Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 10 (2022) pp. 001-009

Prediction of Moroccan Stock Price Based on Machine Learning Algorithms

Abdelhadi Ifleh[1] and Mounim El Kabbouri[1]

[1] Finance, Audit and Organizational
Governance Research Laboratory, National School of Commerce and Management, Settat, Hassan First University of Settat, Morocco
{a.ifleh, mounime.elkabbouri}@uhp.ac.ma

***Abstract*: Stock price prediction one of the most fascinating challenges for both professionals and academicians, especially in high-volatility and high-complexity environments. Traders employ a variety of strategies to forecast stock values. In this study, we used a combination of technical indicators (TI) and the Random Forest (RF) algorithm to forecast Moroccan stock market in several time periods (1, 5 and 10 days), and we compared the findings to those of the Support Vector Machine (SVM) model. For all of the datasets tested, the results demonstrate that the RF technique beats SVM in different time frames. The accuracy, F-Score, Recall, and AUC of the ROC curve were used to assess the robustness of our model.**

***Keywords*:** Moroccan Stock Price, Machine Learning, RF, SVM, ROC curves.

I. Introduction

Predicting asset prices of financial markets is an important issue for researchers and practitioners. The main objective is to have the possibility to take the best trading decision at the right time -baying the assets at the lowest prices and sell them at the highest prices. To attain this objective, researchers and practitioners use generally two approaches, fundamental approach and technical approach.

The first one, it based on news, company's activity and macroeconomic indica-tors to take trading decisions. The fundamental analysts take decisions of buying or selling a stock by comparing the intrinsic value to the actual value. They pro-ceed to buy (sell) a stock if the fundamental value is superior (inferior) than actu-al [1]. The efficient market hypothesis, EMH henceforth, supposes that the intrin-sic value is equal to the market price and any disequilibrium will be corrected by rational traders, the market price will congregate to the intrinsic value [2]. This theory assume that stock market prices are not forecastable, and market prices follow a random walk [3].

While fundamental analysists are based on intrinsic value to take trading deci-sions, technical analysists are based on TI, and they are built via using mathemat-ics and statistics tools on the stock's price (Open, High, Low, Close and Volume) [1].To predict stock's direction and take the right decision, technical analysists mobilize a variety of time series analysis methods like ARIMA, ARCH, and GARCH.

Stock markets are known with their high volatility and risk due to the wide range of factors that impact them. For this reason, traders attend to predict stock prices in order to maximize (minimize) their profit (loss).

At the present time, the development of technologies and the rising power of arti-ficial intelligence, AI henceforth, give a new method which can lead investors and researchers to predict stock prices. Machine learning, ML henceforth, is a field of AI and it is one of the most used algorithms in predicting stock price direction.

We are going to predict the Moroccan stock prices using RF and then compare the results with SVM model. This algorithm is a supervised learning classifier. It uses the historical data to forecast the future direction. We are going to mobilize this model to help traders to make more accurate decisions and beat the Moroccan stock market. In section 2, we will present different using artificial intelligence to predict stock prices. In section 3, we explain the datasets, the features and the ML algorithms. Section 4 discusses the results of the prediction of the model and the comparison between the two classifiers. And we conclude in section 5.

II. Related works

Predicting stock prices using historical data contradicts with EMH assumptions. In fact, all available information is introduced in the price, and any deviation caused by irrational traders will be corrected by rational traders.

Based on literature review, researchers use various methods to predict future movement of prices, Neural Networks, Artificial Neural Networks, Long Short Term Memory, SVM, Decision Trees (DT), RF, K-Nearest Neighbor and others.

Rodrigo et al. [4] in their study have used K-Nearest Neighbor algorithm to determine when to buy and sell stocks in Sao Paulo Stock Market. By combining Piecewise Linear Representation and Artificial Neural Network in different TI, Chang & al. [5] developed a model in order to predict future buy and sell signals, and this model beat other models of other researchers (Giles et al.[6]; Mallick et al.[7]; [5]). Agrawal et al. [8] compared the performance of different models: Optimal-LSTM, Event-LSTM, SVM and Linear Regression, as a result OLSTM had higher accuracy than other models.

Sezer et al. [9] applied Neural Networks on the RSI and moving average (MA) indicators to detect when to buy and to sell. Ou et al. (2009) made a comparison between various models, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) , K-nearest neighbor classification, Naive Bayes based on kernel estimation, Logit model,Tree based classification , neural network, Bayesian classification with Gaussian process, SVM and Least Squares SVM (LS-SVM), to forecast Hang Seng price using historical data, and concluded that SVM and LS-SVM had higher accuracy.

Basak et al. [10] trained and tested the prediction of six features with RF and Gradient BoostedDT models in different time horizons. They found that the predictions are accurate in the medium and long terms than short term. In their turn, Kumar et al. [11] compare the prediction accuracy of five models, RF, SVM, NN, Logit and Discriminant Analysis. They applied those models to forecast the direction of S&P CNX and they found that RF and SVM are more accurate.

