International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Improved Analysis of Stock Market Prediction

Vani N, Rohit H[1], Parvathi C R[2], Anusha G S[3], Tarun P[4]

Vani N, Assistant professor, BGS Institute of Technology
[1]Rohit H, Department of Computer Science and Engineering, BGS Institute of Technology
[2]Parvathi C R, Department of Computer Science and Engineering, BGS Institute of Technology
[3]Anusha G S, Department of Computer Science and Engineering, BGS Institute of Technology
[4]Tarun P, Department of Computer Science and Engineering, BGS Institute of Technology

---

Abstract – Predicting the financial value of the stocks of the company is the aim of this system. We use machine learning methodologies to predict the stock values. A machine learning model called Random Forest Regressor is used. Factors that are considered for evaluation are open, low, high, close, adjacent close and volume of the company.

Key Words: machine learning, random forest regressor,.

1. INTRODUCTION

Stock market forecasting is an act of analysis and an attempt to diagnose the marketability of a company's shares. It serves as a reliable tool for the financial growth of companies traded on the exchange. Commerce and Trade are two economic elements that play an important role in the development of the national economy in a wide range of areas such as industrial markets and investors. This process makes it easy to predict stock prices in the event of a price increase or decrease at any time. The top players are the investors and industries involved in this share exchange process for securities. Behind this lies the concept of pure supply and supply of economic policy. For example, if the demand for shares of a particular company declines, the price of that particular company's stock will always fall for any period.

The Efficient Market prediction is a technical theory, and the experimental challenge is the motivation for achieving efficient results in marketing. The stock price fully reflects fluctuating information about the component and opportunities to make a sufficient profit. Today, many industries and companies are involved in this process daily. This contains a huge data set that is difficult to extract, analyze, and extract information. This is a big task for users of manual processes. Stock market patterns and forecasts are revealed by analysis of the stock market using machine learning models

2. RELATED WORK

A literature review found that the application of machine learning technology to stock market forecasts is being thoroughly implemented. Machine learning techniques have proven to be much more precise and faster than the latest prediction techniques.

Evidence of the work on advancements of machine learning in the field of stock market prediction was done by M. Usmani, S. H. Adil, and S. S. A. Ali and K. Raza [1]. Tae Kyun Lee, Joon Hyung Cho, Deuk SinKwon and So Young Sohn [2] have experimented and stated that the Random forest model gave 54.12 accuracies. K. V. Sujatha and S. M.Sundaram [3] provide insightful techniques for dealing with unusual situations that occur frequently during the functioning of the system and can cause confusion and lead to inaccurate predictions. K. A. Althelaya, E. M. El-Alfy, and S. Mohammed [4] contributed further in this area by conducting experiments and simulations to assess the potential of applying deep learning techniques to predict stock prices. E. Chong, C. Han, and F. C. Park [5] leverage the capabilities of deep neural networks to extract information from stock return time series without relying on predictor prior knowledge, using a deep functional learning-based stock market forecasting model from the data. We extracted abstract features and tested the data. Frequently from the Korean stock market. LSTM networks suitable for learning temporal patterns have been widely used for various time series analysis tasks [6]. LSTMs are preferred over traditional RNNs because they overcome the vanishing (or explosion) gradient problem and can effectively learn long-term dependencies through memory cells and gates. Liu, G. Liao, and Y. Ding [7] performed similar work and designed a model for applying LSTMs to equity forecasts with sufficient scope to improve forecast accuracy.

3. METHODOLOGY

This project requires a numpy library to work with arrays and a pandas library to parse data from a particular dataset. It also uses the matplotlib library to plot the output graph and sklearn.model_selection to split the dataset into. Test and training dataset. sklearn also provides a machine learning model called Random Forest Regressor. It also provides a matrix for implementing multiple losses, scores, and utility functions to measure classification performance.

