Comparison of Various Machine Learning Techniques in Stock Price Prediction

Badal Varshney
dept. of Electrical Engineering
Indian Institute of Technology, Bombay
Mumbai, India
19D070015@iitb.ac.in

Atharva Warhade
dept. of Electrical Engineering
Indian Institute of Technology, Bombay
Mumbai, India
190070012@iitb.ac.in

Aryan Padmakar Dolas
dept. of Mechanical Engineering
Indian Institute of Technology, Bombay
Mumbai, India
190100025@iitb.ac.in

***Abstract*—With the growing awareness of the stock market among the general public today, it has become a very burning reign for analytics. In the following paper, we've analyzed the TATA motors stock prices and statistics information ranging from 2012 to 2021 and checked the performances of various Machine Learning Models to accurately predict the prices in future. We trained Linear Regression, k-Nearest Neighbour, ARIMA, LSTM, GRU, and hybrid (LSTM+GRU) and found that the hybrid model performed the best out of those given above.**

I. Introduction

The stock market has become an increasingly popular area of interest among investors in recent years, with a growing number of people seeking to make informed decisions when investing their money. Predicting stock prices using machine learning is important because it can help investors make better investment decisions, companies and financial institutions to manage risk more effectively, and policymakers and economists to gain insights into the broader economic landscape. Accurate stock price predictions can lead to higher returns and reduced risk for investors, while also informing economic policy decisions. Additionally, by analyzing historical data and identifying patterns, machine learning algorithms can help organizations to understand their exposure to market risk and take appropriate measures to mitigate it.

Stock prediction using machine learning is challenging due to the complex patterns and noisy data associated with stock prices, limited historical data, non-stationary data, and market inefficiencies. The stock market is influenced by a wide range of factors that interact in unpredictable ways, making it difficult to identify clear patterns and predict stock prices accurately. Additionally, stock prices are highly volatile and subject to noise and randomness, making it difficult to distinguish true signals from random fluctuations in the data. Machine learning algorithms require advanced techniques, domain expertise, and a deep understanding of the underlying factors influencing stock prices to predict future prices accurately.

In this paper, we present an analysis of TATA motors' stock prices and statistics from 2012 to 2021, with the aim of predicting future prices using various machine learning models. We believe our findings will greatly interest investors looking to make informed decisions when investing in TATA motors and provide valuable insights into the performance of different machine-learning models for stock price prediction. By analyzing historical data and comparing the performance of various machine learning models, we hope to provide valuable insights into the trends and patterns that underlie TATA motors' stock prices and to help investors make more informed decisions when investing in the stock market.
In our current model, we limit ourselves to using a single feature of the stocks, i.e., close price, to predict the stock's value. We use data manipulation techniques in order to generate specific features to take as input for the given model.

We also explore hybrid models (models made by a combination of two or more models) and see how some types of hybrid models can perform even better than their individual counterparts

II. Background and Related Works

The stock market is a dynamic and complex system that involves the buying and selling of shares of publicly traded companies. The stock prices of these companies are influenced by a wide range of factors, including economic conditions, geopolitical events, company performance, and investor sentiment. As a result, predicting stock prices accurately is a challenging task that has long been a topic of interest for investors and financial analysts.

Traditionally, stock market analysis has relied on fundamental analysis and technical analysis. Fundamental analysis involves examining a company's financial statements and economic conditions to determine its intrinsic value, while technical analysis involves analyzing historical market data to identify trends and patterns. While these methods can be effective, they have limitations in terms of their accuracy and their ability to handle the complexity and noise of stock price data.

In recent years, there has been a growing interest in using machine learning techniques to predict stock prices more accurately. Machine learning algorithms can analyze large volumes of historical data, identify patterns and relationships, and make predictions based on these patterns. The potential benefits of using machine learning for stock market prediction include increased accuracy, better risk management, and improved investment decision-making.

Kim and Han [1] developed a model using a combination of artificial neural networks (ANN) and genetic algorithms (GAs) to predict the stock price index by discretizing the features. The data used in their study included technical indicators and the direction of change in the daily Korea stock price index (KOSPI) from January 1989 to December 1998. They optimized feature discretization using GA, which is similar to dimensionality reduction. However, their model had limitations, such as a fixed number of input features and processing elements in the hidden layer. Additionally, they only focused on two factors in optimization. Qiu and Song also proposed a hybrid GA-ANN model to predict the direction of the Japanese stock market. They used genetic algorithms with artificial neural network models to optimize the process.

