Machine Learning Models for Stock Trend Prediction

DS203 Project

Parth Shrivastava
dept.of Physics
Indian Institute of Technology, Bombay
Mumbai, India
190260033
190260033@iitb.ac.in

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

Kaushik Singirikonda
dept.of Physics
Indian Institute of Technology, Bombay
Mumbai, India
190260025
190260025@iitb.ac.in

***Abstract*—With the growing awareness about the stock market among the general public today, it has become a very burning reign for analytics. It is important to make informed decisions when it comes to investing. In the following paper we've analysed 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.**

I. Introduction

Investing is the act of allocating resources, usually money, with the expectation of generating an income or profit. Legendary investor Warren Buffett defines investing as "the process of laying out money now to receive more money in the future." Investing ensures present and future financial security. It allows individuals or companies to grow their wealth and at the same time generate inflation-beating returns. One also benefits from the power of compounding.

Numerous studies have shown that over long periods of time, stocks generate investment returns that are superior to those from every other asset class. Due to the increasing interest in the stock market and the availability of huge amounts of accessible data, there is a huge incentive to predict the performance of a stock over a period of time.

In our current model, we are limiting ourselves to the use of a single feature of the stocks, i.e. close price, in order to predict the value of the stock. We use data manipulation techniques in order to generate specific features to take as input for the given model. The detailed description of the Data analysis techniques used for data manipulation has been provided below.

II. Related Works

- Kim and Han [1] built a model as a combination of artificial neural networks (ANN) and genetic algorithms (GAs) with discretization of features for predicting stock price index. The data used in their study include the technical indicators as well as the direction of change in the daily Korea stock price index (KOSPI). They used the data containing samples of 2928 trading days, ranging from January 1989 to December 1998, and give their selected features and formulas. They also applied optimization of feature discretization, as a technique that is similar to dimensionality reduction. The strengths of their work are that they introduced GA to optimize the ANN. First, the amount of input features and processing elements in the hidden layer are 12 and not adjustable. Another limitation is in the learning process of ANN, and the authors only focused on two factors in optimization. While they still believed that GA has great potential for feature discretization optimization. Our initialized feature pool refers to the selected features. Qiu and Song also presented a solution to predict the direction of the Japanese stock market based on an optimized artificial neural network model. In this work, authors utilize genetic algorithms together with artificial neural network based models, and name it as a hybrid GA-ANN model.
- Hassan and Nath [2] applied the Hidden Markov Model (HMM) on the stock market forecasting on stock prices of four different Airlines. They reduce states of the model into four states: the opening price, closing price, the highest price, and the lowest price. The strong point of this paper is that the approach does not need expert knowledge to build a prediction model. While this work is limited within the industry of Airlines and evaluated on a very small dataset, it may not lead to a prediction model with generality. One of the approaches in stock market prediction related works could be exploited to do the comparison work. The authors selected a maximum 2 years as the date range of training and testing dataset, which provided us a date range reference for our evaluation part.
- Lee [3] used the support vector machine (SVM) along with a hybrid feature selection method to carry out prediction of stock trends. The dataset in this research is a sub dataset of NASDAQ Index in Taiwan Economic Journal

Database (TEJD) in 2008. The feature selection part was using a hybrid method, supported sequential forward search (SSFS) played the role of the wrapper. Another advantage of this work is that they designed a detailed procedure of parameter adjustment with performance under different parameter values. The clear structure of the feature selection model is also heuristic to the primary stage of model structuring. One of the limitations was that the performance of SVM was compared to back-propagation neural network (BPNN) only and did not compare to the other machine learning algorithms.

III. Data set and Feature Engineering

The dataset contains data regarding stock prices of TATA motors for each day, extracted from the yfinance library. The data file was cleaned for any empty entries, and there were 2434 rows of data. Data set starts from 2012 as the stock was split into multiple sub parts and the price fell, therefore to avoid ambiguity when analysing prices.

While implementing the models(apart form LSTM), we have notices that taking the closing price values of the stocks does not result in good predictions, we have decided to use the following data manipulation techniques in order to generate workable features.

Therefore, instead of only taking the closing prices, we have taken additional set of features like the moving averages and bollinger bands.(apart from LSTM model)

A. Moving Averages

In order to identify the growth of the stock over time, instead of taking the stock price per day, which would result in a far more volatile and inefficient model, we have decided to use concept of moving averages as the features for training the model, this ensures that while predicting the stock prices of the previous days affect the stock price of the predicted day. As this time ordered data is already accounted for in the LSTM model, we can directly take the closing prices in this case, we also note that this gives us good values for predicting the stock prices.

We take the shifted Moving average in order to predict the values as the actual Moving average is not available and is up for prediction.

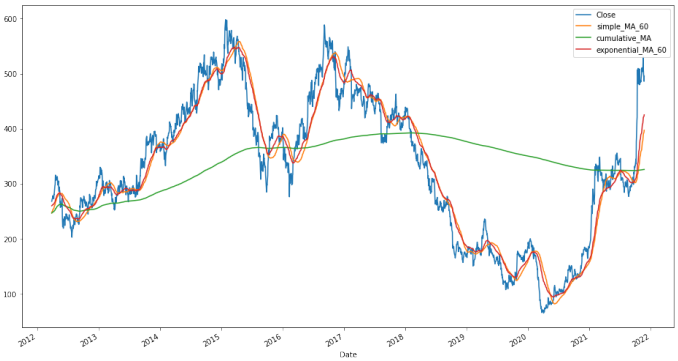

Fig. 1. Moving averages plotted along with the actual data.