We used a machine learning model called Random Forest Regressor to predict stock prices. This model gets at least a year of input dataset of the previous stock price for each stock. Pandas is used to read the provided CSV file and parse the provided dataset. Then remove the unwanted data that is present in the dataset so that the dataset is clean and clear to the machine learning model. Then divide the total data into 20% training data and 80% test data set. You now have two datasets, training data, and test data. Training data is used to train machine learning models, and test datasets are used to test machine learning models. After splitting the data, fit this dataset to the machine learning model Random Forest Regressor. This model provides predictive data to be compared with the test data and plots the test data and the graph of the machine learning model output data.

A. Random Forest Regressor

The Random Forest Regressor means to data estimators. It adapts the numerical determination structure to the various subsamples of the given data. It controls overfitting. The prediction accuracy is improved Algorithm.

Step 1: From the dataset pick N random records.
Step 2: Based on N records, build a decision tree.
Step 3a: From your algorithm, choose the number of trees and repeat steps 1 and 2.
Step 3b: In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output).

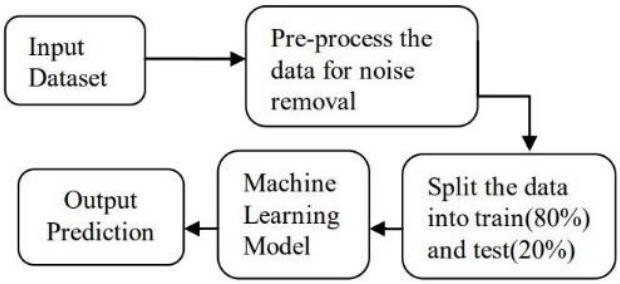

Fig -1: Numerical dataset Algorithm Architecture

B. Dataset

The information required for this project is historical data. This includes the date of each trading day, opening price, loss price, high price, low price, and trading volume. Traders use this data to measure stock volatility.

Table -1: Dataset table

Features	Meaning
Date	The date of the stock value.
Open	Open value of the stock at the start time of the trading day.
High	The highest value of the stock, on the trading day
Low	The lowest value of the stock, on the trading day
Close	The closing value of the stock, on the trading day
Adj Close	Closing price after distributing the dividends of stock's value
Volume	Quantity of traded stocks in the market over a period

Table -2: Sample data from the dataset collection.

Date	Open	High	Low	Close	Adj Close	Volume
2020-09-30	9.07	9.23	9.03	9.09	9.09	1099300
2020-10-01	9.13	9.16	8.98	9.15	9.15	738700
2020-10-02	8.96	9.09	8.93	9.07	9.07	822300
2020-10-05	9.11	9.26	9.1	9.2	9.2	703200
2020-10-06	9.83	9.96	9.63	9.7	9.7	2148800
2020-10-07	9.63	9.79	9.6	9.73	9.73	1308800
2020-10-08	9.65	9.67	9.59	9.67	9.67	1134100
2020-10-09	9.55	9.6	9.45	9.58	9.58	1004600
2020-10-12	9.33	9.4	9.29	9.34	9.34	816100
2020-10-13	9.2	9.2	9	9.08	9.08	1094900
2020-10-14	8.95	9.07	8.92	8.92	8.92	1036100

In the above Table 2 shows the samples of data collected from the collection the dataset. It contains columns such as date, open, high, low, close, adjacent close, and volume.

C. Envisage The Data

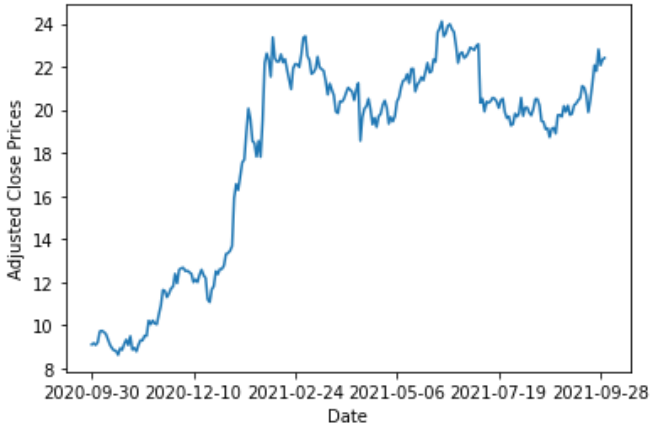

Fig -2: Visualizing the fetched data

Stock data is loaded into a data frame and converted into a CSV file (comma separate value). We plot the line chart of the adjusted close prices over time. The graph shows the data fetched from 30th September 2020 to 28th September 2021.

