

Stock Returns Prediction with Moving Averages: Insights from VF Corporation Data

Zhexin Zhang *

Qian Huang Senior Middle School, Changzhou, China

* Corresponding Author Email: ccchris1933@gmail.com

Abstract. This study focusses on solving the problem of predicting the stock price in complex market. The traditional MAs have limits in capturing the percentage dynamic changes and non-linear features. By using VF Corp's daily data from 2018 to 2023, this study turns the MAs into relative percentage change indicators. Three models (Linear Regression, Random Forest and Gradient Boosting) are tested which shows all these three models predication power is limited. Linear Regression owns the least negative R^2 which is -0.0074, but its MSE (0.0003) is too low to believe. However, the Random Forest and Gradient Boosting have the problem of overfitting and their Test R^2 is -0.1140 and -0.1158. The mainly limitations are the reliance on the historical price and focusing on one stock price. As the accuracies of three models are bad, it reminds that in the future work, it should contain deep learning and more factors inputs to strengthen the accuracy.

Keywords: Stock return prediction, machine learning, moving average.

1. Introduction

Nowadays as the international relations become conflicted, the stock price shows chaos to predict. In this situation, stock price forecasting plays a critical role in guiding investment decisions while reducing basic risks in investment. Move Averages (MAs) is the basic technical analysis method. And the stock traders use the cross point of the short-term MA (e.g. MA5) and the long-term MA(e.g.MA50) to decide whether to buy or sell [1]. However, there are two important key points undermining their effectiveness. The first one is over-reliance on absolute price. The traditional MAs use raw price data to predict. As a result, it cannot show the moment change of the real-world trading which based on percentage. For example, a \$10 stock increase of \$1(a 10% change) is more significant than a \$100 stock increase of \$1(a 1% change). However, the traditional MAs shows no difference between two conditions. Another one is that stock is unable to capture nonlinearity. The financial markets are influenced by trends, investors' emotions and external events (e.g., national policy, geopolitical news). Therefore, the simple linear model cannot capture the nonlinear model

In the early field of stock forecasting, the previous method was using MAs and hybrid models [2]. Zuo and Kita applied Bayesian networks to P/E ratio prediction and this proves that feature engineering (e.g., clustering) can enhance pattern capture [3]. Pai and Lin propose an Auto - Regressive Integrated Moving Average & Support Vector Machine (ARIMA-SVM) hybrid, showing linear-nonlinear model integration improves forecasts [4]. Shahi et al. compared Long Short - Term Memory & Gated Recurrent Unit (LSTM/GRU), highlighting deep learning's potential for stock data. However, there are gaps to fill. The first one is MA limitations. Most studies use absolute MA values, ignoring percentage which is important in world trading. Another is that the real-world risk influences stock forecasting. There are few studies directly bridging MAs forecasting and risk management (e.g. stop-loss signaling). This study improves that make MAs be relative change indicators (e.g., MA5-pct) to be like realities. Unlike prior work, this study pays more attention to predicting the risks (e.g., using volatile MA changes for risk quantification) than predicting accuracy. By using VF Corp data, this study considers the real-world risks and use MAs to build a better model.

This study turns traditional MAs into relative percentage changes (e.g. MA5-pct) to match real-world trading. This study tested 3 models including Linear Regression, Random Forest and Gradient Boosting by using 5 years of VF Corp data to see how they capture markets patterns. This study focusses on risks and uses MA volatility to find stop-loss chances in a noisy market. In this case of

study, this study contributes to using better methods to fit market behaviors, considering the risks in the market and comparing the accuracy of the models in forecasting.

2. Method

2.1. Dataset Preparation

This study uses daily trading data of VF Corporation (ticker: VFC) from the Yahoo finance API (yfinance library in Python) from January 1, 2018, to December 31, 2023(6 full years) [5], containing 1510 observations (excluding holidays/non-trading days). The target of this study is developing a model to predict the basic trend of the stock price after training by these data [6].

