^i = \begin{cases} 1, if\ C_{(t)} > C_{(t-1)} \\ 0, if\ C_{(t)} \leq C_{(t-1)} \end{cases} \quad (1)$$

For each sample, the previous day close, open, high and low values are used as input. Data regarding the closing price

direction of past days can also be added as additional features. They are called *lag* features and are widely used in time series forecasting, as mentioned by [8]. The closing price direction of the previous five days was added to the dataset.

TABLE I - SUMMARY OF STOCKS ANALYSED.

Stock	Data range	Number of samples	Percent. of class 1 samples
Facebook (FB)	09/07/2012 – 28/11/2018	1644	52.4%
Apple (AAPL)	01/02/2005 – 28/11/2018	3482	52.5%

Further input features were added next. Technical indicators are used by researches and financial market analysts to support stock market trend forecasting. There are several technical indicators available in the literature. The work from [4], [7] and [11] was used as a reference. Common indicators between these three papers were selected and calculated for Facebook and Apple stocks. The final list and their respective formulas are given in Table II.

Deterministic data is used as features in [7] and the same concept is applied to this dataset. Technical indicators provide a suggestion of the stock price movement. For each technical indicator, its daily value is analysed, and a class is assigned for that day. Class "1" is given if the indicator suggests upper trend, class "0" for a downtrend. In other words, financial market analysis is performed at a simplistic level, in the attempt to translate what the continuous value means. For each technical indicator of Table I, the continuous value is converted into one of the two classes according to its interpretation. How each technical indicator trend was defined is described next.

Some indicators follow the same trend of the stock price, meaning that if the indicator value is increasing with time, the stock is also likely to be in an 'up' trend. This is achieved by comparing the indicator current and previous day value and assigning value "1" or "0" accordingly. This analysis is valid for Stochastic Oscillator (SO), Moving Stochastic Oscillator (SSO), Moving Average Convergence Divergence (MACD) and Accumulation / Distribution (ADO) indicators. The daily value of Moving Average (MA), Weighted Moving Average (WMA) and Exponential Moving Average (EMA) needs to be compared to the stock closing price for interpretation. If the closing price is higher than the moving average value, an 'up' trend is characterized, and number "1" is assigned. A positive value for Momentum (M) indicates a rise in the stock price. Number "1" is added for positive and "0" for negative values. Relative Strength Index (RSI) values range from 0 to 100 and are used to characterise overbought or oversold periods. If a stock is overbought, above 70, it means its value may decrease soon and class 0 is added. Class 1 is for the oversold region, when RSI values are below 30 and prices are likely to rise. For values between 30 and 70, class 1 is allocated if $RSI_{(t-1)}$ is higher than $RSI_{(t-2)}$. For the Commodity Channel Index (CCI) a similar analysis is done, but the range used is 200 and -200.

For a given country or region, the stock market index characterises the performance of its financial market and the overall local economy. As part of a globalised economy, foreign markets daily performance can influence the behaviour of the selected American stocks. Asian and European financial markets open and close before the US market, where the AAPL and FB stocks are traded. For this reason, the same day performance of these markets could contribute to the machine learning model predictions. Similar to [8], six global indexes were added as features, with their closing direction as up or down, class "1" or "0", respectively. Table III provides information regarding the indexes used. Data for these indexes were also retrieved from Yahoo! Finance [13].

TABLE II - LIST OF TECHNICAL INDICATORS USED.

