Analysis of various ML techniques for Stock Market Prediction

Aman
IIIT, Delhi
aman19014@iiitd.ac.in

Kushagra Gupta
IIIT, Delhi
kushagra19056@iiitd.ac.in

Shashank Dargar
IIIT, Delhi
Shashank19107@iiitd.ac.in

Yash Aggarwal
IIIT, Delhi
yash19480@iiitd.ac.in

Abstract

Stock Market speculation aims to determine the future valuation of stocks. It can be considered a complex problem because of two reasons. First, the growth of different industries and sectors over time results in variations in stock prices. Second, the market is itself volatile to historical events, leading to price fluctuations. The uncertainty also makes the problem dynamic and ever-changing. Albeit the risk, vast profits are involved, which are a significant motivation to study the field. Sound technical analysis of the market and historical data improves predicting future prices and increasing profits. The stock market is an essential part of an economy, and the fluctuations in stock prices affect a large subset of the population.

Due to the risk involved in stock prediction, Building intelligent AI systems to predict future prices becomes an important issue. The project aims to analyze various ML algorithms and their predictions of future stocks for various companies.

1. INTRODUCTION

Our project aimed to analyze various predictive models on the stock market trends of the selected companies and try to find the best prediction technique that is effective for stock market prediction. Stock market trends are nonlinear, and much data is generated worldwide every day, causing extreme speculation globally.

To be specific, for a given company, we obtain its historical stock prices, opening, closing, and daily highs and lows records and try to predict the direction and magnitude of the trend of the price using the trained models. We analyzed our results and tried to compare various learning approaches for predictions, and this comparative analysis will allow us to determine the more reliable attributes that can be trusted to predict future stock market trends.

We plan to analyze predictions on machine learning models based on Stochastic Gradient Descent(SGD) regression, Support vector machine (SVM), Long-short term memory (LSTM), and Convoluted Neural Network (CNN).

2. LITERATURE SURVEY

We see that a few papers (Lee)[1] focussed on using a Support Vector Machine (SVM) based model coupled with a hybrid method for feature selection to predict the stock trend. The author improved the performance by successfully identifying the non-informative features and removing them from calculations, thereby reducing the time spent on unnecessary computations. The matrices used for checking the usability of a particular feature were F-scores and the Supported Sequential Forward Search (SSFS) method, coupled with feature selection techniques like ID3 and correlation. In the SVM model, grid search on 5-fold cross-validation was performed to prevent overfitting without much overhead. The model successfully demonstrated filtering of non-informative features and consequently showed how feature selection has a powerful impact on the training time and accuracy. It paved the way for better SVM models. However, the model only predicted binary values, +1 and -1, which meant whether the stock would go up or down, which can be improved to take full advantage of the model achieved.

Another thread of work [2] (Sharma et al.) surveyed various Stock market prediction systems based on regression. The prediction methods included regression, both linear and polynomial regression. They also explored Sigmoid based and Radial basis function (RBF), which uses the distance from a particular point, with distances varying from euclidean to Lukaszyk–Karmowski metric. RBF regression can also be interpreted as a simple neural network. For support vector classifications, a similar technique gets used as a kernel.

We see that in some cases [4] the author tries to predict future stocks and then use multiple Portfolio Optimization techniques to improve investment strategies by investing in multiple stocks. They use stock indices such as the Average Directional Index(ADX) and the Stop and Reverse Index(SAR). ADX is used to measure the strength of the stock trend, whereas the SAR Index captures the stock trend over a given period. For time series forecasting, they have used Support Vector Regression.

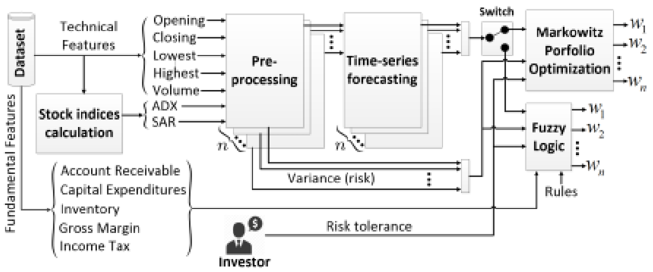

Figure 1. ML Model for Portfolio Optimization used by NekoeiQachkanloo et al. [4]..