Khaidem et al. [12] used to predict the direction of different stocks in different time horizons using RF to train and test their six extracted features, and they found out that this model is more accurate in the long term and it is confirmed with ROC curve. Imandoust et al. [13] compared between DT, RF and Naïve Bayesian in predicting the direction of the Tehran Stock Exchange Index. They used three types of features, TI, fundamental indicators and both of them. As a result, DT was found sharper than other models using the different inputs. Shen et al. [14] used different type of features. They utilized various major international financial assets as inputs for predicting the direction of some indexes and they mobilized the SVM model to train and predict the model. Consequently, this model gave good outcomes compared to other models and it beat two proposed strategies.

III. Methodology

In this research we aim to predict Moroccan stock market movement using algorithms of ML, RF. Our objective is to help traders to make good decisions at the right time to minimize their losses and maximize their gains. Also, we intend to reject the EMH by exploiting historical data to determine the future.

We are going to predict five stocks listed in Casablanca Stock Exchange (CSE), AttijariWafa Bank (ATW), Banque Marocaine pour le Commerce et l'Industrie (BMCI), Bank of Africa (BOA), Crédit Immobilier et Hôtelier(CIH) and Banque Centrale Populaire (BCP), all these stocks are from the banking sector. The historical data was extracted from the official websites of CSE and the Deposit and Management Fund. We have selected ten years of historical data between 04/01/2010 and 10/04/2020, and it, furthermore, includes the open, high, low and close prices and volume. We have proposed our method to predict Moroccan stock price and it is divided into 6 main steps. In step 1, we import our data. In step 2, we extract features from the imported data. In step 3, we split our data into training and testing data, 80% and 20% respectively. In step 4, we train our data with RF. In step 5, we predict and evaluate our predictions with proposed accuracy metrics. In the last one, we compare the accuracy of RF with SVM.

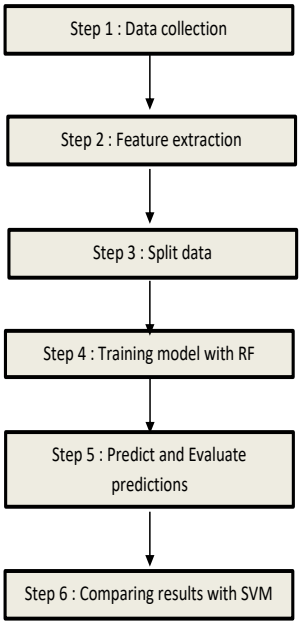

Figure 1: Our proposed method

A. Feature extraction

In financial markets traders use TI to predict if the prices are going to increase or decrease [15]. In this research, we are going to calculate some of the important TI using our historical data. All TI which we have to use to train and test our model are presented below:

1) Relative Strength Index (RSI)

One of the most popular TI, it oscillates between 0% and 100%. When itis under 30%, we say that the asset is oversold and that we should look for an occasion to buy the asset and vice versa when it is above 70% [15].

$$\mathrm{RSI} = 100 - \left[\frac{100}{1+\mathrm{RS}}\right] \quad \textbf{(1)}$$

$$\mathrm{RS} = \mathrm{Average}\left[\frac{\text{X day's Up closing Price}}{\text{X day's Down closing price}}\right] \quad \textbf{(2)}$$

2) Simple Moving Average (SMA)

It gives an idea about the current price relative to a mean of previous N-days. It also allows to identify the trend [15]; we can find other different types of MA, like exponential, smoothed, and so on.

$$\mathrm{SMA} = \sum_{x=1}^{X} \frac{\text{X day's closing Price}}{X} \quad \textbf{(3)}$$

3) Moving Average Convergence Divergence (MACD)

This This indicator is represented by three elements:
- Firstly, is MACD line which is the difference between the exponential moving averages (EMA) of two different periods usually 26 days and 12 days.
- Secondly, is the MACD EMA line calculated over 9 days. It is called the signal line (SL).
- Thirdly, is the histogram which represents the difference

between the MACD and the SL.
If the MACD is above the SL, there is an upward trend. In the opposite case there is a bearish trend (Clement Thierry (b) 2018) and (Anthony Busière 2019).

$$\mathbf{MACD} = \mathbf{EMA_{12}} - \mathbf{EMA_{26}} \quad \textbf{(4)}$$

$$\mathrm{SL} = \mathrm{EMA}(\mathrm{MACD}_9) \quad \textbf{(5)}$$

4) Bollinger Bands (BB)

The Bollinger Bands are channels of evolution (envelope) which are drawn on both sides of the MA. They are calculated by adding (upper band) or subtracting (lower band) to the MA twice the value of the standard deviation of the stock's price, calculated over the period of the MA. As a result, prices have a 98% chance of being inside the Bollinger Bands [15].
Bollinger bands were developed by John Bollinger in 1980. They are used to measure price volatility. The distance between the bands means that the price is volatile and the narrowing of the bands means that the price is not volatile.
Thus, the Bollinger bands determine the direction of the trend. If the price oscillates between the upper band and the MA then the price is in an uptrend. And when it oscillates between the MA and the lower band, then it is in a downward trend.

$$\mathrm{MA} = \sum_{x=1}^{X} \frac{\mathrm{X\ day's\ closing\ Price}}{\mathrm{X}} \quad \textbf{(5)}$$

$$\mathrm{Upper\ Band} = \mathrm{MA} + 2\mathrm{StDv} \quad \textbf{(6)}$$

$$\mathrm{Lower\ Band} = \mathrm{MA} - 2\mathrm{StDv} \quad \textbf{(7)}$$

StDv is the standard deviation.