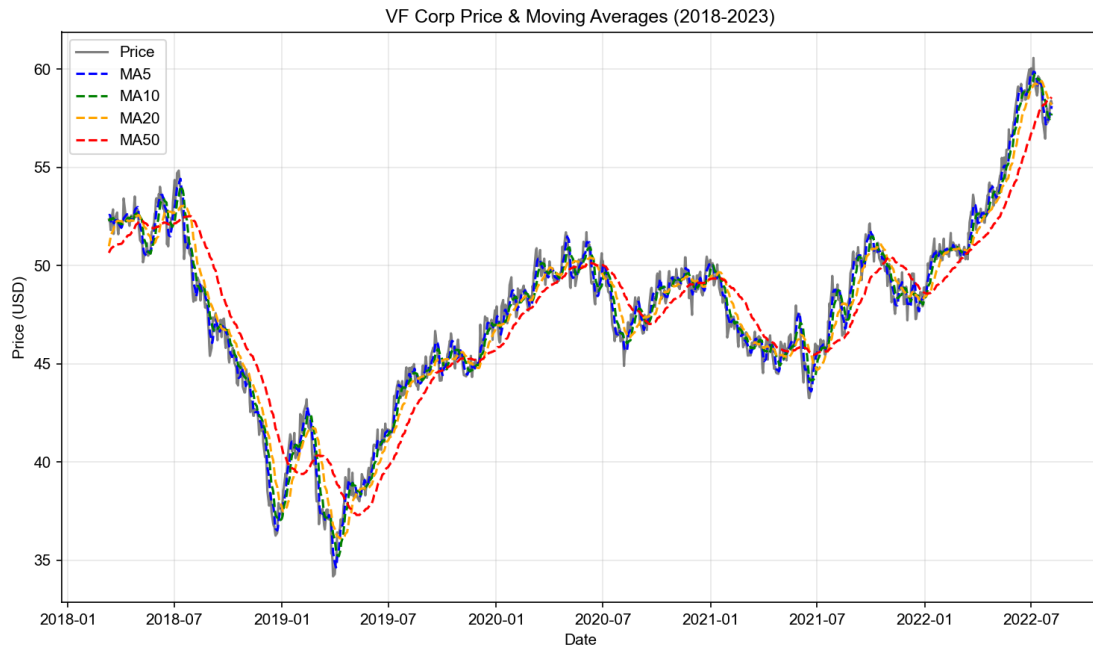


Fig 1. VF Crop Price & Moving Average (2018-2019) (Picture credit: Original)

This Fig. 1 illustrates VF Crop's stock price with MAs (MA5, MA10, MA20, MA50) from 2018 to 2023. After about 2019-4, it shows an increasing but undulating trend. And these smooth MA lines and price trends give numerous continuous and predictable patterns (linear/non-linear). Then the regression like gradient boosting can easily capture these patterns to predict profits. This study uses 70% data to train and 30% data to test to observe the accuracy of the model.

2.2. Machine learning-based prediction models

2.2.1 Linear Regression

Linear regression is a basic statistical method which deals with dependent variables and independent variables [7]. By searching the optimal straight line or hyperplane to minimize the differences between observed values and predicted values. Its mathematical expression is $y = \beta_0 + \sum \beta_i x_i + \epsilon$ where β_i represents coefficients and ϵ is the error term. It divides the data set into train set and test set with using ordinary least squares (OLS) to minimize residuals and use R^2 to evaluate the result. Its advantages are easy to predict and expand and fast to calculate. But it also has limits like it is hard to deal with non-linear model and its performance isn't good in complex data sets.

2.2.2 Random Forest

Random Forest is an ensemble learning algorithm which builds many decision trees while training to enhance the accuracy of the prediction [7]. It gets samples from the original data set and trains a decision tree on each sample, with splitting subsets to ensure diversity. Its workflow includes creating bootstrap samples, training trees to minimize impurity and aggregating results. It can deal with high-

dimensional data without normalization and capture key features of the non-linear data. It also provides importance scores to help to understand variables drive predictions. Although it calculates more difficultly than linear regression, it can contain performance and stability. So, it is a better model to deal with non-linear model than linear regression.

