Stock Price prediction using ML

Anonymous ACL submission

1 Introduction to Stock Market

Stock market is a medium where the financial and business activities around the companies are funded by the retail investestor. By retail investors I mean common people like us, they need not be large scale investors where they are supposed to put a huge chunk of money in the one company but can invest in a diversified way. Before we understand the ML of predicting stock price, we must understand how the stock market actually works. When an investor is looking forward to making an investment in a company where he/she thinks they'll make a good return, they will have to buy a publicly traded company from the stock exchange. Currently, there are 2 stock exchanges in India : NSE (National Stock Exchange) and BSE (Bombay Stock Exchange). To buy a stock of a company, a person will have to approach a stock broker where the investor asks for placement of his/her order on the stock exchange. Only stock brokers are allowed to deal with the stock exchange and the retail investor can't directly approach the stock exchange. The broker takes some commission and he will place the order on the stock exchange and you get the stock of the company in your demat account which are maintained by the CDSL (Central Depository Services Limited) and NSDL (National Securities Depository Limited) in India. When any stock of a company is purchased, there must be a seller and vice versa. For a trade of a stock to happen on the stock exchange, it's essential that both the buyers and the seller agree on the same price. So, let's say that a seller quotes the price of the stock for X and then there is a buyer looking for the same stock for the same price X then a trade will happen. What drives the price of the stock is the agreement on the price of the stock by both buyers and sellers.

2 Problem

It is a well known fact that if we can predict the forthcoming price of a stock, anyone can make a handsome profit. Let's say that a person is predicting that the price of stock will go up in the next day or week or month or even next year, he/she could buy the stock and make a profit on his/her investment if the prediction turns out to be true and if the prediction turns out to be false, the person will have to book loss. The problem we are looking to solve is to predictit the future prices of the stock using ML techniques so that someone could get a good return on his/her investment. This problem can be tackled by ML to some extent but it cannot guarantee 100% results as the stock market is not only driven by previously existing patterns but also from many other factors. Some of the factors which drives stock price are:

1. Company's financial performance for the year.
2. Company's fundamentals : debt, equity, number of shareholders, book value of the share etc.
3. Trend of the stock and also the stock market: Uptrend, downtrend, sideways market etc.
4. Good/bad news for the company.
5. Management of the company.

There can be even more reasons for stock price volatility but these are the main ones. The problem lies in predicting the values of stock from the past trend behaviour stock.

3 Importance of Project

You may have noticed that people call the stock market "Satta bazar" in India or they refer to it as a gambling. Many people resist investing in the

stock market in India by giving an argument that they might lose their money. They don't understand that investing is the most predictive thing if they are well aware of the rules. If this project is able to make good predictions of the future stock prices then by this project, it may cast to the people that investing is a good thing and the stock market is a very good place to compound their money. If more and more people invest in the stock market, the better it is for the companies and countries as they will have money to fund their activities at the same time giving a good return to the investors. So, an investor is indirectly helping his/her country grow its GDP. The motive of the project is to get a better insight of stock prices and to make good predictions of the future prices so that it can benefit people and at the same time making more people aware of the fact that investing is not gambling. Also, since the stock market is tightly linked with the GDP of a country, we can essentially see whether a country is moving towards a recession or depression or whether the country is doing extremely well.

4 Literature Review

There are multiple papers which are published in the domain of stock market forecasting using machine learning. We need to follow the following steps :

1. Collection of data.
2. Processing of the data.
3. Building the model.
4. Testing the model $\rightarrow$ prediction of stock prices.

Collection of data is dealt in "Data collection" section and processing of data is dealt in "Data analysis" section.

4.1 Building the model

Building the model for prediction requires input of stock prices and input data can be provided in several ways to enhance the prediction of stock prices. Following subsection deals with these methods.

4.1.1 Direct Input

When we put a request to get the historical data of what we get from the API is : opening, closing, high, low price and volume of the stock on that particular day. Opening price is the price at which the stock opened that day, similar is the closing price. High price is the highest price at which the stock was traded, equivalent is the low price. Volume is the number of shares which changed hands on that particular day. So, we have these 5 numbers which are ready to be feeded into a RNN (Recurrent Neural Network). RNN is a machine learning technique based on neural network techniques. We will be using LSTM (Long Term Short Memory) layers for RNN. Now, in RNN, our input layer can be of size 4 and the output can be simply a single node which will represent the stock price.