5) Rate of Change (ROC)

ROC is a momentum-based TI referring to the percentage change in price between the current price and the one from a number of periods ago.

$$\mathrm{ROC} = \frac{\mathrm{Price_t} - \mathrm{Price_{t-n}}}{\mathrm{Price_{t-n}}} \quad \textbf{(8)}$$

6) Percentage Price Oscillator (PPO)

The PPO is a momentum-based TI that provides a measure of the change between the fast and slow EMAs as a percentage of the slow EMA. PPO values are not affected by the amount of the stock price. It focuses on the ROC of MA. The PPO values may be comparable for different stocks, although the price differences are significant [15].

$$\mathrm{PPO} = \frac{(\mathrm{EMA_{12}} - \mathrm{EMA_{26}})}{\mathrm{EMA_{26}}} \quad \textbf{(9)}$$

7) Triple Exponential Average (TRIX)

The triple exponential moving average oscillator (TRIX) is a momentum-based TI that fluctuates around zero. It indicates the percentage variation between two triples smoothed EMAs [16].

$$\mathrm{Trix_t} = \mathrm{EMA_t}(\mathrm{EMA_t}(\mathrm{EMA_t}(\mathrm{Close}))) \quad \textbf{(10)}$$

B. Machine learning algorithms

1) Random Forest

RF introduced by Breiman (2001) is a supervised ML algorithm. RF is an extension of the DT model.

We build our RF model based on the conventional approach given in Breiman (2001). There are no changes introduced to the algorithm, as it is estimated that the original RF may have enough capacity to process a wide range of variables in the datasets and result in an unbiased prediction for real-world classification problems, including finance.

Basically, the RF is composed of multiple deep uncorrelated DT built on multiple samples of data (Breiman, 2001). The construction process of a RF is simple. For every DT, we initially randomly generate a subset as a sample of the original data set. We then grow a DT with this sample to its maximum depth of J_{RF}. During this time, m_{RF} features utilized on each split are randomly collected from *p* features. After repeating the process multiple times using the original dataset, n_{RF} DT are produced. The final outcome is a collection of all DT, and classification is conducted through majority voting. The computational complexity can be simply estimated $O(n_{RF}(p * n_{ins} * lign_{ins}))$, where n_{ins} represents the number of instances in the training datasets. Three metrics need to be set to verify the robustness of RF on classification, which are the number of trees n_{RF} , the maximum depth J_{RF} and the number of features m_{RF} in every split [17].

Various extensions may be made to the actual RF algorithm, like the mixed RF-neural network machine, and the multi-label learning machine advanced by RF. These approaches are interesting in that they lead to a broader diversification of the model, and can be used to increase the efficiency and effectiveness of our stock selection in the future[17].

In RF classifier model, predictions are based on majority voting predictions made by a multiple trees. In our case, we will use 100 trees. Concretely, each tree of the RF is trained on a random subset of data according to the principle of bagging (correct the instability of DT by reducing the variance), with a random subset of features according to the principle of "random projections" (reduce the dimensionality of a set of points). The data used in RF is split into subsets, and the commonly splitting criterion used in RF is GINI index.

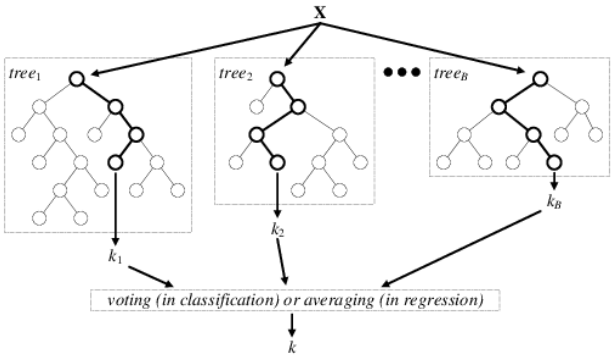

Figure 2. RF architecture

2) Support Vector Machine

SVM is also is popular and reliable algorithms for predicting stock price direction. SVM is a supervised ML technique that we use for both classification and regression. It was created in the 1990s. Its goal is to use a hyperplane in N-dimensional space to divide data points into classes.