D. Data Pre-Processing

This step is the most important part of this project. Data preprocessing is a procedure performed to prepare data for a machine learning model. Preprocessing involves transforming raw data into a format that the model can accept and process. This project aims to have a dataset that the model can accept and the algorithm can understand. The value may be missing from the dataset and the information may be verbose, irrelevant, or noisy. Data cleaning is a form of preprocessing that involves removing missing or inconsistent values and changing the index. The same applies to feature selection, hyperparameter adjustment, and data standardization.

I. Feature Selection

Here, the x and y features are selected to retrieve the model's dataset. x and y features are declared for training and test datasets.

Feature selection is one of the essential concepts of machine learning applications that have a tremendous impact on the model's performance. In future selection, not every column is necessary. These features selected have an impact and

contribute to the prediction output. Unnecessary features decrease the general performance of the test set. A method of selecting futures is finding out the most important features, feature importance. Sklearn has feature importance and a feature selector module that can be implemented. With the feature importance module, a score is given to each feature in the data. Features with the highest scores are the most relevant, and sound output variables are guaranteed. The benefits of using feature selection include improving accuracy, reducing overfitting, reducing training time, and improving data visualization. The more features there are, the model is likely to suffer from overfitting.

II. Divide Into Training and Test Datasets

The dataset should be split into a training dataset and a test dataset before modeling.

Training set: A subset of the dataset is used to build and fit predictive models. Training datasets are generated by creating training dataset scripts that generate training dataset functions from input options and raw stock price data. The data is sent to the model for training. The model learns from this data and drives the train set.

Test set: A subset of the dataset for assessing the future performance of the model. This is a good standard for evaluating a model. The test set is used against the predicted dataset to test the trained model. The model has not seen this part of the set. Used for evaluation purposes.

III. Feature Scaling

This is called data standardization. Sklearn has a feature called the standard scaler that is used to standardize datasets. Standardization is known to improve the numerical stability of the model and improve training speed.

IV. Tuning of Parameters

The parameters are model settings. It is important to adjust the parameters to optimize performance. Set up after training and testing the dataset and before adjusting and predicting. Parameters solve the main problem of machine learning, which is overfitting. This project used random search cross-validation.

For a random forest regression model, the best parameters to consider are:

- n_estimators — number of trees in the forest
- max_depth — maximum depth in a tree
- min_samples_split — minimum number of data points before the sample is split
- min_samples_leaf — minimum number of leaf nodes that are required to be sampled
- bootstrap — sampling for data points, true or false
- random_state — generated random numbers for the random forest

V. Model Application and Prediction

The dataset model is now ready. The first step is to select a random state value and build a tree based on the number of random states. Random forests prevent overfitting by creating random subsets of features and using those subsets to create small trees. To build a random forest, you need to train your data. In addition, the parameters from the parameter adjustments are applied here.

VI. Statistical indicators

Root-mean-square error (RMSE) is the standard deviation (prediction error) of the residuals. The residual measures how far the data points are from the regression line. Mean Absolute Error (MEA) measures the average size of a series of prediction errors without considering the direction of the prediction. Mean squared error (MSE) is the sum of absolute error values. The mean squared error also determines the performance of the model. In this case, errors larger than MAE errors are common. The lower the MSE value, the higher the prediction accuracy.

3. EXPERIMENTAL RESULTS

A. One-Year Prediction

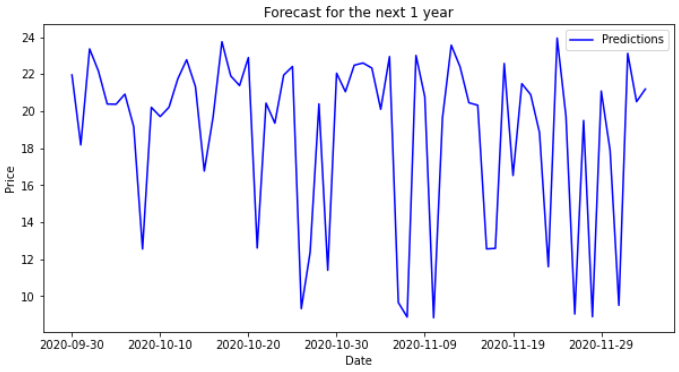

Fig -3: Prediction for next year.