Technical Indicator	Formula
Moving m-day Average (MA)	$MA = \frac{C_{(t-1)} + C_{(t-2)} + \cdots + C_{(t-m)}}{m}$
Weighted Moving m-day Average (WMA)	$WMA = \frac{(m)C_{(t-1)} + (m-1)C_{(t-2)} + \cdots + C_{(t-m)}}{m + (m-1) + \cdots + 1}$
Momentum (M)	$M = C_{(t-1)} + C_{(t-m)}$
Stochastic Oscillator (SO)	$SO = \frac{C_{(t-1)} - LL_{(t-(m-1))}}{HH_{(t-(m-1))} - LL_{(t-(m-1))}} x100$
Moving Stochastic Oscillator (SSO)	$SSO = \frac{1}{m}\sum_{i=t-m+1}^{t}(SO_{(t-1)})$
Exponential Moving Average (EMA)	$EMA = \alpha C_{(t-1)} + EMA(k)_{(t-1)}$ Where, $\alpha = \frac{2}{k+1}$ and, k is the time period (days) used for EMA
Moving Average Convergence Divergence (MACD)	$MACD = MACD(k)_{(t-1)} + \frac{2}{k+1}[(EMA(12)_{(t-1)} - EMA(26)_{(t-1)}) - MACD(k)_{(t-1)}]$
Relative Strength Index (RSI)	$RSI = 100 - \frac{100}{1+RS}$ Where, $RS = \frac{\sum_{i=1}^{m-1}\frac{UP_{(t-1)}}{n}}{\sum_{i=1}^{m-1}\frac{DW_{(t-1)}}{n}}$ $UP_{(t-1)} = upward\ price\ change$ $DW_{(t-1)} = downward\ price\ change$
Commodity Channel Index (CCI)	$M_{(t-1)} = \frac{H_{(t-1)} + L_{(t-1)} + C_{(t-1)}}{3}$, $SM_{(t-1)} = \frac{\sum_{i=1}^{m} M_{(t-i+1)}}{m}$, $D_{(t-1)} = \frac{\sum_{i=1}^{m}\lvert M_{(t-i+1)} - SM_{(t-1)}\rvert}{m}$
Accumulation / Distribution Oscillator (ADO)	$ADO = \frac{(C_{(t-1)} - L_{(t-1)}) - (H_{(t-1)} - C_{(t-1)})}{H_{(t-1)} - L_{(t-1)}}$

C=Closing price at time t, H=Highest price at time t, HH=Highest high price in period t days, L=Lowest price at time t, LL=Lowest low price in period t days.

B. Training Routine

For time series forecasting, training and testing methods must consider chronological order. It is not advised to use random samples to build training and test set as the order of events is a relevant characteristic of the time series model [7]. The approach taken to train and test the machine learning models is referred to as sliding window or walk-forward, and it is illustrated in Figure 1. This method consists of dividing

the dataset into several time frames. The final model and parameters are the ones that provide the best accuracy considering the average accuracy from all windows. The sliding window routine creates training and test set of equal size for each window; the sets go forward in time according to the step size selected. By applying this method, the model is based on more recent data.

TABLE III - LIST OF MARKET INDEXES USED.

Index	Description	Symbol
Nikkei 225	Trades on the Tokyo Stock Exchange and contains main 225 Japanese companies.	N225
Hang Seng	Hong Kong market index contains major local companies.	HSI
All Ordinaries	Index share contains 500 Australian companies.	AORD
Euronext 100	Comprises the 100 largest Euronext stocks. Companies from France, Netherlands, Belgium, Portugal and Luxembourg are part of this index.	N100
SSE Composite Index	Largest Chinese stock exchange, this index represents all shares traded in Shanghai Stock Exchange.	SSE
DAX	Trades on the Frankfurt Stock Exchange and comprises 30 major German businesses.	DAX

The datasets are divided into five windows of equal size. The training and test ratio used is 80-20 for each window. The step size and number of samples for training and testing are described in Table IV.

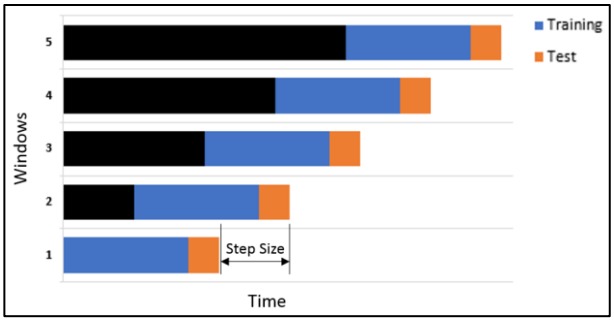

Figure 1 - Sliding window train and test routine.

For the individual windows, the model that provides the best accuracy is kept and the parameters recorded. After finding the best parameters for each window, a second run is performed. This time, the best model of each window is used against all windows, and the accuracy is evaluated. The parameters that provide higher overall model accuracy, by averaging the accuracy of all windows, is defined as the final model solution.