Within a decade, SVM have gained a lot of interest as a new classification technique with good generalization capacity. However, the basic idea of SVMs is to fit input vectors into a high-dimensional feature space and linearly split the feature vectors with an optimal hyperplane in terms of margins, SVMs are promising approaches for financial time series prediction because they employ a risk function composed of the empirical error and a regularized term derived from the principle of structural risk minimization [18]. Considering a training data set which is expressed by the X-matrix ($X_1 \ldots . X_m$) divided into two linearly separable classes with class labels (+1 and 1) stored in the Y-vector ($y_1 \ldots . y_m$), the maximum margin plane can be found by minimizing ($\|w\|_2$):

$$\|\mathrm{w}\|_2 = \mathrm{w}.\,\mathrm{w} = \sum_{\mathrm{i}=1}^{\mathrm{d}} \mathrm{w}_1{}^2 \quad \textbf{(11)}$$

with constraints:

$$\mathrm{y_i}(\mathrm{w}\cdot\mathrm{x_i} + \mathrm{b}) \geq 1 \quad \textbf{(13)}$$

where i = 1,...,m, b ∈ R, and $x_i \in R^d$.

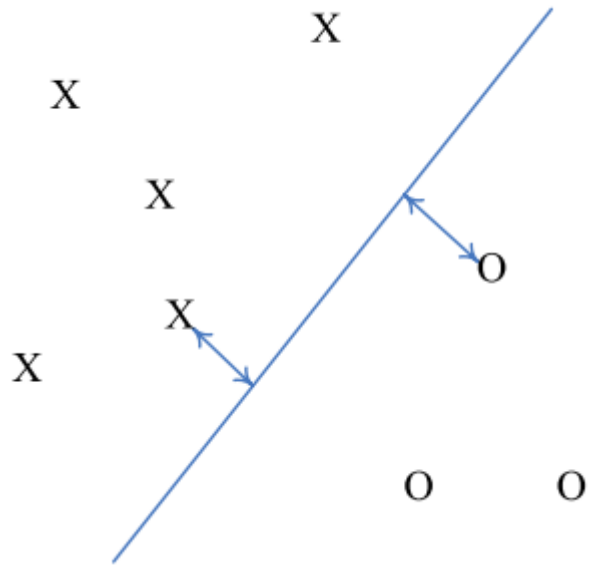

Figure 3. SVM architecture

Fig. 3 illustrates the most simple example of two classes that are linearly separated. The concept of support vectors is depicted by the closest points on the surface dividing two classes, and the margin is given by the distance between the support vectors and the dividing surface.

The decision function has the following form $f(x) = sgn(w\cdot x + b)$, where sgn(·) is a sign function which returns +1 for positive arguments and -1 for negative arguments. This simple classification problem is generalized to a non-separable case by adding surplus variables x_i and minimizing the following quantity:

$$\frac{1}{2}\mathrm{w}\cdot\mathrm{w} = \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{m}} \xi_{\mathrm{i}} \quad \textbf{(14)}$$

where $y_i(w\cdot x_i + b) \geq 1 - \xi_i$ and $\xi_i > 0$.

The above quadratic optimization problem with constraints may be reformulated by using Lagrange multipliers, and the

$$\mathrm{L}(\mathrm{w},\mathrm{b},\xi,\alpha,\nu) = \frac{1}{2}\mathrm{w}\cdot\mathrm{w} + \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{m}} \xi_{\mathrm{i}} - \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_{\mathrm{i}}\left[\mathrm{y_i},(\mathrm{w}\cdot\mathrm{x_i} + \mathrm{b}) - 1 + \xi_{\mathrm{i}}\right] - \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{v_i}\xi_{\mathrm{i}} \quad \textbf{(12)}$$

following Lagrangian is given:

Stationary points of this Lagrangian can be obtained by:

$$\frac{\partial \mathrm{L}}{\partial \mathrm{w}} = \mathrm{w} - \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_{\mathrm{i}}\,\mathrm{y_i}\mathrm{x_i} = 0 \quad \textbf{(15)}$$

$$\frac{\partial \mathrm{L}}{\partial \mathrm{b}} = \sum_{\mathrm{i}=1}^{\mathrm{m}} \alpha_{\mathrm{i}}\,\mathrm{y_i} = 0 \quad \textbf{(16)}$$

$$\frac{\partial \mathrm{L}}{\partial \xi_{\mathrm{i}}} = \alpha_{\mathrm{i}} + \mathrm{v_i} - \mathrm{C} = 0 \quad \textbf{(17)}$$

Two remaining derivatives $(\frac{\partial L}{\partial \alpha}, \frac{\partial L}{\partial v})$ recover the constraint equations. By replacing the formula $w = \sum_{i=1}^{m} \alpha_i \, y_i x_i$ in the Lagrangian, we obtain this simpler dual formulation:

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i \alpha_j \, y_i y_j x_i \cdot x_j - \sum_{i=1}^{m} \alpha_i \qquad (18)$$

