Stock Return Prediction Using Machine Learning Regression Models

Ning Deng

{Z5532838@ad.unsw.edu.au}

School of Computer Science and Engineering, The University of New South Wales, Sydney, New South Wales, Australia

Abstract. This study analyses the hypothesis of whether machine learning regression models can forecast the stock returns of stocks in the short term on the basis of historical market data. An integrated pipeline of time-series is run to facilitate uniform preprocessing, construction of features and evaluation of models. The dataset has 615, 505 observations that can be used representing 505 U. S. stocks between 2013 and 2018. The input features are lagged daily returns over the last five trading days to forecast trading day returns. Experimental results indicate that linear regression has an MSE of 2.17×10^{-4}. Decision tree regression achieves an MSE of 2.19×10^{-4}. The neural network model performs the worst among the three, with an MSE of 3.07×10^{-4}. These results suggest that prediction of short-term returns is very difficult and that more complex model does not necessarily enhance better predictive accuracy.

Keywords: Short-term stock return prediction; Machine learning regression; Financial time series; Model comparison; Historical price data.

1 Introduction

Financial markets are marked with inconsistent and recurring price fluctuations as well as high degree of uncertainty. The prices of stocks are determined by the performance of companies, macroeconomics, adjustments of policies and investor behaviour. The interaction of these factors is complex and it is hard to predict them. Meanwhile, prediction of stock returns in the short run is significant to stock trading, determination of investment in a portfolio, and control of risks. The series of financial returns, however, tend to have poor signal-to-noise ratios and negative dynamics, restraining their predictive powers [1].

Early studies under this field adopted mainly statistical and econometric studies. The support vectors machines and linear regression were extensively used in the financial time-series problem forecasting [2]. Such methods usually presuppose constant statistical characteristics. Such assumptions are not always true in volatile market times as this minimizes predictive reliability.

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365163

As computing methods improved, machine learning algorithms have been presented in the financial forecasting processes. The literature findings indicate that machine learning techniques have the ability to reveal non-linearities in market data and can potentially outperform conventional models in some situations [3]. Sequential financial data has also been modelled using deep learning archtectures to learn about temporal dependencies [4]. However, the empirical evidence is divided especially in short run returns prediction environs.

Consensus with regard to systematic comparison under homogenous experimental settings is still required in light of these limitations. The paper uses a unified modelling pipeline to predict the stock returns in the short term as per the historical stock price. Multiple regression models are developed and optimized and compared to each other on the same preprocessing, feature construction, and evaluation terms. The aim is to test whether there is any significant improvement in the short-horizon return forecasting with higher complexity of the model.

2 Dataset description

The data employed in this work is taken at Kaggle and includes the stock price data of 505 publicly hosted U. S. companies per day. The period covered is between 2013 and 2018 inclusive trading date, opening price, highest price, lowest price, closing price, trading volume, and stock identifier details. There are 619, 040 daily observations (of valid data). However, in most stocks, there are about 1, 259 trading days of which a complete trading is five years. There are just several records held in fewer number of stocks because of later listing dates. Preprocessing is necessary to remove missing values and non-trading days in order to achieve uniformity. Where it is applicable, trading dates are matched.

2.1 Data statistics and exploratory data analysis

The prediction target is obtained as daily returns based on closing prices. Mean daily return of all stocks is nearly zero and the standard deviation is relatively greater. This finding suggests that short-run series of returns have a low predictive content as compared to noise.

Before training stock returns, a statistical analysis is undertaken to explore the characteristics of the stock returns in order to identify the statistical characteristics. In figure 1, the distribution of daily returns is provided in all the stocks. The distribution is also very concentrated around zero with majority of the returns not large in magnitude. Simultaneously, both sides of the distribution visibly have a few extreme values. The above pattern is a direct indication of heavy-tailed characteristics of financial returns data as large price changes are quite uncommon, yet they cannot be neglected.

