Predicting the Direction of Stock Market Prices using Random Forest

Venkat sai Dhushetty

Abstract— **In this paper, we used the historical data and predicted future price of a stock using machine learning algorithm known as ensemble learning. Ensemble learning is a method for solving a specific computational intelligence problem by carefully creating and combining several models, such as classifiers or experts. The main goal of ensemble learning is to increase the classification, prediction, function approximation, etc., The stock price movement was treated as a function of time series and solved as a regression problem. The learning algorithms we used in this paper is LSTM, SVM and Random Forest.**

Keywords— random forest, LSTM, ensemble learning, support vector machines, stock market prediction

I. INTRODUCTION

Predicting stock market prices comes with many uncertainties and various variables that influence the market value on a particular day, such as economic conditions, investors sentiments towards a particular company, political events etc., resulting in random fluctuations in the stock price and a lot of investors feel investments in the share market comes with high risk. Therefore, many researchers have attempted to predict the stock market price change using Machine Learning and Artificial Intelligence Algorithms which provides accurate results. Thus, getting highly accurate results is the most challenging task and thus motivated us to research on this particular topic. In this project, an attempt has been made to predict the stock market prices purely based on the price trend of a particular stock with the help of Neural Networks and Machine learning Algorithms.

Several algorithms have been used in stock prediction such as SVM, Neural Network, Linear Discriminant Analysis, Linear Regression, KNN and Naive Bayesian Classifier. Literature survey revealed that SVM has been used most of the time in stock prediction research. Li, Li, and Yang (2014) [1] have considered sensitivity of stock prices to external condition. The external conditions taken into consideration include daily quotes of commodity prices such as gold, crude oil, nature gas, corn, and cotton in 2 foreign currencies (EUR, JPY). In Dai and Zhang (2013) [2], It was found that logistic regression turned out to be the best model with a success rate of 55.65%. the training data used in their research was 3M Stock data. Very few researchers have used SVM classifier and random forest regressor to predict the stock price change. Thus, we decided to train the ML algorithm with SVM classifier, random forest regressor and compare the results with other machine learning algorithms.

II. METHODS

The Parameters are key to machine learning algorithms. They are the part of the model that is learned from historical training data. Choosing correct parameters for a ML algorithm plays a big role in deciding the model's accuracy and its prediction confidence. Hence, these following parameters are chosen for the following algorithms on the basis of providing highest accuracy through our experimental process. Radial Basis Function Kernel (Rbf) kernel is used for SVM classifier as it is a good approach when the data is not linearly separable, which is defined by:[3]

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

The Random Forest Classifier is from ensemble learning. In random forest, random samples from a given data or training set are selected and the algorithm will construct a decision tree for every training data. Then, voting will take place by averaging the decision tree. Finally, select the most voted prediction result as the final prediction result.[4]

Impurity	Task	Formula	Description
Gini Impurity	Classification	$\sum_{i=1}^{C} f_i(1-f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Entropy	Classification	$\sum_{i=1}^{C} -f_i \log(f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Variance / Mean Square Error (MSE)	Regression	$\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$
Variance / Mean Absolute Error (MAE)	Regression	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i-\mu\rvert$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$

Long Short-Term Memory (LSTM)

LSTM is a recurrent neural network which can process not only single data points, but also entire sequences of data. The problem could be framed as randomly chosen contiguous

subsequences as input time steps and the next value in the sequence as output.[5]

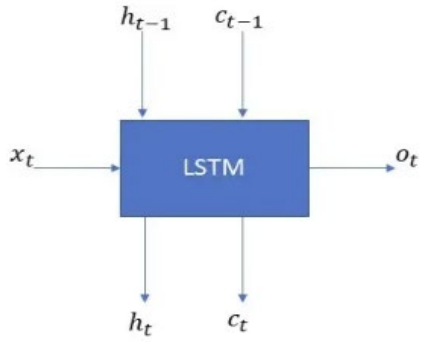

Figure 1: LSTM model

$$f_t = \sigma_g(W_f \times x_t + U_f \times h_{t-1} + b_f)$$
$$i_t = \sigma_g(W_i \times x_t + U_i \times h_{t-1} + b_i)$$
$$o_t = \sigma_g(W_o \times x_t + U_o \times h_{t-1} + b_o)$$
$$c'_t = \sigma_c(W_c \times x_t + U_c \times h_{t-1} + b_c)$$
$$c_t = f_t \cdot c_{t-1} + i_t \cdot c'_t$$
$$h_t = o_t \cdot \sigma_c(c_t)$$

Here, f_t is the forget gate and decides how much past data it should remember.i_t is the input gate and decides how much this unit adds to current state, o_t is the output gate and decides what part of the current cell state makes it to the output. c_t is the cell state, and h_t is the hidden state.

III. IMPLEMENTATION

Our approach for prediction of stock is based on application of SVM and Random Forest Regressor and compare the results with LSTM which author didn't mention in his article.

Algorithm for LSTM:

Step 1: Import the required libraries.
Step 2: Import the training dataset
Step 3: Perform feature scaling to transform the data
Step 4: Create a data structure with 60-time steps and 1 output
Step 5: Initialize the RNN.
Step 6: Add the LSTM layers and some dropout regularization.
Step 7: Add the output layer.
Step 8: Compile the RNN.
Step 9: Fit the RNN to the training set.
Step 10: Load the stock price test data.
Step 11: Get the predicted stock price.
Step 12: Visualize the results of predicted and real stock price.

Algorithm for Support Vector Machines and Random Forest Regressor.