Lee [2] developed a stock trend prediction model using a hybrid feature selection approach and support vector machine (SVM). The research dataset is a subset of the NASDAQ Index in the Taiwan Economic Journal Database (TEJD) from 2008. The feature selection process involved a hybrid approach, where sequential forward search (SSFS) acted as the wrapper. The authors also provided a detailed procedure for adjusting parameters, and the model's performance was evaluated under various parameter values. Another strength of the study is the clear structure of the feature selection model, which can aid in the initial stages of model structuring. However, the study only compared the performance of SVM to a backpropagation neural network (BPNN) and did not include other machine learning algorithms in the comparison.

III. Dataset and Feature Engineering

We have used the TATAMOTORS stock price dataset from Yahoo Financials comprising of the Open, High, Low, Close, Adj Close prices and Volume of stocks traded for each day from 1 January 2010 to 31 March 2023. The dataset is cleaned and filtered for missing values, and duplicate entries are dropped. This gives us a data frame with 3270 entries. We further drop the Open, High, Low, and Adj Close columns as we are only concerned with the closing prices. A plot of closing prices versus time is given below

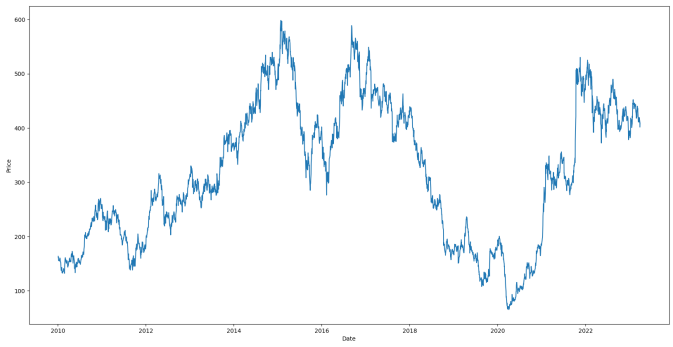

Fig. 1. Closing Prices of TATAMOTORS versus time (from 1 January 2010 to 31 March 2023)

For simpler models, it is intuitive that the closing price alone might not be sufficient to predict future closing prices. For models like simple Linear Regression and kNN, we need to invent additional features from the given features that give a better representation and prediction of data. We add two additional features for these models: Moving Averages and Bollinger Bands.

A. Moving Averages

To develop a more stable and effective model for identifying stock growth over time, we opted to use moving averages as features instead of relying solely on daily stock prices. This approach considers the effect of the previous day's prices on the predicted day's stock price, resulting in a more accurate prediction.

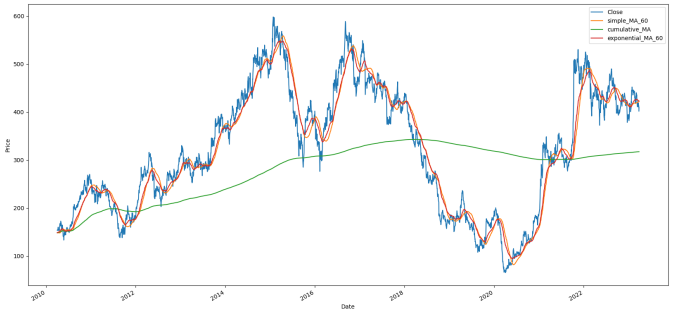

Fig. 2. Moving Averages- Closing Prices of TATAMOTORS versus time (from 1 January 2010 to 31 March 2023)

B. Bollinger Bands

In our model (for Linear Regression and kNN), we have incorporated Bollinger bands, representing moving averages and accounting for standard deviation. This approach ensures that sudden changes in data are considered and prevents biasing. By using Bollinger bands as a training feature, our model becomes more robust and can handle sudden changes in stock prices without throwing off our training parameters. Our model incorporates four Bollinger bands, including positive and negative single and double standard deviations. Additionally, we smooth the data to enable clear trend prediction rather than focusing on small changes in stock prices over a short period.

