Stock Market Analysis

Shreya Maheshwari
Department of CS&E
PES University
Bangalore, India

Vishal Sathyanarayana
Department of CS&E
PES University
Bangalore, India

Rithvik Chandan
Department of CS&E
PES University
Bangalore, India

***Abstract*—Our project finds the stocks which are the most optimal for investment at a given point of time based on their historical data. It will help pick out stocks for investment in such a way that one can minimize risks and maximize profits – by engendering a profitable stock trading decision through technical analysis of the historical data. Raw historical data of stocks was scraped from nseindia.com. Data for splits and bonuses was scraped from moneycontrol.com. We calculate CAGR, number of years of positive returns, number of years of negative returns and the returns for each year for each stock. A score was given based on the aforementioned parameters and top 250 companies were chosen. These were further filtered to find the ones which are in up trend. The result we obtained was a subset of 20 companies which are most optimal for investment (on the latest date in our dataset). Then, we forecast closing prices for each stock in our subset using 2 models: ARIMA and LSTM.**

I. INTRODUCTION

In today's day and age, trading in the stock market is an extremely popular channel of financial investment. There are various investment opportunities such as trading bonds, shares, foreign investments and precious metals to name a few. Investors try to create wealth by buying and selling their investments at a proper time. There are various advantages of investing in the stock market – investment gains, dividend income, diversification and ownership. One of the primary reasons for investing in the stock market is the chance to multiply your money. Over time, companies tend to grow in value, thereby increasing the stock prices by a proportionate amount. Investment in stable companies make profit for investors. Some stocks also provide income in the form of dividends. Dividends represent income on top of the profit that comes from selling the stock. An investor need not put all his money into one stock – he/she can diversify. This helps in building wealth by leveraging different sectors of the economy. Buying shares of a company means taking ownership stake in the company. This implies that investing in the stock market also brings benefits as those which are enjoyed by being one of the business' owners.

The key of a good investment is to invest at a suitable time with minimum risk. The aforementioned is extremely hard to determine due to the dynamic and highly volatile nature of the stock market. Indicators, as calculated using historical data, are extremely helpful in assisting an investor to execute buy and sell decisions. One cannot predict the exact behavior of a stock, as it does not solely depend on the indicators – long term trends, cyclical variations, seasonal variations and irregular movements also play a role. However, technical analysis reduces the uncertainty involved with investment decisions. Even when faced with certain unforeseen circumstances, fundamentally good stocks, ones which exhibit strong indicators, don't tend to fall as much as the rest of the other stocks in the market. Exploiting this can potentially make you high profits with a "low" risk. Even in case of a loss scenario the amount of money lost will be minimized.

Our objective is to develop a model which can indicate the stability of a stock and provide every stock with a certain score using the calculated indicators.

II. WHAT HAVE OTHERS DONE TO SOLVE THE PROBLEM

Reference [1], Kimoto, Takashi et al base their buy and sell timing prediction on Modular Neural Networks. The output of the Network indicated whether a stock needs to be sold or bought. They use a moving simulation method, where the model learns for M months and predicts next L months and the window is continually shifted. In their model, an output value > 0.5 indicated 'Buy' and < 0.5 indicated 'Sell'. However, this overlooks a third, more practical option in "holding" a particular stock.

In [2], Birgul Egeli et al implement an ANN with the traditional Back Propagation (BP) algorithm used to acquire weights of connection to predict the Istanbul Stock Exchange Market Index Values to find out if it does better than the models based on Moving Averages. Although it did do better, there are some limitations like vanishing gradient as they used the Sigmoid function as the Activation function for a neuron in the network. This could lead to slow convergence and increase the chances of getting stuck at a local minimum.

In [3], Lamartine Almeida Teixeira and Adriano Lorena Inácio de Oliveira combine technical analysis and nearest neighbor classification. They use combinations of technical indicators to such as RSI filter, stop loss, stop gain etc. to come up with 22 features to be used as inputs to a k-NN classifier. This however results in a requirement for large computational power as a k-NN classifier does all of its processing during test time as the "distances" need to be calculated w.r.t 22 features along with cross-validation to get an ideal value for k and might still result in lower accuracy as compared to say a Support Vector Machine.

A commonly seen problem with ANN's is that it ignores noise and various non-stationary characteristics in the data. Training of a Back-Propagation algorithm is difficult due to the noise as it is hard to incorporate market variables into the model without making certain assumptions as a result, the network might end up always predicting the most common output.