With the constraints $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{m} \alpha_i \, y_i = 0$.
Given a mapping

$$x \to \phi(x),$$

The dot product in the final space can be substituted by a Mercer kernel

$$\phi(x)\phi(y) \to K(x, y) \qquad (19)$$

Since $\phi(\cdot)$ generally maps x to a much larger dimensional space, it is commonly expressed by defining the kernel implicitly.The above dual formulation thus becomes a preferable approach due to the high-dimensional feature space elicited by the $\phi(\cdot)$ mapping. The decision function for classification problems is then given by:

$$f(x) = sgn(\sum_{sv} \alpha_i y_i K(x_i, x) + b) \qquad (20)$$

C. Evaluation metrics

1) Accuracy

The simplest scoring measure to employ is model accuracy, which can be defined in simple words as the number of "hits" over the number of predictions. However, a more technical definition is the proportion of the sum of true positives (TP) and true negatives (TN) relative to the total population size. This metric, however, does not reflect the predictive power of the model in cases where the data are strongly biased or where we need to give more weight to some correct predictions. As an example, imagine a scenario in which we have a dataset of 100 observations and 99 of them are non-bankrupt firms. Under this scenario, the model can attain superficially high accuracy by making a prediction that all observations are not bankrupt in order to take advantage of the high base rate of non-bankrupt firms in the data set. However, this would contradict the goal we originally set for predicting business bankruptcies in advance. The bankruptcy dataset in fact has the pitfalls mentioned above. The bank failure data are highly biased because they are mostly comprised of non-failed firms and few failed firms. Furthermore, we would be more interested in TP, i.e., correct predictions about bankrupt issuers (TP) instead of correct predictions about nonbank issuers (TN). Thus, accuracy for bankruptcy datasets would be a poor metric for evaluating the performance of the model [19].

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn} \qquad (21)$$

2) Precision and Recall

The best alternatives that can overcome precision problems for unbalanced datasets are precision and recall. These two parameters are commonly used in classification problems in the ML literature. First, precision is the fraction of TP relative to the number of positive predictions. For the bankruptcy prediction problem, this means that precision is the probability that a company predicted to be bankrupt is actually bankrupt. For the bankruptcy prediction problem, this means that precision is the probability that a company predicted to be bankrupt is actually bankrupt. A second way of looking at this problem is that recall is the fraction of TP over the number of all positives in the data set. This means that recall is the probability that a firm in bankruptcy is predicted to be in bankruptcy in the model. Precision and recall are examined together because there is often an opposite relationship between precision and recall. Also, three main methods exist for calculating the average of these scores for multi-class classification problems. The micro average would calculate these scores by aggregating TP and false positives (FP) globally. The macro average would compute these scores for every class, then add them up. Finally, the weighted average would compute these scores for every class and weight them according to the number of observations with those labels. While both of these metrics help us to better estimate model performance, it is also challenging to compare models using both metrics at the same time [19].

$$\text{Recall} = \frac{tp}{tp+fn} \qquad (22)$$

$$\text{Precision} = \frac{tp}{tp+fp} \qquad (23)$$

3) F Score

In order to overcome this limitation of precision and recall, we can use the F-score, a metric that combines both precision and recall in the form of a harmonic mean.
The F1 score which provides equal weights to recall and precision is given as follows:

$$F\_Score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (24)$$

Hence, the F score may be used as a single measure to evaluate the performance of the models and select the model with the highest F1 score [19].

4) Receiver Operating Characteristics (ROC)

Most of the classification models mainly create scores and make a classification according to their comparison to a certain threshold. Consequently, the predictions might change depending on how that threshold changes. The Receiver Operating Characteristics (ROC) curve, otherwise known as ROC curve, is constructed to demonstrate the progression of the TP rate or the recall rate relative to the FP rate as a function of threshold. This visualization aids model developers in reflecting on what threshold they may want to establish as a function of the trade-off between TP rate and FP rate. In Addition, the area under the curve (AUC) or ROC score is a metric that indicates how predictive the model is. A random prediction model will have an AUC of 0.50 since it will have an equal number of true and FP at any threshold. Alternatively, an ideal model would have an AUC of 1, because this means that the model has a 100% TP rate while the FP rate is 0%. Therefore, a model with a larger AUC is a better predictive model, all else being equal, and this metric can be used to

compare different models [19].

IV. Results and Discussion

1) RF model results

Based on the prediction results of our model we can decide if the stocks of our sample should be bought or sold after n days (1 day, 5 days and 10 days). Which means that we can expect after n days if the price of the stocks will rise or fall. The evaluation of the robustness of our prediction is an important task. There are many metrics to evaluate our model. In our case, we are going to use prediction metrics presented above, Accuracy, Recall, F Score and AUC. The tables below present the results of our chosen stocks:

Stock	*Accuracy*	*Precision*	*Recall*	*F_Score*	*AUC*
ATW	65.74	65.74	65.74	65.74	74.00
BOA	74.34	74.34	74.34	74.34	82.00
BCP	62.36	62.36	62.36	62.36	67.00
BMCI	61.64	61.64	61.64	61.64	65.00
CIH	66.12	66.12	66.12	66.12	73.00

Table 1. Classification results using RF for 1 day

Stock	*Accuracy*	*Precision*	*Recall*	*F_Score*	*AUC*
ATW	74.71	74.71	74.71	74.71	83.0
BOA	65.45	65.45	65.45	65.45	73.0
BCP	75.90	75.90	75.90	75.90	83.0
BMCI	74.20	74.20	74.20	74.20	81.0
CIH	75.50	75.50	75.50	75.50	80.0

