Stock Price Prediction: Dataset Size Driven Analysis

Eden Dubrovsky
Computer Science
New Jersey Institute of Technology
Newark, New Jersey
ed239@njit.edu

Senjuti Basu Roy
Computer Science
New Jersey Institute of Technology
Newark, New Jersey
senjutib@njit.edu

Abstract

The goal of this paper is to investigate machine learning models and its effectiveness to predict individual stock prices. Analyzing and understanding the behavior of market indices such as the S&P 500, Dow Jones Industrial Average, Nasdaq Composite Index, etc., can be effectively guided by predicting individual stock, since they are an integral part of these indices and contribute to their overall performance. Predicting closing stock prices has been a challenging task due to various influential factors including political events, unexpected incidents, and the economic conditions of the company. Despite these external factors, there is enough data available from publicly traded stocks that can enable us to predict closing stock prices within a reasonable accuracy threshold. With increased individual computation capability, it has become common to predict stock prices using machine learning algorithms. The paper studies 4 off-the-shelf machine learning models (Least Short-Term Memory (LSTM), Extreme Gradient Boosting Regressor (XGBRegressor), Linear regression, and Support Vector Regression (SVR)) for the time series prediction problem. Predicting future stock prices and has three goals. – firstly, to comprehend their effectiveness by varying training set size, secondly, to assess their training time in relation to the training set size, and lastly, to evaluate the effectiveness of these models and datasets in transfer learning. The evaluation of these models will be based on the root mean squared error (RMSE) and mean absolute error (MAE) metrics. Based on these measures the study discovered that the XGBRegressor model was the most accurate while the linear regression model was the most efficient. This study aims to optimize the amount of training data required to build a precise machine learning model for predicting closing stock prices.

CCS Concepts

• Computing methodologies ~ Machine learning ~ Machine learning algorithms

Keywords

Machine Learning, Stock Price Prediction, RMSE, MAE, Training Data Set Size

ACM Reference format:

Eden Dubrovsky, Senjuti Basu Roy. 2023. “Stock Price Prediction: Dataset Size Driven Analysis,” In *Proceedings of ACM Data Economy conference (Data Economy’ 23). ACM, New York, NY, USA, 6 pages.*

1. Introduction

Market index predictions are closely related to the overall economy since they provide performance indicators of the stock market as a whole. Predictions of these indices can help present insights into the state of the economy which allows investors to make informed decisions to contribute to economic growth and stability. These decisions stem from individual stock predictions. Since market indices such as the Dow Jones Industrial Average are comprised of thirty publicly traded companies, seeing individual stock trends is the first step in understanding the direction of the overall market.

The stock market is a complex system, and predicting its behavior is challenging because of outside influential factors and its non-linear behavior. With the primary interest of investors and traders being to maximize profits, making stock price predictions is a vital aspect of their decision-making process. Because of the advancements in machine learning techniques over several decades, stock price prediction has become much more feasible and accessible. Machine learning has gained immense popularity in the finance industry due to its ability to learn from historical patterns and predict future trends.

These machine learning algorithms can be trained on large datasets to increase their learning and accuracy. However, for the algorithms to be relevant and useful for investors and traders, the predictions must be timely. Since stock market transactions occur at a fast pace, the training time of the algorithm must be low enough to ensure the results are applicable to the market. If the results are not available in a timely manner, the accuracy can become insignificant since the model cannot be used.

This paper aims to explore the effects of varying training set sizes on the performance of the algorithms based on the RMSE and MAE, while also analyzing the impact on the time taken to train the models.

Predicting stock market behavior based on historical data is challenging, mainly due to factors such as data quality, overfitting, and the impact of unexpected events on the stock market. To conduct the experiment, data from Yahoo Finance was utilized for all training set sizes. The data contains several price measures: Open, High, Low, Close, Adjusted Close (Adj Close), and Volume.

Overall, the main goal of this paper is to achieve three objectives. Firstly, to investigate the effectiveness of four machine learning algorithms – LSTM, XGBRegressor, Linear Regression, and SVR – in relation to varying training set size for stock prediction. Secondly, to compare their training times to determine which model is the most efficient. Lastly, to evaluate the effectiveness of these models and datasets in transfer learning, hence identifying the most promising algorithms for future research.