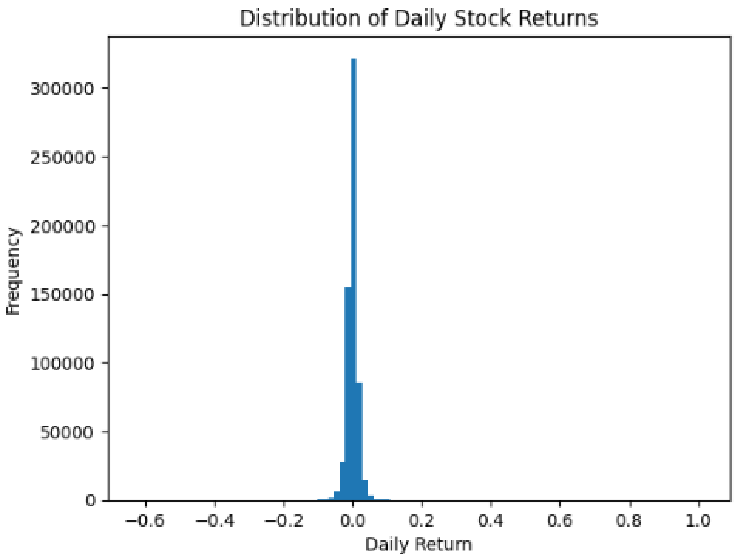

Fig. 1. Distribution of Daily Stock Returns (Picture credit: Original)

Autocorrelation function of lags of daily returns of the twenty trading days is explained in figure 2. The coefficients of autocorrelations are near to zero at majority of the lags, which implies that the dependence of current returns and the one of the past is weak. This property indicates that there is limited short-term linear predictability and this goes to explain why it is difficult to forecast short-horizon returns.

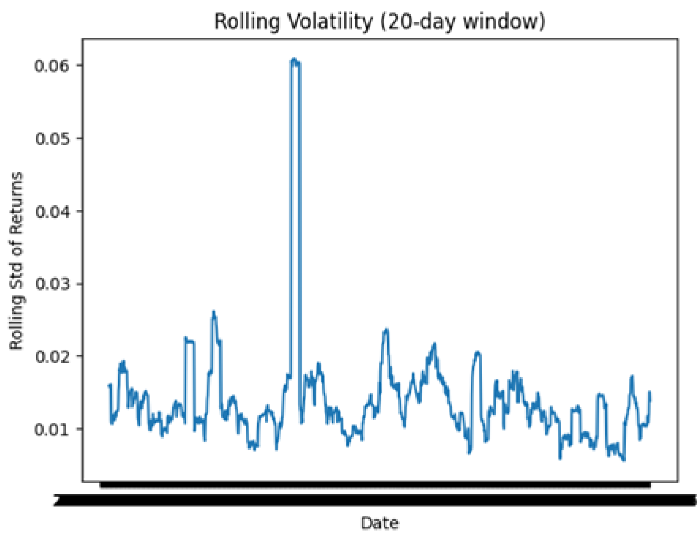

Fig. 2. Autocorrelation of Daily Stock Returns (Picture credit: Original)

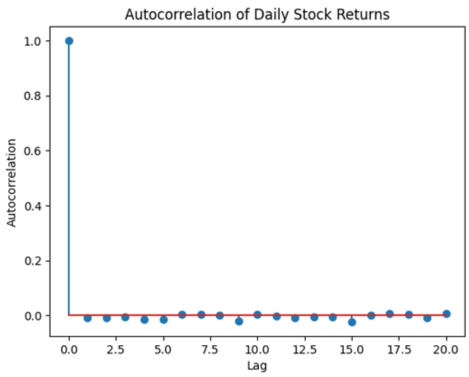

Fig. 3. Rolling Volatility of Daily Returns (Picture credit: Original)

Use of rolling standard deviation of daily returns is presented with respect to the short term volatility which is depicted in figure 3. There is apparent clustering behaviour in the volatility series. Episodes of low volatility are likely to continue and high volatility also exists in groupings. This trend is highly recorded in financial time series and shows the fluctuation of the uncertainty in the market with time.

The big market movements and price irregularities are usually linked to high-volatility periods. Return dynamics are less predictable during such periods making it harder to predict. In this observation, the element of robustness in the analysis of the model performance in various conditions of the market becomes essential.