4.1.2 Timestamped Input

For the stock market market data, we have a time series data which means our data changes with time i.e. let's say we have the following data for the average stock price of the day :

Date (Oct 2020)	12	13	14	15	16	17
Stock Price	47	40	46	42	41	44

Since the stock price in one day is very tightly linked to the stock price of the previous day, we will have to consider time stamped feature data. From the above data, we just can say that the stock price will be 10 on 18th October 2020. It must lie within some rough range of 40 - 50. For that matter of fact we consider time stamped data. So, for the above prescribed stock quotations, let's say we timestamp the data by 2, then our instances will look like :

X (features)		Y (labels)
47	40	46
40	46	42
46	42	41
42	41	44

Although for the project we will be using more than 2 timestamp (usually 100) but here just for illustration, we are showing timestamps of 2. Now, the above data can be distributed to training and testing samples and accordingly we can proceed to the next part of the algorithm which is to build the model. Now, since we have timestamped the data by 100, we have 100 features ready to feeded

into the model. Similar to the previous part, we can build the same RNN model but with an input layer of 100 with an output node of 1 representing the stock price.

5 Evaluation Metric

By now, it should be pretty clear that the task of predicting the future stock prices of a company is a regression task. We feed in some inputs to the model and model provides us with a single value i.e. the predicted stock price. We can compare this predicted stock price with the actual stock price in multiple ways : RMSE (Root Mean Squared Error), MSE (Mean Squared Error), Mean Absolute Error etc. The lesser these errors are, the more accurate our prediction of stock price is.With every instance we feed into the model, the model will gain accuracy and lower the error, so more the data, better the accuracy.

6 Data Collection

There are multiple API (Application Programming Interface) available on the internet using which we can collect the required data of stock prices. I'll have to use multiple APIs if I want to consider stock prices of the companies around the world because I have not found a single API using which I can get data of all the companies around the world. The APIs I have been testing are : "yahoo-finance", "google-finance", "nse" (for Indian stocks). Following is a sample of data I get for Microsoft when I run the API to get history stock quotes of the company (see last page). In the similar way we can get the financials of the company such as dividends, earnings, sales, Earning before tax, interest and depreciation (ebida) which is a ver essential feature. These points drives the stock price the most. Rather than only focusing on stock trends, we have to give equal importance to these features also.

7 Data Collection

The data we get from the APIs is the historic data of stock price in the format : open price, closing price, high and low price. Now, this data is to be transformed such as it can be feeded into the model for prediction of stock prices. To do so, we will first timestamp the data (for 100 times) $\rightarrow$ 99 of the previous historic data + some financial statements of the company will be used as a feature of an instance. The output of the model will be a single number, which is the predicted stock price. To complete this task, first we'll have to scale the stock price data and the financial data of the company between 0 to 1, so that highest stock price will be scaled to 1 and the minimum stock price will be some number greater than 0 but the closest to 0.

Date	Open	High	Low	Close	Adj Close	Vol
2017-01-03	28.950001	29.082500	28.690001	29.037500	27.548414	11512
2017-01-04	28.962500	29.127501	28.937500	29.004999	27.517582	8447
2017-01-05	28.980000	29.215000	28.952499	29.152500	27.657520	8877
2017-01-06	29.195000	29.540001	29.117500	29.477501	27.965857	12700
2017-01-09	29.487499	29.857500	29.485001	29.747499	28.222006	13424
...	...	...	...	...	...	
2017-04-24	35.875000	35.987499	35.794998	35.910000	34.216194	6853
2017-04-25	35.977501	36.224998	35.967499	36.132500	34.428196	7548
2017-04-26	36.117500	36.150002	35.845001	35.919998	34.225719	8016
2017-04-27	35.980000	36.040001	35.827499	35.947498	34.251919	5698
2017-04-28	36.022499	36.075001	35.817501	35.912498	34.218575	8344