2. ML Models for Stock Prediction

The models used in this study were chosen based on popularity in the finance industry. While LSTM is the most used algorithm for this problem, all models demonstrated promising results.

2.1 LSTM Approach

Least Short-Term Memory is a recurrent neural network (RNN) algorithm that is used for time series analysis. LSTM can retain information over extended periods of time making it ideal for stock prediction. The algorithm learns long-term dependencies by processing historical data one time-step at a time and storing the information in its memory cell [1]. This data is then used to make future predictions. The model is trained on a sliding window value, 60, indicating that each day's prediction is based off of the previous 60 days. To accurately train the data, a scaler was used to normalize the input values. This is done to ensure standardization and consistency within the data across all input features.

The LSTM model was optimized through hyper parameter tuning modifying the LSTM units, number of epochs, and batch size. The model contained two LSTM layers with 256 and 128 units in addition to two dense layers with 25 and 1 units.

Limitation: The training of the LSTM model was conducted solely on the Close prices, as opposed to the other models which were trained on a more diverse range of variables including Open, High, Low, Adjusted Close (Adj Close), and Volume.

2.2 XGBRegressor Approach

XGBRegressor is based on the gradient boosting algorithm XGBoost, but handles continuous data as opposed to discrete data. Due to its capacity to handle large amounts of multidimensional data and model nonlinear relationships between variables efficiently, it has become a popular choice for stock price prediction. The model works by building weak prediction models (decision trees) and combining their predictions to make stronger predictions [5]. The model is optimized during the training process by adjusting the weights of the weak models based on the performance.

In this study, XGBRegressor was enhanced by tuning two critical hyperparameters, the n_estimators and learning rate. N_estimators determine the number of decision trees that the algorithm creates, which can also be interpreted as the number of times the algorithm will learn. The learning rate is the speed at which the model learns. Setting a learning rate above 0.3 could result in over fitting. The optimal model was obtained with 1000 n_estimators and a learning rate of 0.05.

Data Economy'23, June 2023, Seattle, WA USA

2.3 Linear Regression

Linear Regression is a statistical method that identifies a linear relationship between a set of independent variables. In this study, historical stock prices along with technical indicators, and a dependent variable (closing price) are used to generate predictions. The linear regression algorithm is one of the foundational supervised machine learning algorithms for regression using continuous data [3]. This model can be used for time series forecasting and non-time series forecasting [6].

2.4 Support Vector Regression

Support Vector Regression is a type of Support Vector Machine that can be used to predict stock prices based on historical data [7]. The model ultimately finds the line of best fit which is the hyperplane with the maximum amount of data points [4]. SVR can handle nonlinear relationships between independent and dependent variables which is an advantage in terms of stock market predictions. The accuracy of the model can be hyper tuned with several parameters– in this study the parameter C was set to the optimal value of 2. C is the regularization parameter which defines the amount of misclassification each training sample should avoid. Similarly, to the LSTM model, a scaler was applied to the input data to normalize the data. In this model the scaler had a noteworthy effect on accuracy. When using a dataset from 2011, the SVR model with no scaler had an RMSE of 40.9 and MAE of 26.7 while with the scaler the RMSE was 1.41 and the MAE was 0.6. Due to this considerable difference, a scaler was implemented for this particular model.

3. Experimental Evaluation

We compared the effectiveness of four machine learning algorithms described in Section 2 in terms of RMSE (Root mean squared error ($)), MAE (mean average error ($)), and training time efficiency in seconds [7]. The datasets used in this study consisted of historical stock market data for three different stocks – Apple (AAPL), IBM (IBM), and Chevron (CVX). The data sets were collected from Yahoo Finance. The datasets were split into training and testing sets, with the training sets being 80% of the individual datasets. The testing sets comprised 20% of the individual datasets.