B. Bollinger Bands

Bollinger bands represent the Moving averages while also accounting for the standard deviation. This ensures that sudden changes in data are also accounted and would prevent biasing. This training feature makes our model more robust and ensures that the sudden change in stock prices do not completely overthrow out our training parameters. We have taken 4 Bollinger bands, namely a) single standard deviation positive,b) single standard deviation negative, c) double standard deviation positive and d) double standard deviation negative.

They also smoothen the data in order to enable to clear trend prediction rather than small changes in the stock prices over a small period.

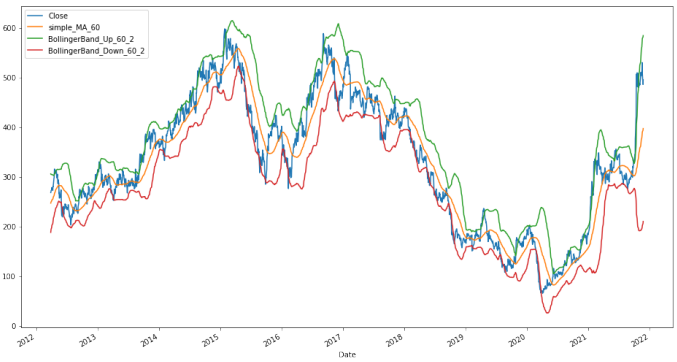

Fig. 2. Bollinger Bands plotted along with the actual data.

IV. Model Description

A. Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. The goal is to minimize the mean squared error between the observed targets and the predicted targets in our model.

The only features being used are the dates and the training is based on the previous stock values, therefore one may not expect the model to be very precise.

Fig. 3. Prediction using Linear Regression.

B. k-Nearest Neighbours

The K-Nearest Neighbor is a model that can be used for both classification and regression. Inn code, gridsearch to find the best parameters. Then the training data is fit into the model and the test data is predicted, and rms value is obtained.

Fig. 4. Prediction using kNN model.

C. Linear Regression with updated features

We use the same Linear Regression model, but this time we also include the Moving Average and Bollinger band features which training. We can clearly see that this results in a better model. Barring LSTM, this is the best model that we have obtained to predict values of stocks.

Fig. 5. Prediction using Linear Regression with updated features.

D. k-Nearest Neighbours with updated features

We use the same k-Nearest Neighbours model, but this time we also include the Moving Average and Bollinger band features which training. We can clearly see that this results in a better model. This update though shows an improvement over taking just the closing prices, it is not as good as the linear regression model as shown above, this provides us with mediocre results.

E. auto-ARIMA

Auto-arima which is a time series model was used. A time series model, in this case, is expected to perform better than linear regression in theory. It describes the correlation between data points and takes into account the difference of the values. In the code, the data was loaded, preprocessed and made

Fig. 6. Prediction using kNN model with updated features.

univariate, and fit in the model. On plotting it was observed that it performed better than linear regression but not better than linear regression with updated features.

Fig. 7. Prediciton using auto-ARIMA model.

F. LSTM

Long Short Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. These networks have chain like structures(time ordered) and each repeating structure has four single neural networks, having very specific functions.

- First neural network determines which data is to be thrown away from the incoming cell state
- The Second Neural network determines the new data that is to be stored in the cell state. Firstly there is a sigmoid layer which decides what values need to be updated, followed by a tanh layer which converts it to a vector in order to add it to the cell state.
- The third layer updates the cell states with the previously derived features
- The final layer determines what output is to be given from the cell state.

We have used the LSTM package from tensorflow.keras in order to implement the model.

V. Results

In our work, we have compared the efficiency of different models in order to predict the future stock prices. We have noticed that without any data manipulation (Moving Averages

Fig. 8. Predictions using LSTM.

and Bollinger Bands) the LSTM model outperforms all the other models by a huge margin. This is a result of the fact that an LSTM is an optimized form of a Recurrent Neural Network which is already time ordered and hence considers all the previous values while predicting the current price of stock. We also notice that when we take the features with moving averages and Bollinger bands, the linear regression model also gives us a prediction with minimal error from the actual price of the stock, though not as good as the LSTM model but is a lot quicker and cheaper computationally.

The LSTM model is able to predict the sudden rises as well as the sudden falls in stock prices and infact does well on a long run. This is not suitable for short term prediction. There a lot of factors that influence the value of a stock such as government policies, pandemics, economic collapse etc. Our model does not take into account these phenomenon, thus there is always a margin of error to this.

VI. Future Work

There may be some variables outside of previous stock prices that affect the prices. To account for some of them, in addition to our model, we can use NLP models to identify key terms in the contemporary news that changed may have affected the model and add features to our current model and get better performance in predictions.

References

[1] https://ieeexplore.ieee.org/document/1578783
[2] https://www.sciencedirect.com/science/article/pii/S0957417400000270?via%3Dihub
[3] https://www.sciencedirect.com/science/article/pii/S0957417409001560?via%3Dihub
[4] https://towardsdatascience.com/machine-learning-for-stock-prediction-a-quantitative-approach-4ca98c0bfb8c