Table 2. ***Classification results using RF for 5 days***

Stock	*Accuracy*	*Precision*	*Recall*	*F_Score*	*AUC*
ATW	94.39	94.39	94.39	94.39	98.00
BOA	82.00	82.00	82.00	82.00	87.00
BCP	86.24	86.24	86.24	86.24	90.00
BMCI	85.03	85.03	85.03	85.03	90.00
CIH	92.15	92.15	92.15	92.15	92.00

Table 3. *Classification results using RF for 10 days*

The results of classification show that the performance of predictions increases progressively with the increase of the window-width. This increase over a time frame might be explained by the fact that the features capture more information within a long-time span. In 1-day predictions we obtain higher evaluation metrics in BOA stock. In 5-days predictions all metrics are above 74% in all stocks except BOA which are equal to 65,45%. In 10-days predictions all metrics are above 72% in all stocks

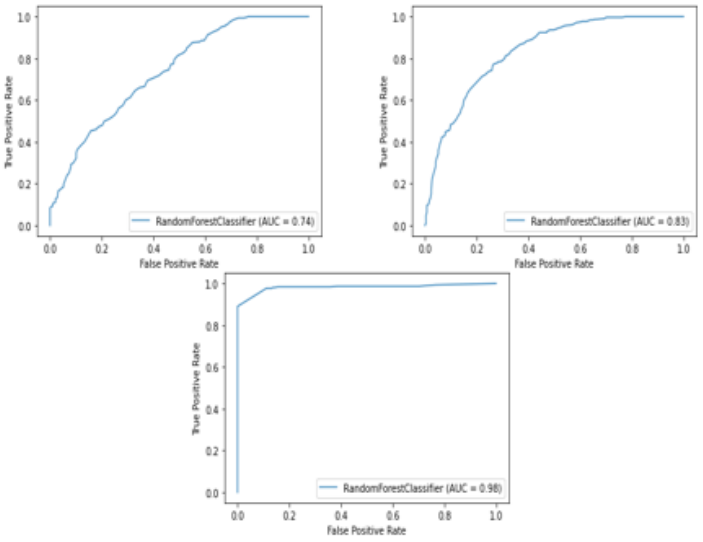

Figure 4. ROC curves corresponding to ATW dataset for 1. 5- and 10-day trading windows

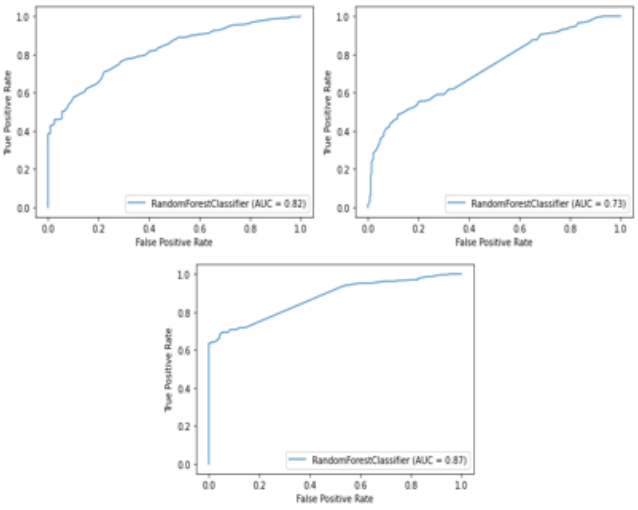

Figure 5. ROC curves corresponding to BOA dataset for 1. 5- and 10-day trading windows

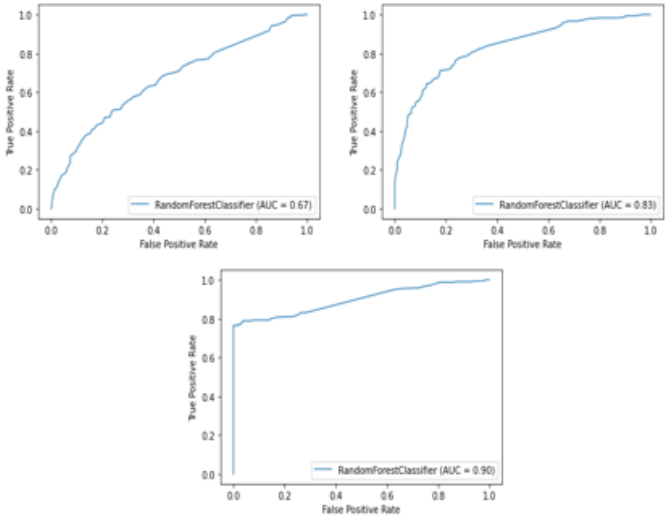

Figure 6. ROC curves corresponding to BCP dataset for 1.5- and 10-day trading windows