Start Date	Training Set Size	Testing Set Size
10/10/2011	2309	578
10/10/2005	3518	880
10/10/2000	4522	1131
10/10/1995	5536	1385

Fig. 1. Start Date and Dataset Size

Each model was trained and tested on each dataset five times, and the average RMSE and MAE were computed accordingly.

4.1 Impact of Dataset Size

To investigate the effects of varying dataset sizes on the accuracy of the machine learning models we trained and tested on progressively larger datasets.

The first experiment computed the RMSE, MAE, and training time of the four algorithms. The results are as follows:

Start Year	2011	2005	2000	1995
RMSE	3.46	2.85	2.60	2.26
MAE	2.78	2.15	1.93	1.55
Training Time (s)	630.13	990.15	1230.72	1294.86

Fig 2. RMSE, MAE, and Training Time with LSTM

Start Year	2011	2005	2000	1995
RMSE	0.38	0.35	0.24	0.19
MAE	0.27	0.20	0.12	0.09
Training Time (s)	5.66	5.83	6.71	8.35

Fig 3. RMSE, MAE, and Training Time with XGBRegressor

Start Year	2011	2005	2000	1995
RMSE	0.37	0.38	0.31	0.31
MAE	0.28	0.25	0.20	0.18
Training Time (s)	0.20	0.26	0.27	0.29

Fig 4. RMSE, MAE, and Training Time with Linear regression

Start Year	2011	2005	2000	1995
RMSE	1.41	1.28	1.09	0.89
MAE	0.60	0.44	0.34	0.26
Training Time (s)	0.57	0.92	1.18	1.62

Fig 5. RMSE, MAE, and Training Time with SVR

The listed charts correlate to the following line graphs which display the trends of the evaluation metrics. The line graphs show the relationship between dataset size, based on the start year, RMSE, MAE, and training time. The dataset size can be determined based on the year as shown in Fig. 1. Figures 6-9 show the tendency of the error measures and training time in relation to increasing dataset sizes.