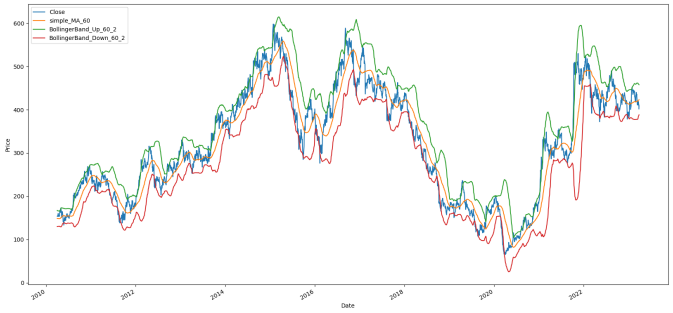

Fig. 3. Bollinger Bands- Closing Prices of TATAMOTORS versus time (from 1 January 2010 to 31 March 2023)

IV. Experiment

Here we look at the approach and performance of various models used in our project and compare them analytically in the next section.

A. Linear Regression (with added features)

Linear regression is a statistical approach for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and tries to estimate a linear equation that predicts the dependent variable based on the independent variables. The goal of linear regression is to find the best-fitting line through the data points that minimizes the sum of the squared errors between the predicted values and the actual values.

Here we use the dates, the moving averages, and the Bollinger bands as features, and training is based on previous stock values. The following plot shows the actual versus predicted values of the stock's closing prices.

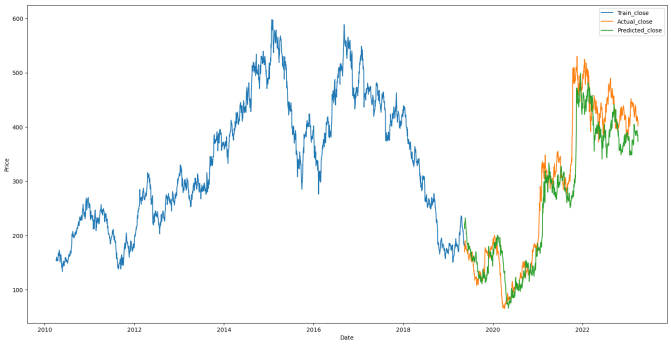

Fig. 4. Linear Regression- Closing Prices of TATAMOTORS versus time (from 1 January 2010 to 31 March 2023)

B. kNN

The K-Nearest Neighbors (KNN) algorithm is a supervised machine learning algorithm for classification and regression tasks. In KNN, the value of an unknown data point is predicted based on the k-nearest data points in the training set. The value of k is usually an odd number to avoid ties, and the distance metric used to calculate the distance between the data points is typically the Euclidean distance.

C. ARIMA

ARIMA (AutoRegressive Integrated Moving Average) model is a time series analysis method used to forecast future values based on historical data. It combines two techniques: AR (AutoRegressive) and MA (Moving Average). The AR model uses a regression equation that predicts the future value of a time series based on its past values, while the MA model predicts future values by considering the moving average of past errors. The integrated component (I) involves differencing the time series to make it stationary, meaning that its statistical properties do not change over time. Note that we stop using Moving Averages and Bollinger Bands as explicit features for this and subsequent models.

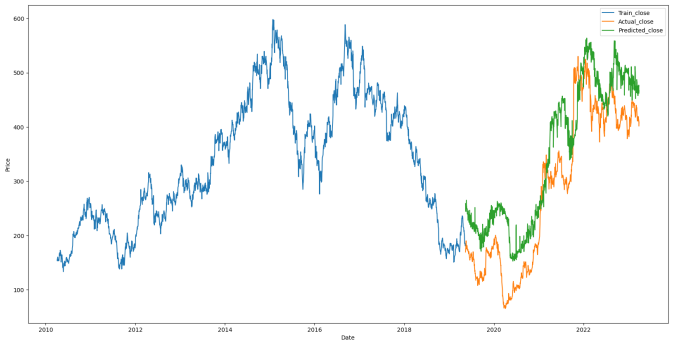

Fig. 5. kNN- Closing Prices of TATAMOTORS versus time (from 1 January 2010 to 31 March 2023)