3. DATASET DESCRIPTION

The data used in this project is of the five companies, namely Amazon, Tesla, Coke TCS, and ThomasCook ([7] , [8], [9], [10]), from September 2016 to September 2021, which is a series of data points indexed in time order or a time series. Our goal was to predict the closing price for any given date after training. All data was collected using Yahoo Finance.

The data consists of 7 attributes: Date, Open, High, Low, Close, Adjusted Close, and Volume.

- The Date column represents the date of the observation.
- The columns Open and Close represent the starting trading price and the final trading price of the stock on a particular day.
- The High is the highest trading price of the stock during a particular day. The Low is the lowest price of that day.
- Volume is the number of shares traded in the day.
- The Adj Close price column is a calculation adjustment made to a stock's closing price. It is more accurate than the closing price.

The data consists of the stock trading information of the last five years.

4.1.PREPROCESSING AND FEATURE SELECTION

Normalisation of the data has been done using MinMaxScalar of sklearn's library.

Moving Average: We have used moving average instead of actual prices for the purpose of smoothing the curves and to reduce overfitting of models due to noise. Moving average is believed to give better results in many studies [12] .

Features selection: Attributes volume and closing price have been removed.

Reason :

1. Volume has dropped, since it is the metric to measure the number of stocks traded in a day and since we are analysing the (adjacent) closing price of stocks hence the feature of volume is of no use.

2. Adjusted closing price gives us a better idea of the overall value of the stock and helps make informed decisions whereas closing stock price tells the exact cash value of a share of stock.

4.1.1 VISUALIZATION

To visualize the data we used the matplotlib library and plotted all the attributes of the data with the number of days available. The trend for each attribute was found to be similar

Below is the plot of all the companies. Each plot is a closing price vs index (Increasing Date) graph. Preprocessed data is used.

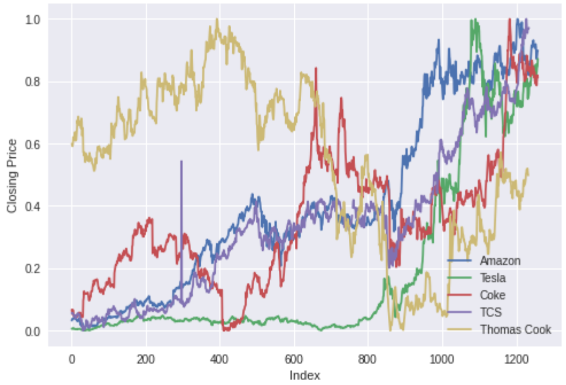

Figure 2. Training set data plots for various companies.

4. HYPERPARAMETER TUNING

We manually tried to test our models by dividing the training set into training and validation set, we tried with different values of hyperparameters. The final testing predictions were done with the best hyperparameters obtained. For SDG, we used Sklearn's grid search to find the ideal parameters.

5. METHODOLOGY

Our datasets consist of historical share market price data, and we divided our data into two sets, the training and the validation set. For the test set, we used the latest data beyond our dataset and this helped us to provide ideal conditions for testing predictions.

We considered each company independently. Considering all the stocks independently reduces noisy outliers, which can be seen when data is seen collectively.

For the purpose of analysing various ML based models, we trained four different models on the collected data; each model focuses on the pattern of the trends in a different way and thus was expected to give a different result. These results provided insight into the critical attributes that should be taken into account for stock market forecasting.

The models were trained independently for each of the companies such the hyperparameters are tuned for giving the best performance for the given company.
The models that we considered for our analysis were:

5.1 Stochastic Gradient Descent Regression(SGD): SGD regression was included as the baseline model in our study for this task. We used the moving average , which uses average prices over some intervals instead of the actual price of that day. It introduces smoothness to rather dynamic curves and helps successfully model the data in the regression. The advantage of the SGD regression method is that it does not need the assumption of distribution for the error term. SGD regression models were rather light-weight and can be used as baseline to evaluate the more complex models.

5.2 Support Vector Machine (SVM): Support Vector machine models the given data in hyperspace and uses multidimensional hyperplanes for predictions on the given testing examples. Support vector machines give better results on data with a lot of attributes independent of each other.