TABLE IV - TRAINING, TEST AND STEP SIZE USED FOR EACH STOCK.

Stock	Number of Windows	Train (samples)	Test (samples)	Step Size
Apple	5	815	205	615
Facebook	5	460	115	260

C. Feature Analysis

Due to reasonable dataset size and performance, it was decided that there were no requirements to perform feature engineering such as PCA or regression feature elimination. A general analysis is performed for the final best performing model. The 42 input features are incrementally added to a dataset and a new model is created. The classifier runs and the accuracy recorded. This is done for each window separately. This aims on having a better understanding of how much the additional features were enhancing the model. This is illustrated in Figure 2 and 3.

D. Support Vector Machines (SVM)

Support vector machines can be used for classification or regression problems. SVM performs classification by identifying a maximum margin of separation between the classes, a hyperplane or a line for a two-dimensional case [7]. A kernel function can be applied for non-linear separable models, and the initial input vectors are mapped to a higher dimensional feature space. The objective is to obtain the minimal number of misclassifications while maximising the hyperplane margin. The margin gamma, γ, is defined as the distance of the closest sample from the hyperplane. A penalty parameter C (cost) controls if the model will prioritise margin maximisation or classification error reduction. As value C rises, higher misclassification cost is applied for samples located inside the margin boundaries [14]. C values were trailed in the range of 10 – 100, higher values did not demonstrate meaningful accuracy increase. For γ, the range of 0.1 – 0.0001 was iterated for each window, with a step size of $10*\gamma$. The selected parameters for the individual windows are shown in Table V. Radial basis function (RBF) kernel provided best results in comparison with Gaussian or linear kernels.

E. Random Forest

Random forest is a popular ensemble method of machine learning that consists of several decision trees classifiers. In decision tree algorithms, the input space is recursively partitioned until a pure local model is defined in each region of the input space. The splits are performed where the highest impurity reduction can be achieved in order to obtain a final region that contains samples of a single class. The result is a tree structure, where each specific region is a leaf or a node of the model [1]. Random Forest uses several binary trees, and each tree uses a random sample of input features and data cases. The final classification is obtained by averaging the result of each binary tree. This is the basic principle behind the bagging technique, also known as bootstrap aggregation. The entropy and information gain is calculated to define node split quality. The parameters tuned for random forest are number of trees, maximum depth of each tree and number of features to be randomly selected at each split. Values in the range of 1 up to 100 were trailed to determine the appropriate number of trees, with an increment of 1. Primary tests showed that higher values were overfitting the model to the training data. The number of features (n) of each bootstrap sample was defined as $\sqrt{n}$. The depth of the trees was also iterated in the same range as the number of trees. Top parameters shown in Table VI.

F. Logistic Regression

Logistic regression is a statistical machine learning algorithm used to solve classification problems. It performs a binary classification by modelling the probability of an event occurring depending on the given input features. Classifications are performed by estimating the probability of an observation belonging to a specific class. This probability is calculated by using a logistic function [15]. As it is the case

for SVM, penalisation parameter C can be varied for model tuning. For this model, lasso penalisation was applied, and different C values were tested. The best C value varied significantly between the windows, for this reason, average values will also be considered in the final model. Table VII describes the best cost values for this algorithm.

III. EXPERIMENTAL RESULTS

An initial test was performed to find the best set of parameters for each window. The results are displayed in Table V-VII.

TABLE V - SVM BEST PENALISATION AND GAMMA VALUES FOUND FOR INDIVIDUAL WINDOWS.

Stock	Best C values	Best gamma values
Apple	10, 20, 30*,40	0.0001, 0.001, 0.01
Facebook	1, 30, 40, 60	0.001, 0.01

* resulted from the average of all values found

TABLE VI - RANDOM FOREST BEST NUMBER AND DEPTH OF TREES FOUND FOR INDIVIDUAL WINDOWS.