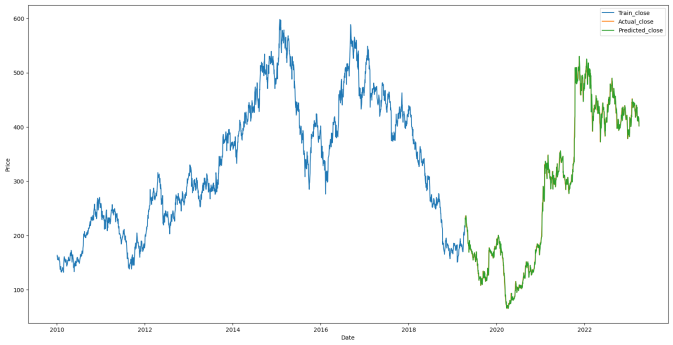

Fig. 6. ARIMA- Closing Prices of TATAMOTORS versus time (from 1 January 2010 to 31 March 2023)

D. LSTM

The Long Short-Term Memory (LSTM) model is a type of recurrent neural network (RNN) commonly used for sequential data analysis and prediction. Unlike traditional feedforward neural networks, which only take in a fixed-size input, RNNs can handle inputs of varying lengths. LSTMs are designed to address the vanishing gradient problem that can occur in RNNs, making it difficult for the network to learn long-term dependencies in the data. LSTMs achieve this by incorporating a memory cell that can store information for long periods of time and selectively update and forget information using gate mechanisms. This makes them particularly useful for time series prediction tasks, such as stock price prediction, where the relationships between data points over time can be complex and non-linear.

E. GRU

The Gated Recurrent Unit (GRU) model is a type of recurrent neural network (RNN) introduced in 2014 by Cho et al. GRUs are similar to LSTMs, another type of RNN, in that they are designed to work with sequential data, such as time series or text data. However, GRUs have fewer parameters than LSTMs, making them faster to train and more computationally efficient. Like LSTMs, GRUs have the ability to selectively remember and forget previous inputs, allowing them to learn from long-term dependencies in the data.

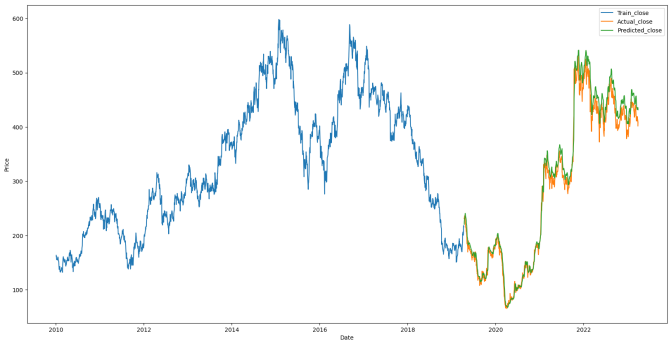

Fig. 7. LSTM- Closing Prices of TATAMOTORS versus time (from 1 January 2010 to 31 March 2023)

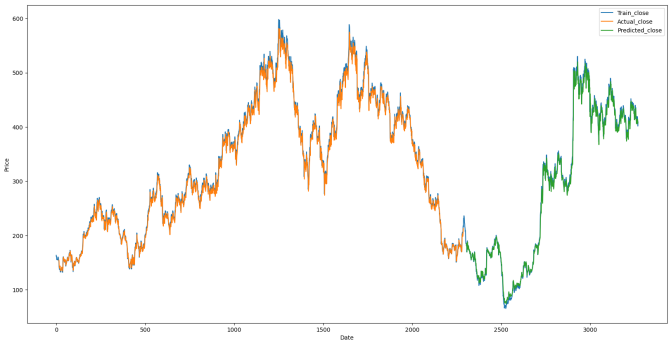

Fig. 8. GRU- Closing Prices of TATAMOTORS versus time (from 1 January 2010 to 31 March 2023)

F. LSTM + GRU Hybrid Model

The LSTM (Long Short-Term Memory) + GRU (Gated Recurrent Unit) hybrid model is a combination of two popular types of recurrent neural networks (RNNs) used for sequence modeling. This model incorporates the strengths of both LSTM and GRU, while addressing some of their limitations. The LSTM network is known for its ability to remember long-term dependencies, while the GRU network is known for its computational efficiency and simplicity. The hybrid model combines these two networks to create a powerful sequence modeling tool that is able to handle long-term dependencies while also being computationally efficient.