Unlike an ANN, the Autoregressive Integrated Moving Average (ARIMA) is a model which is used explicitly for time series data and has a certain structure to it. In [4], Ayodele A. Adebiyi. Et al build an ARIMA model to provide investors successfully with a short-term prediction to aid them with their decision.

This model however does not take into account any technical indicators as it is completely dependent only on

previous target value data. This could lead to inaccurate results in the long run.

Due to this, researchers tend to move towards hybrid solutions by synergistically combining both, the ANN model and the ARIMA model. Studies have been conducted to see if a hybrid model yields better results than individual models [5]. Similarly, Ping-Feng Pai Et al [6] showed that promising results can be obtained from a hybrid model of ARIMA and SVM [6].

Another very effective non linear approach to forecast time series is LSTM. LSTMs were developed to deal with the exploding and vanishing gradient problems that can be encountered when training traditional RNNs. Roondiwala et al. in [7] has used RNN-LSTM model on NIFTY-50 stocks with 4 features (high/close/open/low price of each day). They have used 21 days window to predict the next day price movement

III. PROBLEM STATEMENT

To develop a model based on historical data of the stocks listed on the National Stock Exchange in order to find sustainable stocks for safe investments.

A. Collection of Data

Historical data of the stocks listed on NSE was collected from nseindia.com using a web scraping software called GetBhavCopy. One text file was generated for each day that trading was carried out. Each file consists of 7 columns as follows: the name of the company, the date (in yyyymmdd format), open price, highest price on that day, lowest price on that day, close price and volume. We have collected data for the last 25 years i.e. from 1995 to 2019. We have consolidated all the data from these files into one CSV file by running a python script. This CSV file is used as our main dataset.

Information about splits and bonuses in stocks was collected by web scraping from the website www.moneycontrol.com.

B. Cleaning of Data

The data that was collected from the software contained equities as well as indexes. Our primary focus is equity stocks, so the records containing indexes were dropped. Furthermore, some records contained '-' and some records had 8 columns which were dropped as well.

On analyzing the graphs of price vs date of a few companies, we realized there were certain steep drops in the same. On investigation, this was attributed to splits and bonuses in company stocks.

Splits: All publicly traded companies have a set of shares that are outstanding. A stock split is a decision by a company's board of directors to increase the number of shares that are outstanding by issuing more shares to current shareholders. A stock's price is also affected by a stock split. After a split, the stock price will be reduced since the number of shares outstanding has increased. For example: in a 2-for-1 split, the share price will be halved.

Bonuses: Bonus shares are shares given to existing stockholders in proportion to the number of shares they hold. A 1:1 bonus means that a shareholder will get one share for each share held by him. Usually, after the bonus issue, the share price of the company gets adjusted according to the bonus ratio. For example, if the price before bonus is Rs 200 and a company issues bonus shares in the ratio of 1:1, the post-bonus share price will be Rs 100, which means that the total market value (2 x Rs 100 = Rs 200) remains the same.

We handled this by finding the date after which a stock had been split and multiplied the closing price with the appropriate split factor.

We accounted for a stock splitting multiple times by multiplying by the latest split factor after the corresponding date. All records of companies whose data was missing in between have been dropped.

The effect of this is a sudden drop in the closing price of the stock over a very short period of time(~1day). This has been illustrated in Figure 1 in the form of a line graph.

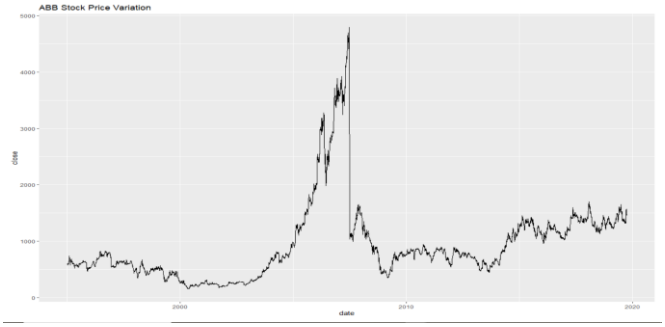

Figure 1: General Trend for ABB stock between 1995 and 2020.