Stock	Best Number of Trees	Best Depth of Trees
Apple	1, 2, 7, 10	1,2,3,4
Facebook	1, 30, 40, 60	2,3

TABLE VII - LOGISTIC REGRESSION BEST C VALUES FOUND FOR INDIVIDUAL WINDOWS.

Stock	Best C
Apple	1, 10,45*,70,100
Facebook	1, 10,20*,50

* resulted from the average of all values found

The results were evaluated using accuracy and f-measure scores to measure the performance of the different machine learning techniques applied. The formulas for both are defined in Eq. (4) and (5). For F-Measure, parameter β is defined as 1 indicating that PPV and SE have the same weight. The input values of F-measure are retrieved from test set results.

$$PPV = \frac{TP}{TP + FP} \tag{2}$$

$$SE = \frac{TP}{TP + FN} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$F - Measure = \frac{(1 + \beta^2) \times PPV \times SE}{\beta^2 PPV + SE} \tag{5}$$

The parameters reported in Table V-VII are tested across all five windows for both stocks. Table VIII, Table IX and Table X present the final three best results for SVM, random forest and logistic regression models considering accuracy and f-measure results.

It is noticeable that the results from Apple are not quite satisfactory as Facebook. For Apple stock, all machine learning methods have similar result reaching 60%. Random forest performs slightly better than logistic regression and SVM. For Facebook better accuracies were found, SVM resulted as the top model with 78.4% of accuracy, followed by logistic regression which achieved results above 75%. As contrary from Apple, for Facebook random forest has shown inferior results. The depth of trees values are similar for both stocks, and the results indicate that test accuracy is penalised as the depth of trees is increased. Which is coherent, as deeper the tree more likely it is to overfit. In Table X, logistic regression results present a more apparent difference between training and testing sets than SVM or random forest results.

TABLE VIII - TOP THREE BEST PERFORMING SVM MODELS.

Stock	Gamma	C	Training Accuracy	Testing Accuracy	F-Measure
Apple	0.001	30	61.58	59.23	60.45
	0.001	40	61.85	59.03	60.48
	0.001	10	61.58	59.13	60.14
Facebook	0.001	60	79.75	78.39	74.71
	0.001	40	80.05	77.86	81.99
	0.001	50	80.23	77.86	81.54

TABLE IX - TOP THREE BEST PERFORMING RANDOM FOREST MODELS

Stock	Trees	Depth	Training Accuracy	Testing Accuracy	F-Measure
Apple	1	3	61.11	57.87	62.39
	10	1	60.72	59.43	60.68
	10	3	66.19	58.65	60.46
Facebook	10	2	68.72	65.20	68.18
	40	2	70.63	66.44	65.83
	20	3	75.63	64.68	65.18

TABLE X - TOP THREE BEST PERFORMING LOGISTIC REGRESSION MODELS.

Stock	C	Training Accuracy	Testing Accuracy	F-Measure
Apple	100	64.30	59.03	60.30
	70	64.28	58.84	60.15
	45	64.38	58.55	59.97
Facebook	1	80.14	75.92	76.88
	20	81.20	75.92	77.25
	50	80.80	75.39	76.91

Ten stocks were analysed in the work of [10], including Facebook and Apple. Our results are partially consistent with their work. Applying random forest and using a trading window of three days, [10] obtained an accuracy of 65.3% for Apple and 67.7% for Facebook. Our results have found lower accuracy for random forest. Considering Facebook with SVM, higher accuracy was achieved. This is expected since we utilised a one-day trading window.

Figures 2 and 3 illustrate how individual windows perform as features are incrementally added to the model. The models used are the SVM with best parameter combination. Even though the incremental addition of features is a superficial analysis since different combinations of input features can provide different results, it is interesting to visualize this data. For Apple, Figure 2, it is noticeable the flat regions and how the models respond better to the final features. From the graph, it is visible that each window uses the features differently and this oscillation might have contributed to Apple's poor performance. The results suggest that it's required a different selection and set up of the indicators based on the characteristics of the stock being analysed. Figure 3 presents the same analysis with Facebook stock. All windows seem to

roughly follow the same pattern, and feature 21 (Momentum) seems to improve the model considerably.