5.3 Convolution Neural Network (CNN): 1-D CNN uses the CNN kernel in only one direction (rather than the more standard two dimensions, as seen in image processing). 1-D CNN fits our requirements of time series forecasting in stock prediction perfectly.

5.4 Long short term memory (LSTM): Long Short-Term Memory models are extremely powerful time-series models. They are based on Recurrent Neural Networks (RNN). LSTM models prove to be a good choice for our problem statement because we need to maintain a feedback connection for keeping track of dependencies of the current data on past values.

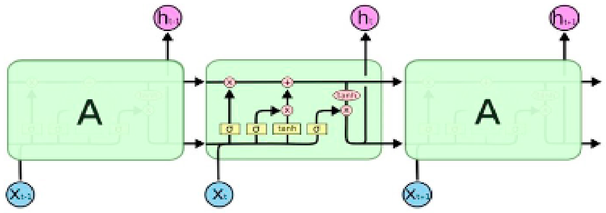

Figure 3. A Generic LSTM Implementation [13].

6. RESULTS AND ANALYSIS

Since our task was a regression task, We analysed the root mean square error (RMSE) and the R2 score for all the models above. RMSE score gives us the clear indication of the difference between the actual and the predicted values, small RMSE values indicate that we were able to predict the future stock prices with a lot of precision.
R2 score on the other hand resembles how well the model was able to incorporate the relationship between the dependent value and the input features provided. The range of R2 values is from -1 to 1. R2 score of 0 represents that the model was not able to model the relationship at all.

Models → Companies↓	**SGD**	**SVM**	**CNN**	**LSTM**
TESLA	**0.0026**	0.0074	0.0027	0.0042
AMAZON	0.0069	**0.0068**	0.0075	0.0083
COKE	**0.0011**	0.0071	0.0023	0.0029
THOMASCOOK	0.0089	0.0127	**0.0074**	0.018
TCS	**0.0052**	0.0102	0.0066	0.0077

Table 1: Root Mean Squared Error (RMSE) values for models vs companies

Models → Companies↓	**SGD**	**SVM**	**CNN**	**LSTM**
TESLA	**0.9762**	0.8987	0.9742	0.96
AMAZON	0.8874	0.8289	**0.9130**	0.86
COKE	**0.9879**	0.8733	0.9502	0.96
THOMASCOOK	0.8700	0.7083	**0.8911**	0.78
TCS	**0.9499**	0.8487	0.9296	0.93