IV. PROPOSED SYSTEM

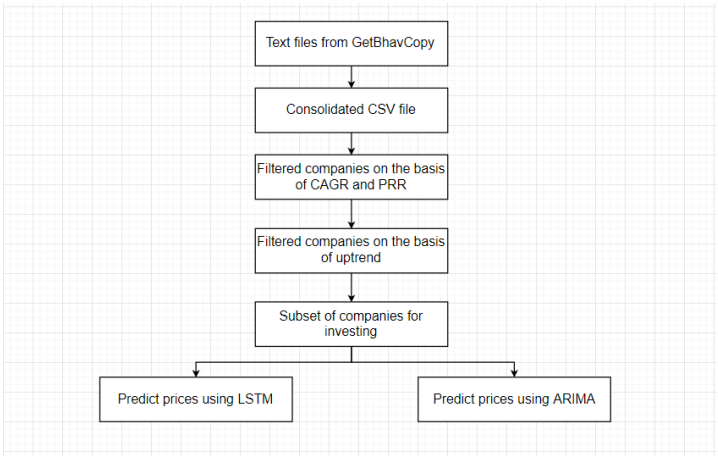

Figure 2: Block diagram of the general process followed

V. FEATURE ENGINEERING

i) CAGR: We calculated the Compound Annual Growth Rate for each company since inception. This would give us an overview of how much an investment, if made on day 1, would be worth today.

$$CAGR = (EB/BB) \wedge (1/n) - 1 \qquad (1)$$

EB: Ending Balance

BB: Starting Balance

n: Number of years

Higher the value of CAGR, more the company has grown over the years of its existence.

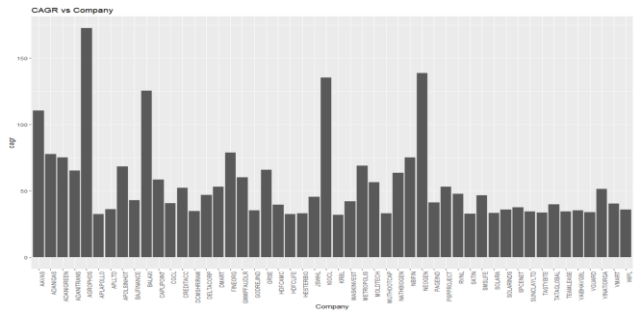

Figure 3: Bar graph for CAGR of top 50 companies.

ii) Positive returns ratio: is defined as the number of years of positive returns divided by total number of years

$$PRR = NYPR/n \qquad (2)$$

this gives us an insight into how many years has the company actually performed well. If it has performed well more than 75% of the time – it makes it extremely reliable.

The number of positive years was found as the closing price on the last trading day of the year and the closing price on the first trading day of the year. If their difference was positive, one was added to the number of years of positive returns. If not, one was added to the number of years of negative returns. Total was the sum of these two values.

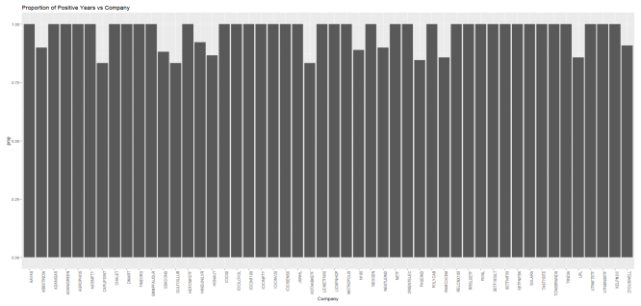

Figure 4: Bar graph for Proportion of positive returns of top 50 companies

By just looking at the CAGR and the aforementioned ratio, one cannot say that the company is good for investment as of today because this result takes into account all the data that is available for that company. More recent data should be given more weightage.

iii) Weighted Returns: to solve the above issue, we calculated another score that takes into account the time that has passed by since the return was generated. We gave each return a weight of *1/no_of_years_passed* and then calculated a score

A combination of these factors would help us determine which companies are stable – those with high value of CAGR and proportionately higher number of years of positive returns as compared to the total number of years.

Moving Average (MA) creates a series of averages of different subsets of the complete dataset. Represented as MA(x) where x denotes the subset size. The first element of the moving average is obtained by taking the average of the initial fixed subset of the number series. Then the subset is modified by "shifting forward"; that is, excluding the first number of the series and including the next value in the subset. It is used extensively with time series data to smooth out any short-term fluctuations in the data so as to highlight the long-term trend.

We want to look into companies that have given a high CAGR as well as good returns in the recent years - so we calculated another score which took into account both of these parameters. We then filtered out the top 250 companies. Among these companies, the best ones to invest in presently, would be the ones in uptrend which is calculated using the following algorithm –

If MA_50 > MA_100 > MA_200:

uptrend = 1

else:

uptrend = 0

MA_50: 50 day Moving Average

MA_100: 100 day Moving Average

MA_200: 200 day Moving Average

We consider only those companies which have the uptrend flag set to 1.