A heatmap of laggings of the (return) variables is built to study the possibility of correlation redundancy. The figure 4 depicts the correlation between input features. The majority of the pairwise correlations are not too high, which means that multicollinearity is not too high. This finding indicates that lagged returns contain complementary channel short-term information, in contrast to redundancy.

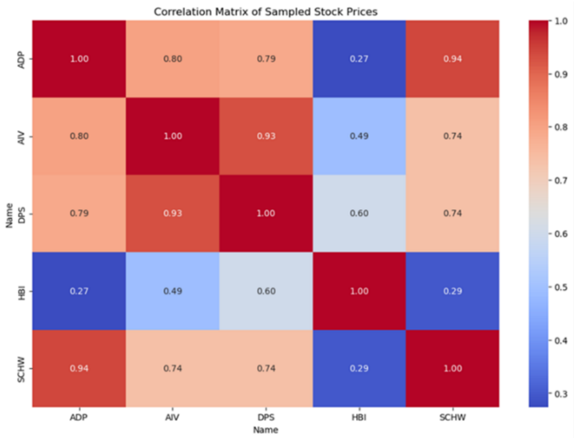

Fig. 4. Correlation matrix of sampled stock prices (Picture credit: Original)

2.2 Return representation

This research is a prediction of the daily returns based on stock returns, as opposed to unfiltered stock prices. A return is used to measure changes of relative prices between two trading days in succession and gives a scale-invariant measure of a price movement. The comparison of stocks at varying levels is made possible with the use of returns and the effects of long-term trends are mitigated.

The daily return at time $t$is defined as

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{1}$$

P_t is used in Equation (1) to refer to the closing price at time t. Such a representation using a returns-based reducing effect of long-run upward trends due to inflation and economic expansion and helps the model to emphasize the short-term price movement patterns.

2.3 Feature engineering

The inputs of the model are lagged return features. In particular, the input variables include returns on the last one to five trading days. This design of design captures short term momentum and reversal effects of financial market studies. In stock prediction activities, feature-weighted methods have also been studied [5].

Normalization of features is used to normalize the input variables to the equal numeric ranges. This is a preprocessing stage that helps make training more stable and provides better optimization. There are no external sources of information (news sentiment or macroeconomic indicator). This design decision restricts the variety of information however, the analysis will still be limited to price-related indicators. This design is concerned with price-based signals in which heterogeneous data sources are not incorporated [6].

3 Methodology

In this paper, three types of regression are compared, namely, linear regression, decision tree regression, and neural network regression. The models are chosen in order to give the higher levels of model complexity and versatility. The time series forecasting task in finance has received a considerable amount of research on deep learning [7]. Recent studies have also systematically summarized applications of deep learning in the stock market prediction [8].

The linear regression is used as bench-mark. It presupposes a linear correlation between input characteristics and the target variable. This assumption enhances increased explanatability and a fixed point, but deters the non-linearity of the model in explaining financial return data.

Decision tree regression presents non-linear decision boundaries by splitting recursively the feature space according to input thresholds. It is a structure that gives the opportunity of the interaction effects between the lagged return features. In order to minimize the risk of overfitting, the depth of the maximum tree is limited, thus enhancing generalization and minimizing noise sensitivity tendencies that are typical of financial time series.

The Feed-forward architecture used in the neural network regression has several hidden layers. There have been systematic reviews of deep learning architectures in more recent forecasting works [9]. Complex lagged returns versus future returns relationships are approximated by the use of non-linear activation functions. Early stopping is used in the training to avoid overfitting and regulating overfitting behaviour.

4 Experiment

4.1 Experiment setup

The experiments are conducted using Python 3.9. The scikit-learn library is a dependency of model implementation. It is calculated using a personal computer in an Intel Core i7 processor with 16GB. The default hyperParameters are accepted unless otherwise. The decision tree will have a maximum depth of 5. The neural network applies the ReLU activation functions, two hidden layers, 64, and 32 neurons respectively. Early termination is facilitated in order to avoid overfitting.

4.2 Results and discussion

Table 1. Model Performance Comparison.