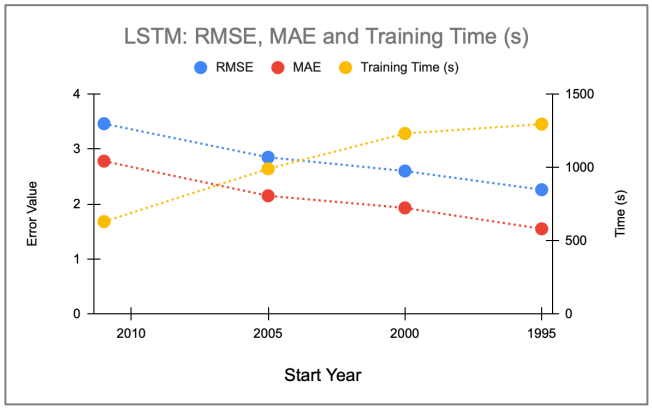

Fig 6. RMSE, MAE, and Training Time with LSTM

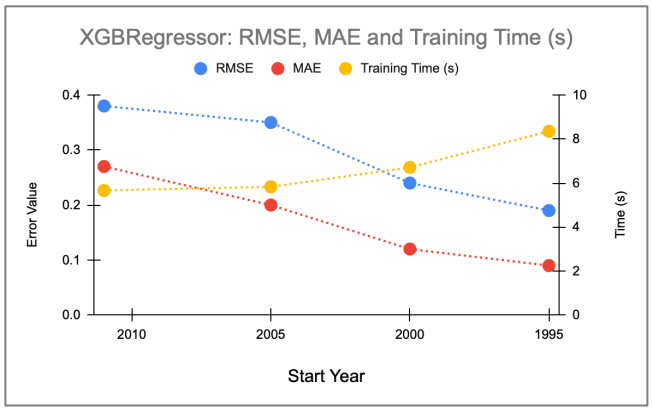

Fig 7. RMSE, MAE, and Training Time with XGBRegressor

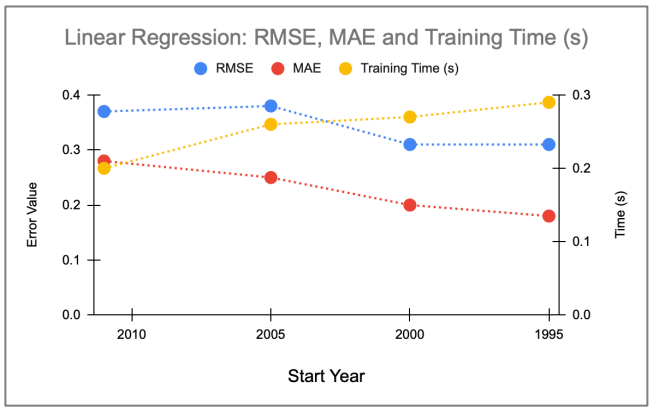

Fig 8. RMSE, MAE, and Training Time with Linear regression

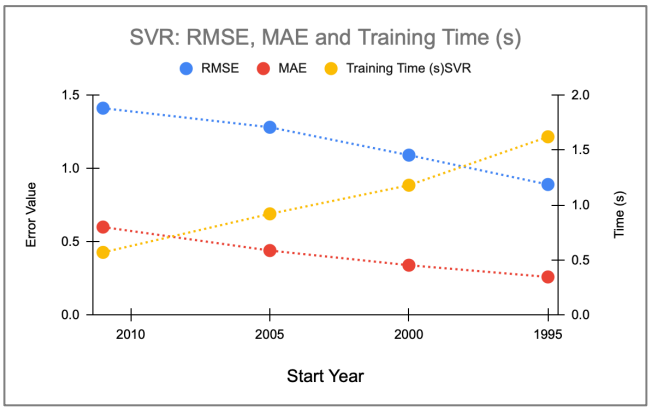

Fig 9. RMSE, MAE, and Training Time with SVR

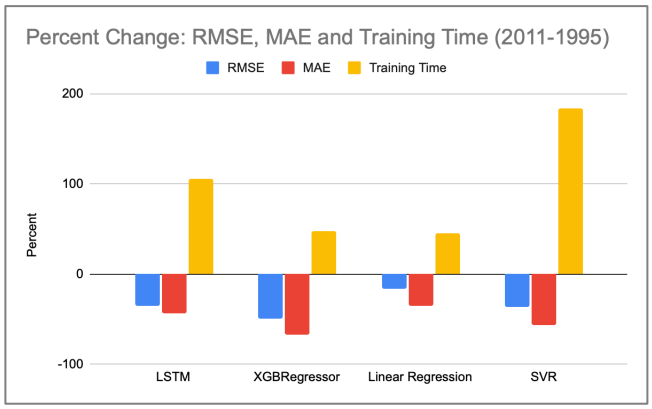

Fig 10. Percent Change: RMSE, MAE, and Training Time

Computation efficiency is a vital aspect of stock price prediction because of the time limitation. The stock market always has various levels of volatility. It is highly dynamic so accurate predictions in minimal time maximize profitable trades for investors and traders. A model with a high training time could produce accurate results that quickly become irrelevant at the time the model is ready to make predictions. In stock trading, even small delays (milliseconds) can result in significant loss, so speed is equally as important as accuracy.

4.1.1 Analysis

In this study we compare accuracy and training times for stock predictions using LSTM, XGBRegressor, linear regression, and SVR. Figures 6-9 show the changes in RMSE, MAE, and training time as the dataset size increases.

The results show that the accuracy of the XGBRegressor model outperformed the other models at all dataset sizes. This was followed by the linear regression model, the SVR model, and then the LSTM model.

In relation to training time, linear regression had the lowest training time followed by XGBRegressor, SVR, and then LSTM.

All the models showed overall improvement in RMSE and MAE scores as the dataset size increased. However, the models had increased training times as the dataset size increased. Ultimately, the XGBRegressor model and linear regression model are optimal for closing stock prediction based on their low error rates and low training times.

The trade-off between training time and accuracy is an ongoing challenge in machine learning. More complex models require longer training times, while simpler models have lower training times but sacrifice accuracy. Figure 10 examines the percent change in RMSE, MAE, and training time from 2011 (smaller dataset) to 1995 (larger dataset).