	company	cagr	w_returns	pos	neg	score	flag
1	MOLDTKPAC	25.01311	33.62301	4	1	29.31806	1
2	N100	22.63962	48.00286	8	1	35.32124	1
3	ZYDUSWELL	23.12901	64.45780	10	1	43.79341	1
4	NESTLEIND	26.13476	65.92028	9	1	46.02752	1
5	PETRONET	26.73138	71.22165	12	4	48.97651	1
6	GSKCONS	33.20563	66.97604	15	2	50.09083	1
7	MARICO	40.48017	61.35788	13	2	50.91902	1
8	HINDUNILVR	46.56926	67.31981	12	1	56.94454	1
9	ASIANPAINT	21.12668	95.71493	19	6	58.42081	1
10	NIITTECH	17.07257	103.75377	12	4	60.41317	1
11	HEXAWARE	35.75515	94.24921	14	4	65.00218	1
12	PIDILITIND	22.87393	112.48351	19	6	67.67872	1
13	BERGEPAINT	27.21060	120.66647	21	4	73.93854	1
14	ABBOTINDIA	44.12495	104.19604	9	1	74.16050	1
15	CAPLIPOINT	62.83921	92.72845	5	1	77.78383	1
16	PIIND	56.19719	99.83562	8	1	78.01640	1
17	DIVISLAB	38.53633	120.04381	13	4	79.29007	1
18	KOTAKBANK	39.31847	120.74495	14	3	80.03171	1
19	HONAUT	31.45589	129.37907	13	2	80.41748	1
20	VINATIORGA	59.17592	142.69898	9	2	100.93745	1

Figure 5: Subset of 20 companies

VI. MODELLING

We have forecasted the stock closing prices of these companies using 2 models: ARIMA and LSTM to give us a general idea of how the stock may vary in the near future.

A. ARIMA

In statistics and econometrics, and in particular time series analysis, an Autoregressive Integrated Moving Average (ARIMA) model is a generalization of an Autoregressive Moving Average (ARMA) model. This model is applied to either better understand data or to forecast future points in the series.

This model is applicable when the initial time series shows evidence of non-stationarity, where an initial differencing step can be applied one or more times to eliminate the non-stationarity. There are seasonal and Non-seasonal ARIMA models that can be used for forecasting. We use the Non-

seasonal one as there is no seasonal component in our data as the market is highly volatile.

A series is said to be stationary if it's mean and variance are constant over a period of time. This model needs 3 parameters: P, D and Q, where p is the periods to lag for. This helps adjust the line that is being fit to forecast the series. D is used to transform the given series into a stationary time series. D refers to the number of differencing transformations needed to be applied. Q refers to the lag of the error component of the model. Error component is that component which is not explained by trend or seasonality of the series.

The Auto correlation (ACF) and partial auto correlation (PACF) plots help determine the values of the aforementioned parameters. The Autocorrelation function plot will let you know how the given time series is correlated with itself. The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags. If the PACF plot drops off at lag n, then use an AR(n) model and if the drop in PACF is more gradual then we use the MA term.

In Figure 6, we can see that ARIMA's forecast after 30 days has an upward drift which is indicated by the dark blue line in the elliptical region (which is a very good approximation of how much the true value can fluctuate from the forecasted value) and these companies will almost always follow this trend. Out of the filtered 20 companies, we can further filter them by looking for such upward drifts in the ARIMA forecast graphs. Figure 7 indicates an ARIMA graph which just forecasts a constant value with no drift. There is no way to determine from this if the close price of the stock is going to increase or fall after 30 days. These are slightly riskier but still offer a fairly good chance of gaining profit as it is a company that was filtered as being reliable in nature.

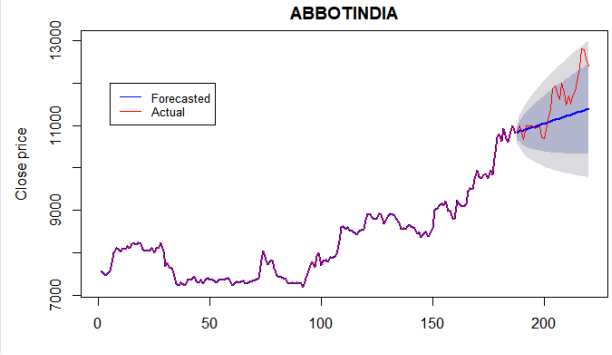

Figure 6: ARIMA forecast for ABBOTINDIA

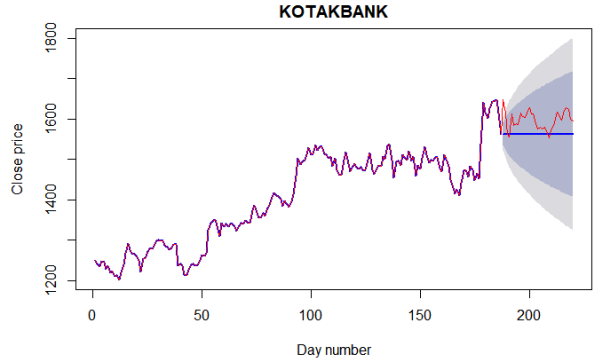

Figure 7: ARIMA forecast for KOTAKBANK

B. LSTM

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

LSTM networks particularly excel at making predictions based on time series data, since there can be lags of unknown duration between important events in a time series event.