2.2.3 Gradient Boosting

Gradient Boosting is an ensemble machine learning method that builds a powerful predictive model by sequentially combining multiple weak models [8]. It starts with an initial prediction (e.g., the target variable's mean) and calculates residuals—the differences between actual and predicted values. Each subsequent model is trained specifically to predict these residuals. The predictions from new models are scaled by a learning rate (e.g., 0.01–0.3) and added to the ensemble to incrementally improve accuracy. This iterative process continues until meeting a stopping criterion, such as reaching a maximum number of trees. The algorithm minimizes a loss function (e.g., mean squared error for regression) using gradient descent, allowing it to model complex linear and nonlinear patterns effectively while handling noisy data robustly. However, it requires careful tuning of hyperparameters (e.g., tree depth, learning rate) to avoid overfitting and is computationally more intensive than parallel methods like Random Forest.

3. Results and Discussion

Table 1. Test results of three models

	CV R^2	Test R^2	Test MSE
Linear Regression	-0.0122	-0.0074	0.0003
Random Forest	-0.0732	-0.1140	0.0004
Gradient Boosting	-0.1283	-0.1158	0.0004

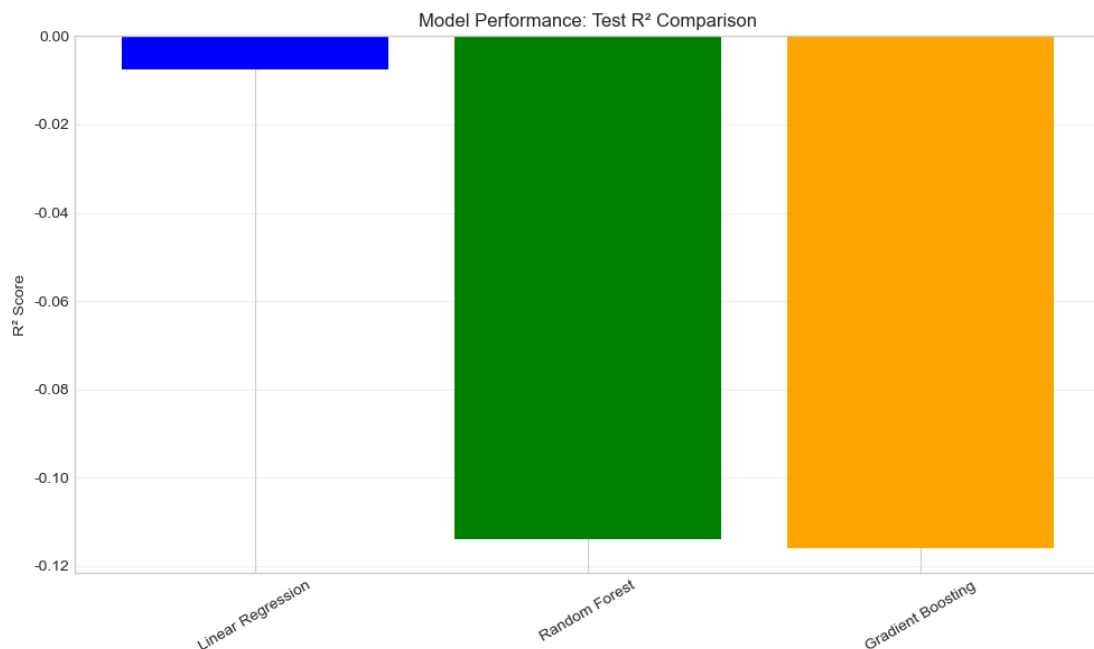


Fig 2. The test R^2 of three models (Picture credit: Original)