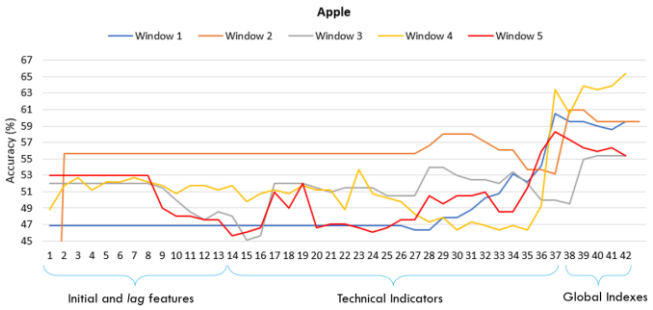

Figure 2 - Analysis of Apple's model response to additional features.

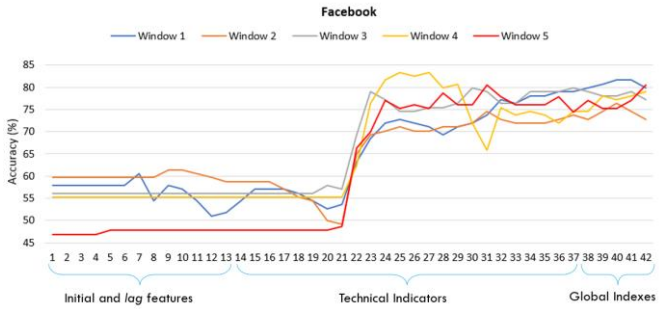

Figure 3 - Analysis of Facebook's model response to additional features.

Figure 4 presents both stocks prices variation over the years. This helps to understand the results obtained. From Mid 2012 to beginning 2018 both tech companies showed an uptrend market. Apple and Facebook exhibited an uptrend with Facebook showing a more consistent uptrend while Apple had a stepper and more oscillatory uptrend. These behaviours help to explain why the accuracy of Facebook prediction in Table X is higher than Apple. Based on the behaviour exhibited by Apple it was expected that the introduction of the technical indicators would increase the accuracy although this was not verified in the model (graphical analysis). This can be related to the oscillatory behaviour of the stock and the time scale used for the indicators.

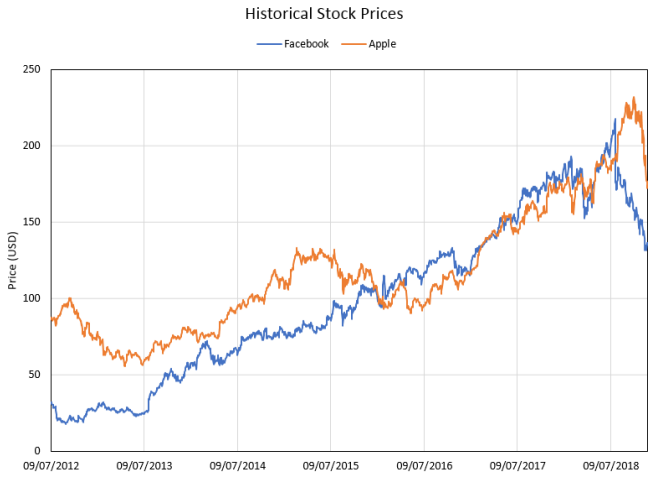

Figure 4 – Historical closing price of Apple and Facebook stocks.

In [12] study on S&P 500 index, the SVM outperforms random forest model. Logistic regression has shown better results than random forest for [8]. In contrast, the work from [1], [7], [10] reported random forest as the best performing machine learning model for the stocks they analysed. This variation in which model performs best suggests that the appropriate machine learning method is linked to the stocks being analysed as well as features being used. Our findings show SVM with RBF kernel as the most reliable model for AAPL and FB, since it can achieve good results for both stocks.

IV. DISCUSSION AND CONCLUSION

The financial stock market is an interesting time-series model that allows the exploration of machine learning techniques. It is interesting to highlight how the usage of different input features improve stock prediction accuracy. This suggests that there will be always room for model improvement.