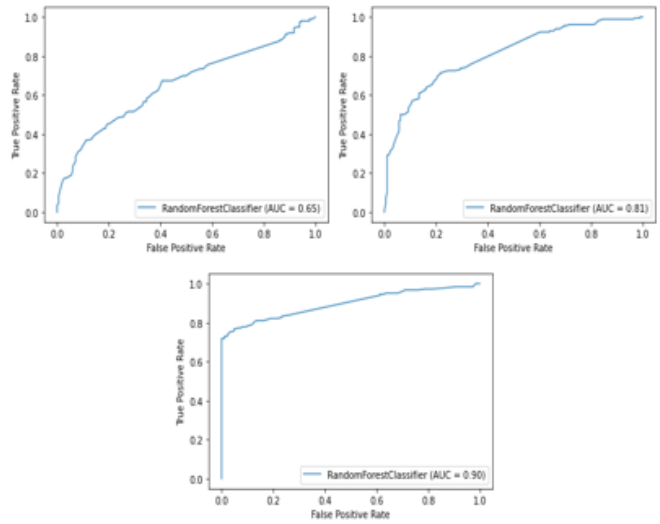

Figure 7. ROC curves corresponding to BMCI dataset for 1. 5- and 10-day trading windows

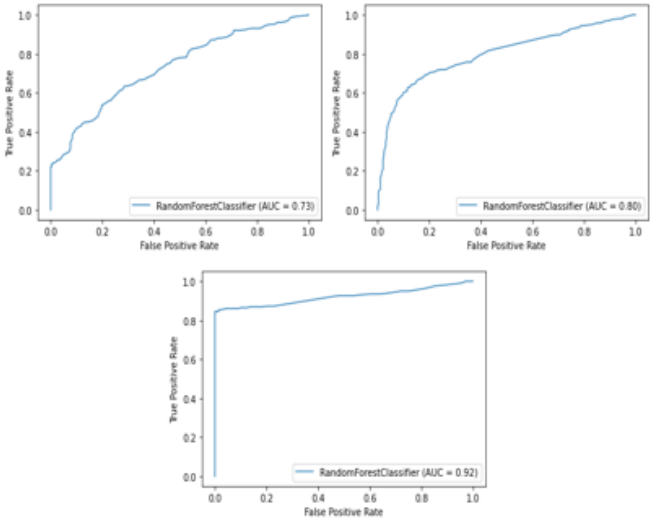

Figure 8. ROC curves corresponding to CIH dataset for 1. 5- and 10-day trading windows

From the ROC curves we can see that the 10 days predictions model produces best results for all stocks. And the AUC is an important tool to evaluate the performance. If it is closer to 1, it means that the classifier model is excellent, and if it is near 0.5 it means that the classifier produces random results. In our case, AUC is above 0.87 for all five datasets for 10 days model. So, this model seems to yield excellent results.

2) Comparison of accuracy with SVM classifier

In this section we compare the accuracy of our model with SVM model. The tables and the graph below compare the two models and show that RF outperforms SVM in the different time windows for all the datasets.

Stock	RF-Accuracy	SVM-Accuracy
ATW	65.74%	55.56%
BOA	74.34%	62.42%
BCP	62.36%	59.48%
BMCI	61.64%	60.05%
CIH	66.12%	59.24%

Table 4. Accuracy Comparison between RF and SVM models for 1-day predictions

Stock	RF-Accuracy	SVM-Accuracy
ATW	74.71%	60.88%
BOA	65.45%	53.99%
BCP	75.90%	57.87%
BMCI	74.20%	53.85%
CIH	75.50%	55.90%

Table 5. Accuracy Comparison between RF and SVM models for 5-days predictions

Stock	RF-Accuracy	SVM-Accuracy
ATW	94.39%	61.51%
BOA	82.00%	57.83%
BCP	86.24%	63.32%
BMCI	85.03%	65.24%
CIH	92.15%	57.66%