This study uses R^2 shown in Fig. 2 and MSE to test three models (Linear Regression, Random Forest, and Gradient Boosting) performances in VF Corporation historical data price and MAs. The Table 1 shows that Linear Regression achieved a Test R^2 of -0.0074 with the smallest Test MSE (0.0003), while Random Forest and Gradient Boosting exhibited more negative Test R^2 values (-0.1140 and -0.1158, respectively) and slightly higher MSE (0.0004). The visualization of the Test R^2 shows these trends in addition way which shows linear regression model's R^2 close to zero

where expanding its explanatory power is weak. Although tree-based models are designed to capture complex data, its test result is worse than linear regression. These results align with the Efficient Market Hypothesis (EMH) [9], which consider that stock prices reflect all the information and make future benefits become semi-random. Because of emotion of investors, liquidity shocks and other factors, the market has become random. This explained why the traditional models become useless in prediction. The wave of short-term prices is influenced by unpredictable events which lead R^2 be negative. In this situation, no models can be beyond naive mean prediction. Moreover, tree-based models may face the risks of overfitting in high chaos situations which test R^2 is more smaller meaning that they ‘remember’ the chaos (short-term waves), instead of concluding the normal trends. Gradient Boosting make this issue worse. All models have their unique advantages and disadvantages. Though Linear Regression is easy to understand and explant, its performance is bad in capturing the nonlinear interactions between MAs and benefits. Its MSE is small because of the results are near the averages, so it’s a stable but useless baseline. Random Forest and Gradient Boosting are design for capturing the complex patterns, but they show overfitting in noisy finance data which limits their actual value. This study has obvious limits. This study only depends on historical price and MAs, with ignoring the news sentiment, macroeconomic indicators which influence the actual profits. Using one simple stock (VF Crop) also limits the results in other market’s generalizability. Moreover, all the models are based on the unreliable hypothesis (historical pattern will repeat), however in finance area these hypotheses are often destroyed by some events like black swan events which undermine the predictive clout of the model. This study’s future work should be bridged with deep learning like Recurrent Neural Networks (RNNs) or Long Short – Term Memory (LSTMs), this kind of model preformat well in capture sequential pattern of time -series data. Containing words emotions, inflation rates, interest rates these factors can enhance features and capture non-price return drivers. The hybrid frameworks which mix stability of the linear regression and ability of capturing non-linear of tree-based models can also enhance the robustness of the model. All in all, the test models’ prediction power is weak which reflects the stock returns’ semi-random nature. Tree-based models face the risks of overfitting; however linear regression’s low R^2 result has misleading. These findings show the issues of traditional models in stock prediction and also highlight that needing advanced methods to face markets’ complexity to build the future work.

4. Conclusion

In this study, MAs is transformed into relative change indicators, with testing Linear Regression, Random Forest, Grand Boosting three models by using VF Corp data. The result illustrates that the prediction ability in all the three models is limited. The Linear Regression has the lowest R^2 but the MSE of the tree-based model is too low to believe which shows the over-fitting problems. This study’s limitations include the reliance on the historical price and focusing on one stock. In the future study, this work will use deep learning (LSTM) and multi-factors inputs to keep the accuracy.

References

- [1] Investopedia. Moving average (MA). Investopedia; 2025. Available from: <https://www.investopedia.com/terms/m/movingaverage.asp>
- [2] Zuo Y, Kita E. Stock price forecast using Bayesian network. *Expert Syst Appl*. 2012;39(8):6729–37.
- [3] Pai P-F, Lin C-S. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*. 2005;33(6):497–505.
- [4] Shahi TB, Shrestha A, Neupane A, Guo W. Stock price forecasting with deep learning: A comparative study. *Mathematics*. 2020;8(9):1441.
- [5] Yahoo Finance. VF Corporation (VFC) Historical Data. Yahoo Finance; 2025. Available from: <https://finance.yahoo.com/quote/VFC/history/>
- [6] Hunter JD. Matplotlib: A 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90–5.

- [7] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785–94.
- [8] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12(Oct):2825–30.
- [9] Encyclopædia Britannica. Efficient market hypothesis (EMH). Encyclopædia Britannica; 2025. Available from: <https://www.britannica.com/topic/efficient-market-hypothesis>