For further enhancement, more advanced parameter optimization selection could contribute to the model's accuracy, especially for SVM model. Also, addition of more technical indicators and better understanding of their relationship with a particular stock would certainly improve accuracy. For example, on this work we focused on using the same indicators for both stocks. An interesting study would be to use different indicators to check if Apple's accuracy would improve. Additionally, different range of moving averages and other technical indicator parameters can be optimized for a specific stock. Other types of input features are also used but not covered in this work. Socio-economic indicators, currency values and even sentiment analysis could be added. Regarding training, increasing the number of windows may also be productive since the model will use more recent data as input. The definition of the window size on this work has been done by trial and chosen same window for both stocks. Selecting a monthly step size for each window can also be an interesting experiment.

Our results have shown that SVM outperforms logistic regression and random forest algorithms for stock movement prediction. The inherent capability of SVM to avoid overfitting contributed to this conclusion. During experimental testing, random forest showed a high tendency to overfit to training data achieving contradictory results between training and test accuracy. Essential contributions from literature allowed the construction of a model that can forecast with similar accuracies of current published papers. As it is the case for market traders, the usage of technical indicators and global indexes have shown to be a powerful strategy to support forecast decisions.

V. APPENDIX

Original dataset and python programming files are available at this link: http://tiny.cc/jncp1y

VI. REFERENCES

[1] M. Ballings, D. V. d. Poel, N. Hespeels and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," in *Expert Systems with Applications*, 2015, vol. 42, pp. 7046-7056.

[2] G. Bontempi, S. Ben Taieb and Y.-A. Le Borgne, "Machine Learning Strategies for Time Series Forecasting," in *Business Information Processing*, vol. 138, M. Aufaure and E. Zimányi, Eds., Brussels, Springer, Berlin, Heidelberg, 2012, pp. 62-77.

[3] V. Kecman, "Support Vector Machines – An Introduction," in *Support Vector Machines: Theory and Applications*, vol. 177, L. Wang, Ed., Springer, Berlin, Heidelberg, 2005, pp. 1-47.

[4] K. N. Devi, V. M. Bhaskaran and P. Kumar, "Cuckoo Optimized SVM for Stock Market Prediction," in *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, India, 2015, pp. 1-5.

[5] K. J. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1-2, pp. 307-319, 2003.

[6] L. J. Cao and F. E. H. Tay, "Support Vector Machine With Adaptive Parameters in Financial Time Series Forecasting," *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1506-1518, 2003.

[7] J. Patel, S. Shah and P. Thakkar, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques," *Expert Systems with Applications*, vol. 42, pp. 259-268, 2015.

[8] Y. Jiao and J. Jakubowicz, "Predicting Stock Movement Direction with Machine Learning: an Extensive Study on S&P 500 Stocks," in *IEEE International Conference on Big Data*, Boston, USA, Dec. 2017, pp. 4705-4713.

[9] S. B. Imandoust and M. Bolandraftar, "Forecasting the direction of stock market index movement using three data mining techniques: the case of Tehran Stock Exchange," *Int. Journal of Engineering Research and Applications*, vol. 4, no. 6, pp. 106-117, 2014.

[10] S. Basak, S. Kar, S. Saha, L. Khaidem and S. Roy Dey, (in press) "Predicting the direction of stock market prices using tree-based classifiers," *North American Journal of Economics and Finance*, (2018), doi: 10.1016/j.najef.2018.06.013.

[11] L. Yu, S. Wang and K. Keung Lai, "Mining Stock Market Tendency Using GA-Based Support Vector Machines," *Internet and Network Economics. WINE 2005. Lecture Notes in Computer Science*, vol. 3828, pp. 336-345, 2005.

[12] M. Kumar and M. Thenmozhi, "Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest," *Indian Institute of Capital Markets 9th Capital Markets Conference Paper*, 2006.

[13] Yahoo, "Yahoo! Finance," 2018. [Online]. Available: https://uk.finance.yahoo.com/. [Accessed 29 11 2018].

[14] K. P. Murphy, Machine Learning: A probabilistic perspective, Cambridge, MA: The MIT Press, 2012.

[15] M. Y. Chen, "Predicting corporate financial distress based on integration of decision tree classification and logistic regression," *Expert Systems with Applications*, vol. 38, no. 9, pp. 11261-11272, 2011.