Table 2: R2 Score values for models vs companies

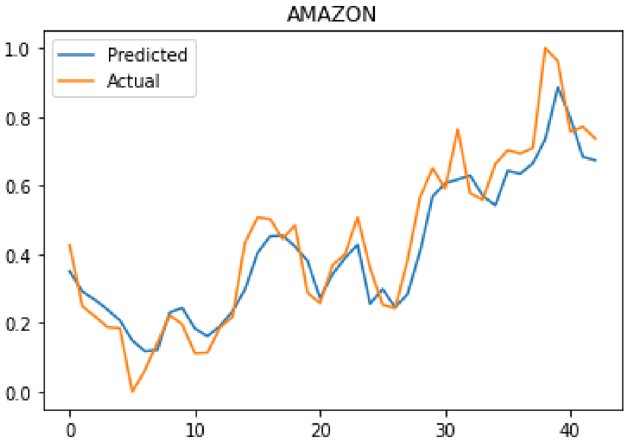

Figure 4. Predictions for SGD model on Amazon test data set

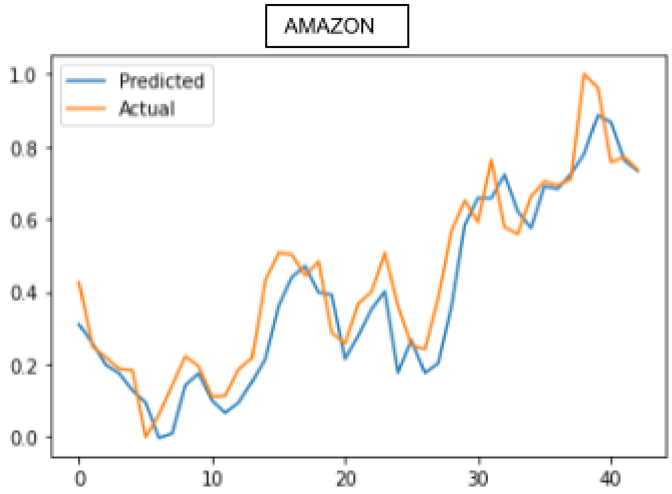

Figure 5. Predictions for LSTM model on Amazon test data set

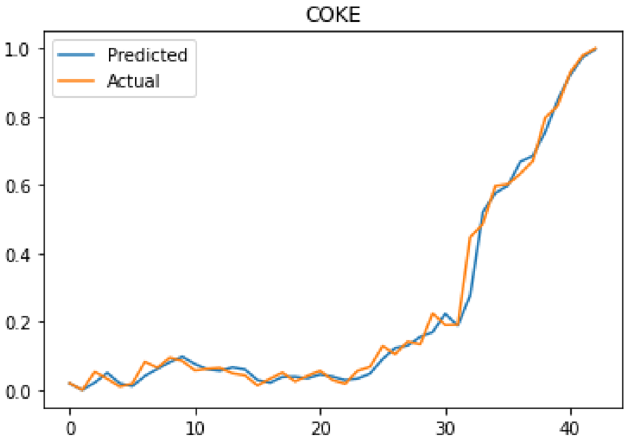

Figure 6. Predictions for SGD model on COKE test data set

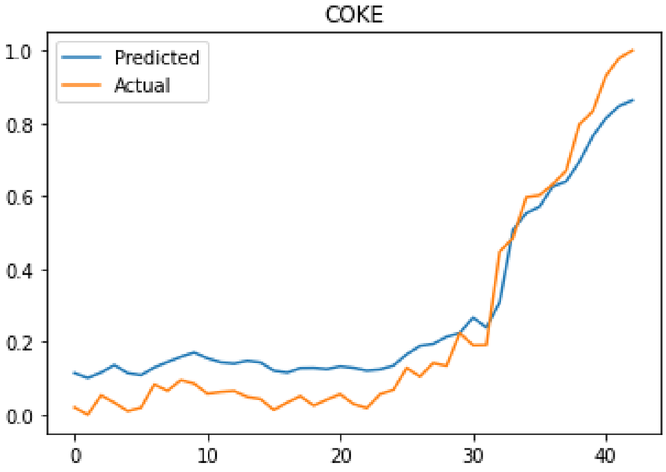

Figure 7. Predictions for SVM model on COKE test data set

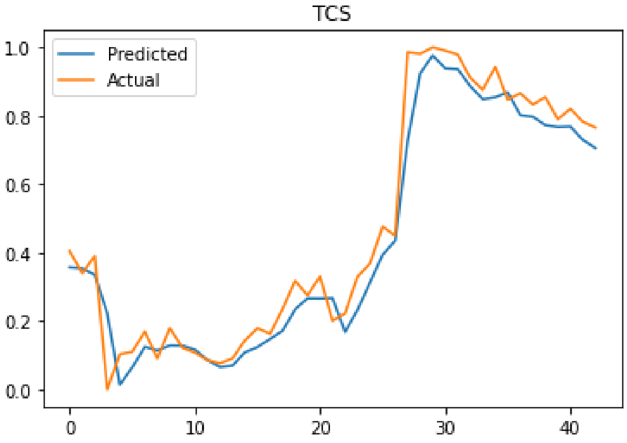

Figure 8. Predictions for SGD model on TCS test data set

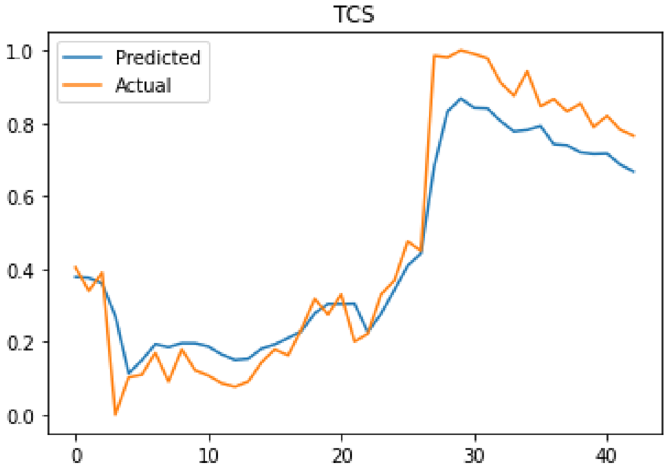

Figure 9. Predictions for SVM model on TCS data set

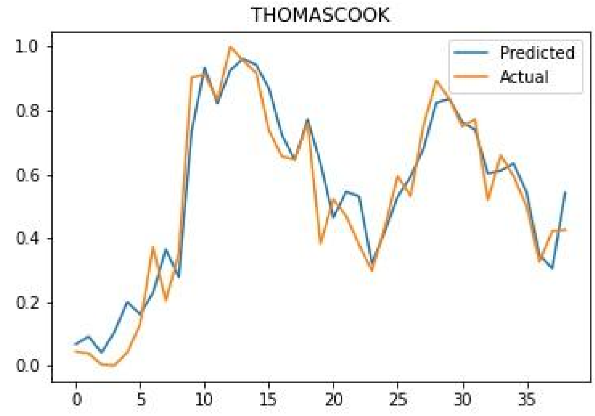

Figure 10. Prediction for CNN model on THOMASCOOK test data set

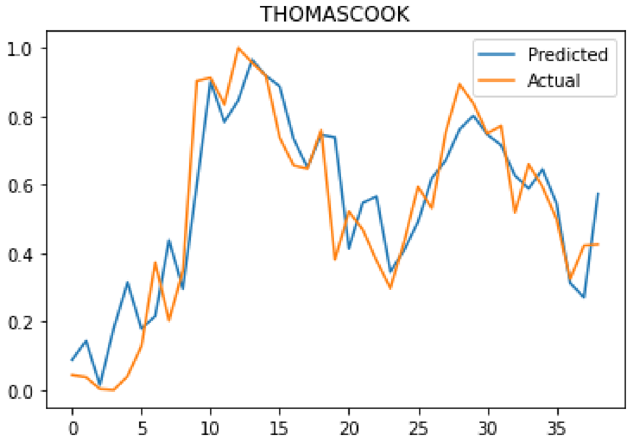

Figure11. Predictions for LSTM model on THOMASCOOK test data set

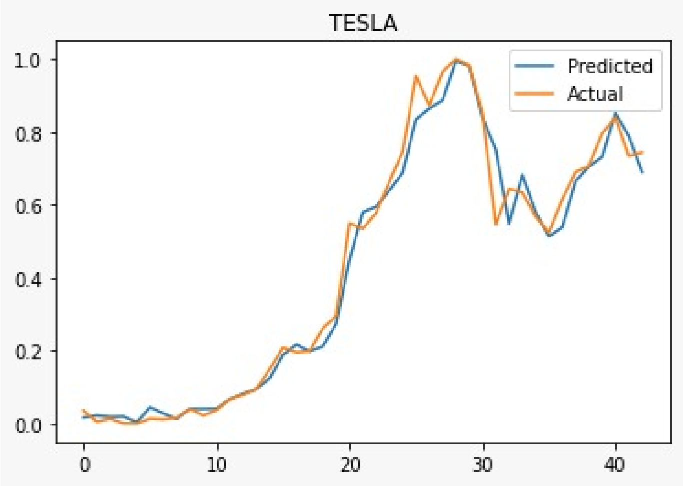

Figure 12. Predictions for SGD model on TESLA test data set

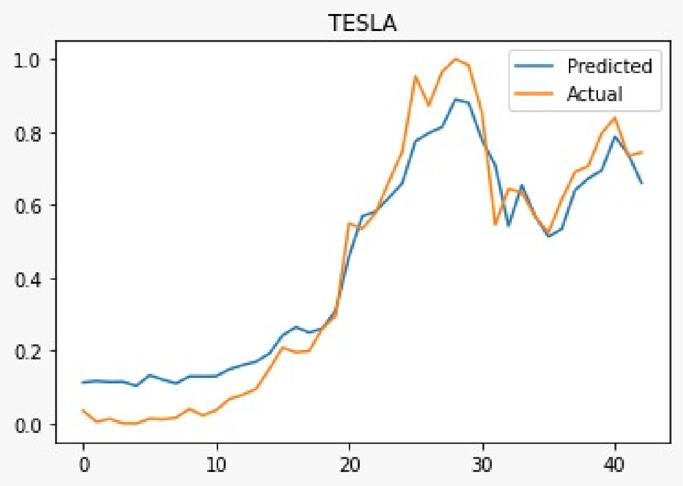

Figure 13. Predictions for SVM model on TESLA test data set

7. CONCLUSION

Our analysis of various ML based models built to forecast prices of stocks of different companies has shown some promising results. The testing results confirm that ML based techniques can be used to predict the closing prices of various stocks with a very low margin of error.
We analysed that simple models like Linear regression models can be used to predict stocks which have a lower amount of variation (average daily variation <10%) , however for dynamic stocks (like Tesla) we require complex modeling as well as some information of recent past (Long-short term memory).