Between 2011 and 1995, the LSTM model showed a 35% reduction in RMSE, a 44% reduction in MAE, and a 105% increase in training time. Similarly, the XGBRegressor model showed a decrease in RMSE and MAE scores by 50% and 67%, respectively, and a 48% increase in training time. In the same time frame, the linear regression model experienced a 16% reduction in RMSE, a 36% reduction in MAE, and a 45% increase in training time. The SVR model demonstrated a decrease in RMSE and MAE scores by 37% and 57%, respectively, and a 184% increase in training time.

This study's findings illustrate the impact the size of the training dataset has on the accuracy of machine learning algorithms when predicting stock prices. The significance of this study is that it highlights the importance of selecting favorable training dataset sizes to achieve accurate and timely predictions. The results can be valuable for researchers, investors and traders in the financial industry who rely on machine learning algorithms to maximize profits from investments.

4.2 Transfer Learning

Transfer learning is a popular approach in machine learning that with regard to stock prediction involves using a previously trained model to predict stock prices of a new dataset. For instance, in this experiment the model was trained on the historical data of IBM stocks, and then tested on the historical data of CVX stocks. This study analyzes the effect of the dataset size on the accuracy results of the CVX prediction. The variability among the dataset size is shown in Figure 11.

Start Date	Training Set Size	Testing Set Size
01/01/2023	53	14
01/01/2022	254	64
01/01/2015	1664	417
01/01/2010	2671	668

Fig 11. Training Dataset Sizes for Transfer Learning

Figures 12-13 illustrate the RMSE, MAE, and training times of the model when using transfer learning.

Start Year	2023	2022	2015	2010
RMSE	3.68	9.32	22	29.96
MAE	2.44	8.15	18.44	24.45
Training Time (s)	4.49	4.45	6.1	6.86

Fig 12. Transfer Learning with XGBRegressor

Start Year	2023	2022	2015	2010
RMSE	9.3	10.66	23.17	32.84
MAE	8.17	8.68	18.94	25.6
Training Time (s)	0.39	0.43	0.54	0.63

Fig 13. Transfer Learning with Linear regression

Figures 12-13 are represented as line graphs in Figures 14-15 to show the trend of RMSE, MAE, and training time when using transfer learning.

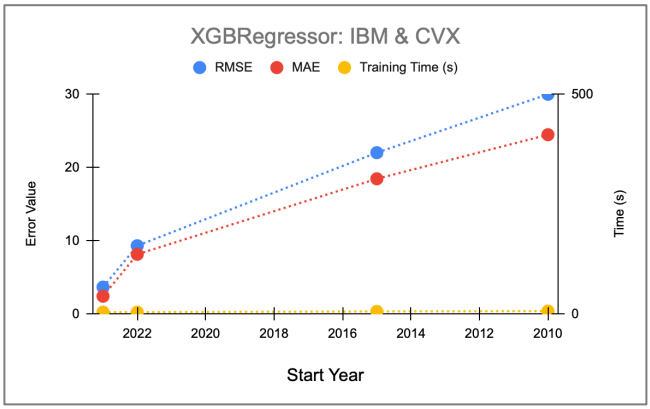

Fig 14. Transfer Learning with XGBRegressor

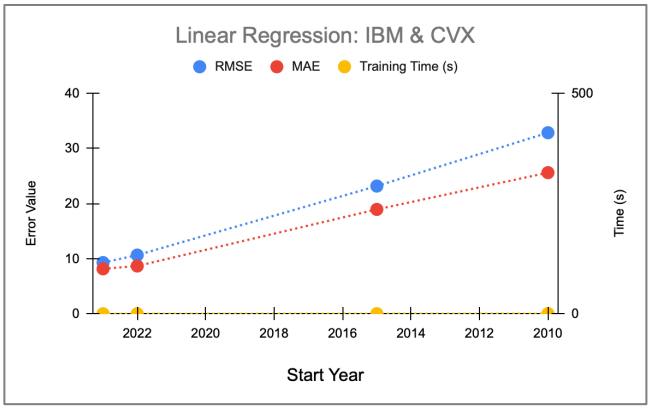

Fig 15. Transfer Learning with Linear regression

4.2.1 Analysis

A transfer learning model trained on IBM stock data and tested on CVX stock data was evaluated in terms of its performance with respect to the size of the dataset. The dataset starting from 2023 provided the most accurate results, especially with the implementation of the XGBRegressor model. As the dataset size increased, the accuracy of both model became comparable but XGBRegressor still outperformed the linear regression model.