Table 6. Accuracy Comparison between RF and SVM models for 10-days predictions

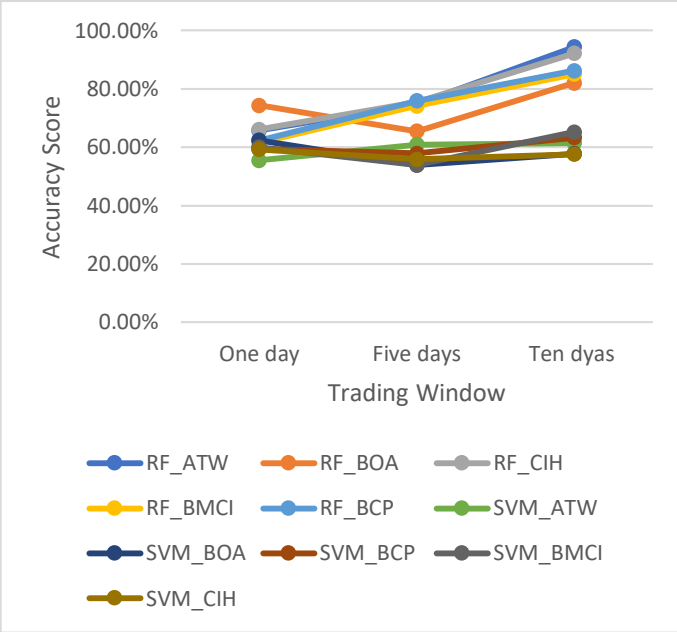

Figure 9. Accuracy Comparison between RF and SVM models

V. Conclusion and future works

The aim of any trader is to predict future stock prices with high precision in order to make more profits. But stock markets are known for their complexity and instability. In fact, developed technologies provide traders with new tools to forecast the market to make more gains and restrict their losses. ML is one of those tools and it proves its effectiveness in financial markets. In this paper, we have used RF classifier to predict price direction and help traders to make the right decision at the right moment. The results prove the robustness of our model. It also seems more robust than SVM model in different trading windows. The robustness of RF model was verified by various metrics like accuracy, precision, recall, F-Score, and AUC. The accuracy of prediction of all the used datasets varied between 61% and 94%, and ROC curves confirm the robustness of the model used.

The RF model might be used by traders to help them confirm the direction of prices of stocks used in this paper, and help them to minimize their risks and losses and to maximize their profits as well.

In future research, we intend to predict stock price movement in short time window and in data. Also, we look forward to comparing multiple models such us K-Nearest Neighbor, DT, Naïve Bayesian, Logit, to name but a few. Last but not least, we intend to discover the ability of Deep Learning in predicting stock price movement.

References

[1] Andrea, P.: Technical analysis and sentiment embeddings for market trend prediction. Expert Systems With Applications, (2019).

[2] Fama E.F. : Random walk in stock market prices. Financial Analysts Journal, (1965b), vol. 21.

[3] Bachelier L. : Théorie de la spéculation. Annales scientifiques de l'École normale supérieure 17 (1900): 21-86, https://doi.org/10.24033/asens.476.

[4] Rodrigo C. Brasileiro, Victor L. F. Souza, Bruno J. T. Fernandes, Adriano L. I. Oliveira.: Automatic Method for Stock Trading Combining Technical Analysis and the Artificial Bee Colony Algorithm. IEEE Congress on Evolutionary Computation June 20-23, Cancun, Mexico, (2013).

[5] P.C. Chang, C.Y. Fan, C.H. Liu, Integrating a piecewise linear representation method and a neural network model for stock trading points prediction, IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews 39 (1) (2009) 80–92.

[6] Giles C.L., Lawrence S., Tsoi A.C.: Noisy time series prediction using a recurrent neural network and grammatical inference, Machine Learning 44 (1–2) (2001) 161–183.

[7] Mallick, V.C.S. Lee, Y.S. Ong.: An empirical study of genetic programming generated trading rules in computerized stock trading service system, in: 5th International Conference Service Systems and Service Management – Exploring Service Dynamics with Science and Innovative Technology, ICSSSM 2008, (2008), pp. 1–6.

[8] Agrawal, M., Khan, A.U., & Shukla, P.K. Stock Price Prediction using Technical Indicators: A Predictive Model using Optimal Deep Learning. International Journal of Recent Technology and Engineering. (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2, (2019).

[9] Sezer, O.B., Ozbayoglu, M.A., & Dogdu, E.: A Deep Neural-Network Based Stock Trading System Based on Evolutionary Optimized Technical Analysis Parameters. Procedia Computer Science, 114, 473-480. (2017).

[10] Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S.R.: Predicting the direction of stock market prices using tree-based classifiers. The North American Journal of Economics and Finance, 47, 552-567. (2019).

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

[12] Khaidem, L., Saha, S., & Dey, S.R.: Predicting the direction of stock market prices using Random Forest. Applied Mathematical Finance. (2016).

[13] Imandoust, S., & Bolandraftar, M. (2014). Forecasting the direction of stock market index movement using three data mining techniques: the case of Tehran Stock Exchange. Journal of Engineering Research and Applications. ISSN : 2248-9622, Vol. 4, Issue 6(Version 2), June (2014), pp.106-117

[14] Shen, S., Jiang, H., & Zhang, T.: Stock Market Forecasting Using Machine Learning Algorithms. (2012).

[15] Tüfekci, Z., & Abul, O. (2020). Distinguishing True and False Buy/Sell Triggers from Financial Technical Indicators. 2020 Innovations in

Intelligent Systems and Applications Conference (ASYU), 1-6.
[16] Ekapure, S., Jiruwala, N., Patnaik, S., & Sengupta, I. (2021). A data-science-driven short-term analysis of Amazon, Apple, Google, and Microsoft stocks. ArXiv, abs/2107.14695.
[17] Tan, Z., Yan, Z., & Zhu, G. (2019). Stock selection with Random Forest: An exploitation of excess return in the Chinese stock market. Heliyon, 5.
[18] Huang, C., Yang, D., & Chuang, Y. (2008). Application of wrapper approach and composite classifier to the stock trend prediction. Expert Syst. Appl., 34, 2870-2878.
[19] Powers, D.M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. ArXiv, abs/2010.16061.