B. One-Month Prediction

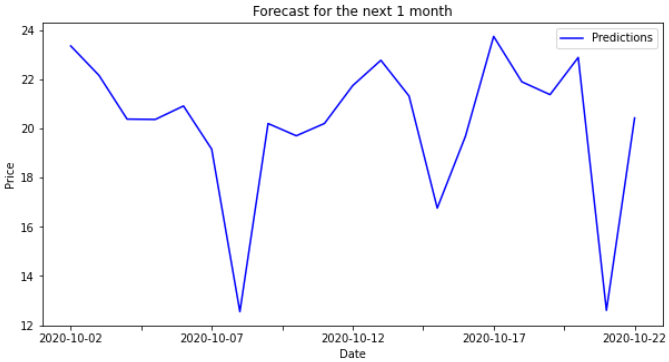

Fig -4: Prediction for next month.

C. Five Days Prediction

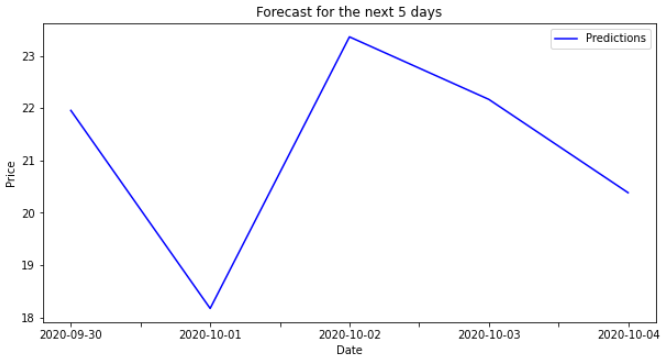

Fig -5: Prediction for the next five days.

4. CONCLUSIONS

The performance ranges from mathematical analysis for prediction to sentiment analysis, financial news articles, and expert ratings. However, the stock market is so volatile that there is no perfect and accurate forecasting system. This algorithm is also suitable for those who need to develop models quickly. It provides a pretty good indicator of the importance it puts into your function. Overall, Random Forests are mostly fast, easy, and flexible. With the help of the Python environment, large data can be processed very efficiently without changing the methods of existing procedures. Experiments and results are obtained using numeric data in Python. Future work may continue using the deep learning algorithm.

REFERENCES

1. M. Usmani, S. H. Adil, K. Raza, and S. S. A. Ali, "Stock market prediction using machine learning techniques," 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2016, pp. 322-327
2. Tae Kyun Lee et al, ”Global Stock Market Investment Strategies Based On Financial Network Indicators Using Machine Learning Techniques”, Volume: 117, Issue: 1, Date: 2019.
3. K. V. Sujatha and S. M. Sundaram, "Stock index prediction using regression and neural network models under nonnormal conditions," INTERACT-2010, Chennai, 2010, pp. 59-63
4. K. A. Althelaya, E. M. El-Alfy and S. Mohammed, "Evaluation of bidirectional LSTM for short-and long-term stock market prediction," 2018 9th International Conference on Information and Communication Systems (ICICS), Irbid, 2018, pp. 151-156
5. E. Chong, C. Han, and F. C. Park, “Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies,” Expert Systems with Applications, vol. 83, pp. 187–205, 2017.
6. D. H. D. Nguyen, L. P. Tran, and V. Nguyen, “Predicting stock prices using dynamic lstm models,” in International Conference on Applied Informatics. Springer, 2019, pp. 199–212.
7. S. Liu, G. Liao, and Y. Ding, "Stock transaction prediction modeling and analysis based on LSTM," 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, 2018, pp. 2787-279