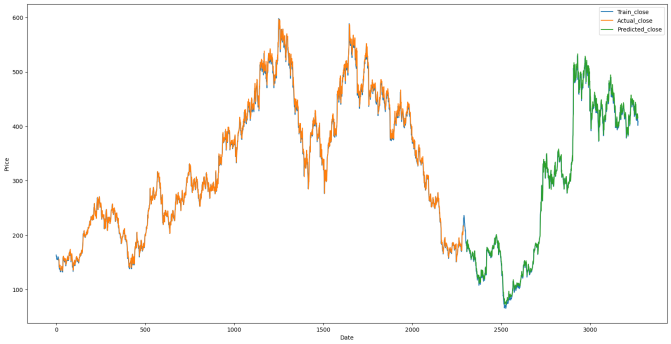

Fig. 9. Hybrid- Closing Prices of TATAMOTORS versus time (from 1 January 2010 to 31 March 2023)

V. Results

TABLE I
PERFORMANCE METRICS FOR DIFFERENT MACHINE LEARNING MODELS USED FOR STOCK PRICE PREDICTION

Models	LR	KNN	LSTM	GRU	Hybrid
Train RMSE	31.087	23.605	0.0404	9.010	8.038
Train MSE	966.439	557.227	0.0016	81.188	64.620
Train MAE	24.549	17.970	0.0323	6.685	5.932
Test RMSE	50.158	77.975	20.2553	9.692	9.127
Test MSE	2515.856	6080.134	410.2774	93.941	83.302
Test MAE	35.820	68.882	15.1916	6.723	6.284
Train var. RS[1]	0.930	0.959	0.9842	0.995	0.996
Test var. RS[2]	0.900	0.883	0.9861	0.995	0.996
Train R2	0.930	0.959	0.9657	0.994	0.995
Test R2	0.872	0.694	0.9791	0.995	0.996
Train data MGD	0.009	0.005	0.0052	0.0007	0.0005
Test data MGD	0.0389	0.110	0.0042	0.001609	0.001406
Train data MPD	2.821	1.691	0.0027	0.225	0.182
Test data MPD	8.875	23.334	1.1680	0.3209	0.2820

As we can see from the table above, the Hybrid model (LSTM + GRU) performs the best by giving the best results across all performance metrics.

VI. Conclusion

In this work, we have compared and compiled the results of training various machine learning models for stock price prediction. We found that training simple models to obtain a reasonable tolerance is possible with added features. It is important to note that without added features, simple models like Linear Regression (that can only model linear relationships within data without the use of appropriate basis functions) do not produce satisfactory results meaning the data is highly non-linear.

The hybrid model of LSTM and GRU performs better than other models and their individual counterparts due to their complementary strengths.

LSTM is good at preserving long-term dependencies, but it has difficulty with handling short-term memory. Conversely, GRU is better suited for capturing short-term dependencies due to its simpler architecture. Combining both models into a hybrid model allows for the benefits of both architectures to be utilized, resulting in better performance.

Additionally, the hybrid model has more parameters, which allows it to learn more complex relationships between the input data and the output, thus making it better suited for capturing the underlying patterns in the stock market data.

[1]Train data explained variance regression score
[2]Test data explained variance regression score

REFERENCES

[1] 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), Warsaw, Poland, 2005, pp. 192-196, doi: 10.1109/ISDA.2005.85.

[2] 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, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2009.02.038.

[3] D. Zhang and M. Li, "Research on the Stock Return Predictability with Combination of Machine Learning," 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2021, pp. 5-10, doi: 10.1109/MLBDBI54094.2021.00009.

[4] Ya Gao, Rong Wang, Enmin Zhou, "Stock Prediction Based on Optimized LSTM and GRU Models", Scientific Programming, vol. 2021, Article ID 4055281, 8 pages, 2021. https://doi.org/10.1155/2021/4055281

[5] Github link- https://github.com/badal091/EE769

[6] Link to the video