The results show that as the size of the dataset increases, both the error rate and training time of the models increase. It is logical to see this outcome. The larger the dataset, the greater the variation in stock market prices and trends between two companies.

The training model optimized for the IBM stock, may not have been able to capture the nuances of the CVX stock as the datasets became larger and more complicated.

While transfer learning models can be useful for predicting stock prices, it is important to consider the size of the dataset and the trade-offs in terms of accuracy. The reason for the lower error rates observed in smaller datasets could be attributed to the fact that there is less time for the two stocks to diverge significantly. The use of transfer learning with other sources of data, such as news, could enhance the models ability for more accurate predictions. In the future, additional data sources along with stock correlation can be included in transfer learning experiments.

5. Conclusion

This research paper aimed to investigate the impact of dataset sizes on the accuracy and training times for four machine learning algorithms, along with the effectiveness of transfer learning in the field of stock price prediction. Through running various experiments and analyzing the results, we have found that the size of the training dataset has a significant effect on the accuracy of these algorithms. We have investigated the accuracy and efficiency of various algorithms to determine their priority for use. The results also provide insight for future studies to examine in greater detail the relationship between accuracy, training time, and training size by exploring factors such as the type and quality of data for various models.

Stock market prediction plays an essential role in facilitating future market health and promoting data economy. Accurate predictions of stock prices can help investors and traders create better investment strategies resulting in increased market efficiency. It also enables financial entities to mitigate risks and optimize portfolio management, which leads to more stable financial systems. Additionally, stock market prediction can help promote data economy by encouraging innovative solutions that use the large amounts of data generated by the stock market.

In conclusion, stock market prediction is a critical application of machine learning that has potential to positively contribute to future market health and data economy. More accurate stock market prediction eventually lead to new opportunities, innovation, and growth within many industries.

7. References

[1] K. J, H. E, M. S. Jacob and D. R, "Stock Price Prediction Based on LSTM Deep Learning Model," *2021 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Puducherry, India, 2021, pp. 1-4, doi: 10.1109/ICSCAN53069.2021.9526491.

[2] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar, "Stock Closing Price Prediction using Machine Learning Techniques," *Procedia Computer Science, Volume 167, 2020,* Pages 599-606, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.326.

[3] I. Bhattacharjee and P. Bhattacharja, "Stock Price Prediction: A Comparative Study between Traditional Statistical Approach and Machine Learning Approach," 2019 4th International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, 2019, pp. 1-6, doi: 10.1109/EICT48899.2019.9068850.

[4] F. L. Marchai, W. Martin and D. Suhartono, "Stock Prices Prediction Using Machine Learning," *2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, Semarang, Indonesia, 2021, pp. 79-84, doi: 10.1109/ICITACEE53184.2021.9617222.

[5] Y. Hu, L. Shao, L. La and H. Hua, "Using Investor and News Sentiment in Tourism Stock Price Prediction based on XGBoost Model," *2021 IEEE/ACIS 6th International Conference on Big Data, Cloud Computing, and Data Science (BCD),* Zhuhai, China, 2021, pp. 20-24, doi: 10.1109/BCD51206.2021.9581619.

[6] B. Panwar, G. Dhuriya, P. Johri, S. Singh Yadav and N. Gaur, "Stock Market Prediction Using Linear Regression and SVM," *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE),* Greater Noida, India, 2021, pp. 629-631, doi: 10.1109/ICACITE51222.2021.9404733.

[7] Fernando Berzal and Nicolfás Matín. 2002. "Data mining: concepts and Techniques" by Jiawei Han and Micheline Kamber. *SIGMOD Rec. 31*, 2 (June 2002)

[8] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4

[9] Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In Thirteenth annual conference of the international speech communication association.