Step 1: Load the important libraries.
Step 2: Import dataset and extract the X variables and Y separately.
Step 3: Divide the dataset into train and test.
Step 4: Initializing the SVM classifier and Random Forest Regressor model.
Step 5: Fitting the SVM classifier and Random Forest Regressor model.
Step 6: Predicting the Test Set result for Both models.
Step 7: Evaluating Both model's performance.
Step 8: Visualizing the training and test set result

Dataset

The dataset used in the code includes everyday open price, close price, highest price, lowest price, and trading volume of every stock. we analyzed the present latest data of a particular stock and created the own dataset from the values of open price, close price, and dates, then we compared the model predicted values and actual values which are from the yahoo finance and then plotted the graph between both the values with dates on x-axis and Price in USD on y-axis. From the plots obtained we can see that our model's accuracy is above 90%.

Hyperparameters for various models

- For SVM Classification, we have used (RBF Kernel, C=1e3, gamma=0.1)
- For Random Forest Regressor, we have used (n_estimators=100, criterion= 'absolute_error', random_state=1, n_jobs=10)
- For LSTM model, we have used (50, return_sequences=True, input_shape=(x_train.shape[1], 1)`

.

IV. RESULTS

In this Section, Comparison of predictions between different models have been illustrated and suggestions are given in order to improve the performance for better results, we extended our work by using LSTM and carried out the trained process.

	LSTM	**Random Forest**	**SVM**
Accuracy	**95%**	**75%**	**76%**

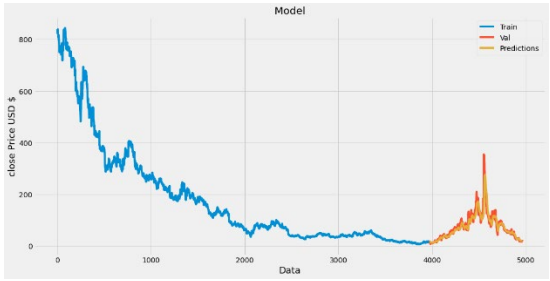

Figure 2 LSTM model Prediction with AMAZON stock Dataset

In the Figure 2, the prediction of the stock price has been displayed above, illustrating blue line as train data, red line is actual values and orange line indicates the predicted stock price change by the LSTM model. Therefore, this model has obtained an accuracy of 97%.

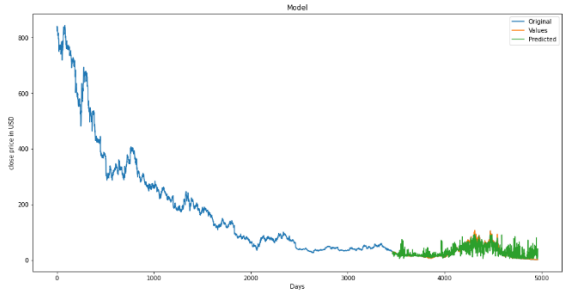

Figure 3 Stock Prediction using Random Forest Regressor on AMAZON dataset

In the Figure 3, illustrates the prediction of the stock price, where blue line as train data, red line is actual values and green indicates the predicted stock price change by the Random Forest Regressor model. Therefore, this model has obtained an accuracy of 75%.

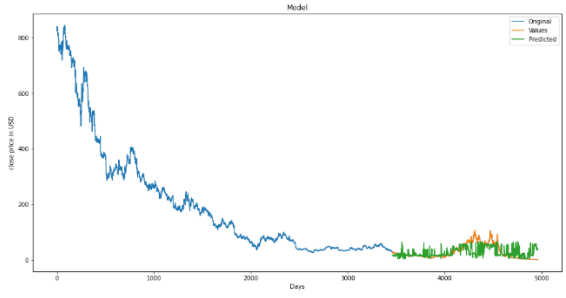

Figure 4 Stock Prediction using SVM Classifier on AMAZON Dataset

In the Figure 4, illustrates the prediction of the stock price where, blue line as train data, red line is actual values and green indicates the predicted stock price change by the Random Forest Regressor model. Therefore, this model has obtained an accuracy of 76%.

V. LESSONS LEARNED

We learned a lot about SVM concepts in this course. While working on the topic, we browsed through a large amount of literature mentioning training using SVM classifiers. Therefore, we added SVM classifier in the code during training and compared the performance with random forest classifier to observe the performance. Although SVMs have many advantages, they have some disadvantages from a practical point of view.

In this paper, we used LSTM Neural Network, Understanding and implantation of LSTM was so challenging and learned that it is used for time-series data processing, prediction, and classification. LSTM has feedback connections, unlike conventional feed-forward neural networks. Such a recurrent neural network can process not only single data points, but also entire sequences of data. The problem could be framed as randomly chosen contiguous subsequences as input time steps and the next value in the sequence as output.

VI. REFERENCES

[1] Haoming Li, Zhijun Yang and Tianlun Li (2014). Algorithmic Trading Strategy Based on Massive Data Mining. Stanford University.

[2] Yuqing Dai, Yuning Zhang (2013). Machine Learning in Stock Price Trend Forecasting. Stanford University.

[3] https://shuzhanfan.github.io/2018/05/understanding-mathematics-behind-support-vector-machines/#:~:text=SVM%20Lagrange%20problem&text=%CE%B1%20is%20called%20the%20Lagrange,%2Bb)%E2%88%921%5

[4] https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

[5] https://towardsdatascience.com/tutorial-on-lstm-a-computational-perspective-f3417442c2cd#:~:text=LSTM%20equations,-The%20figure%20below&text=The%20LSTM%20has%20an%20input,the%20next%20time%20step%20LSTM.