From the varied array of companies that we took in our study, it is also clear that the randomness and unpredictability of the stock market was clearly visible from the case of Thomas Cook and Co. since the company stocks fluctuated to a great extent in the last two years, none of our models were able to capture the relationship very accurately. The case of ThomasCook and Co. proves that more research can be carried out in this field and the growth potential is immense.

8. FINAL OUTCOME

We analysed various machine learning-based models and their ability to predict the behavior of stocks in the near future by analyzing the stock trends over the given data set. We compared various models, some relatively simpler such as SGD Regression, and some complex models such as LSTM. However, while drawing comparison between them, we realized that even though LSTM was able to make more generalizations, the predictions obtained from the simpler models such as SGD, CNN and SVM are much more accurate and bear lesser loss. This shows that we can perform better with simpler models as well.

INDIVIDUAL TASKS

TASKS	TEAM MEMBERS
Data scraping and designing	Shashank and Aman
Constructing dataset and transformation	Yash and Kushagra
Data Visualization	Yash and Kushagra
Feature Extraction	Shashank and Aman
Researching and Training Models	Yash, Shashank, Aman, and Kushagra
Model Training: SGD and SVM	Kushagra and Aman
Model Training: LSTM	Shashank
Model Training: CNN	Yash
Analysis and Performance of Models	Shashank and Aman
Hyperparameter tuning	Yash, Shashank, Aman, and Kushagra
Report Writing	Yash, Shashank, Aman, and Kushagra

9. REFERENCES :

[1] Ming-Chi Lee, Using support vector machine with a hybrid feature selection method to the stock trend prediction, Expert Systems with Applications, Volume 36, Issue 8, 2009, Pages 10896-10904

[2] A. Sharma, D. Bhuriya and U. Singh, "Survey of stock market prediction using machine learning approach," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), 2017, pp. 506-509.

[3] M. R. Hassan and B. Nath, "Stock market forecasting using hidden Markov model: a new approach," 5th International Conference on Intelligent Systems Design and Applications (ISDA'05), 2005.

[4] NekoeiQachkanloo, H. et al., 2019. Artificial Counselor System For Stock Investment. In Innovative Applications of Artificial Intelligence (IAAI-19). 27 January . IAAI-19 Conference, Honolulu, Hawaii, USA, 2019.: AAAI Press., p. 8.

[5] http://cs229.stanford.edu/proj2009/PhamChienLim.pdf

[6] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6

[7] https://www.kaggle.com/rpaguirre/tesla-stock-price

[8] https://www.kaggle.com/hershyandrew/amzn-dpz-btc-ntfx-adjusted-may-2013may2019

[9] https://www.kaggle.com/tarunpaparaju/apple-aapl-historical-stock-data

[10] https://www.kaggle.com/camnugent/sandp500

[11] https://finance.yahoo.com/

[12] https://www.akademiabaru.com/doc/ARBMSV14_N1_P35_41.pdf

[13] lah, C. (2015). Understanding lstm networks–colah's blog. Colah. github. io.