Model	MSE	R^2
Linear Regression	2.166×10^{-4}	-0.0011
Decision Tree	2.192×10^{-4}	-0.0128
Neural Network(MLP)	3.069×10^{-4}	-0.4181

The three regression models have out-of-sample performances that are shown in Table 1. The linear regression has a value of MSE of 2.17 x 10 -4 and R 2 of -0.0011. The decision tree regression results in an MSE of 2.19 x 10 -4 and an R 2 of -0.0128, which shows a somewhat worse performance of generalization. The neural network model does not do any better with a MSE of 3.07x10 -4 and an R 2 of -.4181.

The negative values of the R 2 of all the models show that all the models fail to explain the variation in returns better than a simple mean benchmark. Despite the fact that more flexible models can in principle describe nonlinear patterns, short run stock returns seem to have very weak predictive characteristics. A more complicated model is not better in such an environment. Rather, increased flexibility can increase noise, and give rise to overfitting.

Such results can be considered in agreement with other earlier empirical studies. Greater parallelisms between machine learning models and stock prediction have been demonstrated in other previous empirical studies [10]. As Fischer and Krauss demonstrate, even deep learning models, under the impression that they could work well in-sample, are prone to system changes in the financial markets [4]. Gu et al. also report that the machine learning models that are used to represent the asset pricing issues are unstable a lot of times because of the noisy and weak predictive signals in the asset returns data [1].

The volatility in the market also impacts the quality in prediction. The errors of prediction are expected to rise during times of high volatility when the returns dynamics are less predictable.

Conversely, the peaceful times in the market carry less prediction errors. The trend suggests that performance can be predicted based heavily upon the prevailing market conditions and may be taken in at face value.

5 Conclusion

This paper analyzes the issue of short term stock returns forecasting over historical price information. The short-horizon predictions are not an easy one to undertake due to low signal-noise-ratios, market dynamics and interplay of factors affecting the market, which therefore are complex.

A single modelling path is built such that data preprocessing, feature building and evaluation are applied uniformly among models. Three regression models with various degrees of complexity are assessed using lagged return features and they include the linear regression, decision tree regression and neural network regression. This design makes it possible to examine model behaviour systematically in the same experimental situations, as opposed to examining it simply in terms of predictive accuracy.

The exploratory data analysis indicates various well-known properties of financial returns data that include heavy-tailed distribution, low autocorrelation, and volatility concentration. Those properties point to the natural inability to make accurate predictions based on the short-term returns and present valuable context about the meaning of the results obtained in the experiment. The results of empirical experiments indicate that linear regression can provide quite consistent albeit weak performance, whereas more complex models can represent richer patterns at the price of higher variance. The observation on this trade-off between the complexity and robustness of the model is common in the literature that has been tested and completed.

There are a number of shortcomings in this research. The analysis makes use of the price based information that it does not use any external factors like the macroeconomic indicators or market sentiment. Also, hyperparameter optimization is also purposefully constrained, and this could confine the performance of more sophisticated models. Future studies can examine a more complex set of features, different model architectures, and more sophisticated training approaches to enhance their robustness in different market conditions.

References

[1] Gu, S., Kelly, B., & Xiu, D.: Empirical asset pricing via machine learning. The Review of Financial Studies 33(5), 2223–2273 (2020)

[2] Kim, K.J.: Financial time series forecasting using support vector machines. Neurocomputing 55(1-2), 307–319 (2003)

[3] Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R.: Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications 42(20), 7046–7056 (2015)

[4] Fischer, T., & Krauss, C.: Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research 270(2), 654–669 (2018)

[5] Chen, Y., & Hao, Y.: A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. Expert Systems with Applications 80, 340–355 (2017)
[6] Patel, J., Shah, S., Thakkar, P., & Kotecha, K.: Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications 42(4), 2162–2172 (2015)
[7] Sezer, O.B., Gudelek, M.U., & Ozbayoglu, A.M.: Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing 90, 106181 (2020)
[8] Jiang, W.: Applications of deep learning in stock market prediction: recent progress. Expert Systems with Applications 184, 115537 (2021)
[9] Lim, B., & Zohren, S.: Time-series forecasting with deep learning: a survey. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 379(2194) (2021)
[10] Vijh, M., Chandola, D., Tikkiwal, V.A., & Kumar, A.: Stock closing price prediction using machine learning techniques. Procedia Computer Science 167, 599–606 (2020)