We divide the data in such a way that our training set for each company always has a fixed length. This allows for the creation of fixed-shaped tensors and therefore yields more stable weights. We use Adam's optimizer as a replacement to the traditional stochastic gradient descent which helps handle sparse gradients on noisy problems.

Figure 8: LSTM prediction against actual values observed for ABBOTINDIA

VII. RESULTS

Our solution to the problem was modeled on available data until October 4th 2019. To test the reliability of our model, we define a parameter Profit Percentage.

$$PP = ((FCP - ICP)/ICP) * 100\ \% \qquad (3)$$

FCP: Final Close Price
ICP: Initial Close Price

The initial close price is the sum of all close prices of 1 stock of every company in our subset of 20 companies as of October 4th 2019. The final close price is the sum of all close prices of 1 stock of every company in our subset as on November 21st 2019.

We further filtered this data to only 3 companies: N100, PIIND, ABBOTINDIA. These are the companies which showed an upward drift in the ARIMA forecast of 30 days. Profit percentage was calculated in both of these cases. The results are tabulated as follows:

Stocks invested (October 4th 2019)	Profit Percentage (November 21st 2019)
1 stock of each company in our 20-company subset	4%
1 stock each of N100, PIIND, ABBOTINDIA	14%

Table 1: Results

The profit percentages are calculated based on investments made over just one month i.e. from Oct 4th to November 21st. Indian banks provide interest rates in the range of 4.50% p.a. to 7.0% p.a. on fixed deposits and in the range of 3.50% p.a. to 7.0% p.a. on savings accounts. Investment in sustainable stocks with upward drift (predicted by ARIMA) suggested by our model provides a 14% return on the principal amount in just 1 month with low risk. This is much higher than what a bank offers in a whole year.

Some problems that we encountered with LSTM were: The LSTM model takes very long to train attributed to its high computational requirement. This model worked extremely well with our historic data split into training and testing sets as we can see in Figure 8. However, we were unable to reproduce the same for real time test data. Moreover, LSTM would only allow us to work with very few companies due to its high computational requirements. The aim of our project was to find a set of companies which are suitable to invest for profits at minimal risks and not just one company. This was better modeled by ARIMA, which takes considerable lesser time to train as it does not use computationally heavy components such as neural networks. We have obtained more than satisfactory results using this model as illustrated in Table 1.

VIII. Conclusion

Our model successfully finds sustainable stocks to invest in at a given time which are highly likely to give good returns in reasonable time with minimal risk. Our results show that the profits obtained are double in 1 month than what any bank can provide in a year.

IX. Individual Contribution

Shreya Maheshwari PES1201700025: Handling bonuses and splits in the data, Calculating CAGR, returns and score for companies for filtering. Data collection and cleaning.
Vishal Sathyanarayana PES1201700183: Scraped bonus and split data from moneycontrol.com. Modeled stock data using LSTM. Calculated viability of these models by finding the profit considering real-time data.
Rithvik Chandan PES1201700014: Found uptrend using simple moving averages and modeled the data using ARIMA, worked on Report.

References

[1] Kimoto, Takashi, et al. "Stock market prediction system with modular neural networks." *1990 IJCNN international joint conference on neural networks*. IEEE, 1990.

[2] Birgul Egeli, Asst. "Stock market prediction using artificial neural networks." *Decision Support Systems* 22 (2003): 171-185.

[3] Teixeira, Lamartine Almeida, and Adriano Lorena Inacio De Oliveira. "A method for automatic stock trading combining technical analysis and nearest neighbor classification." *Expert systems with applications* 37.10 (2010): 6885-6890.

[4] Ariyo, Adebiyi A., Adewumi O. Adewumi, and Charles K. Ayo. "Stock price prediction using the ARIMA model." *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*. IEEE, 2014.

[5] Merh, Nitin, Vinod P. Saxena, and Kamal Raj Pardasani. "A comparison between hybrid approaches of ANN and ARIMA for Indian stock trend forecasting." *Business Intelligence Journal* 3.2 (2010): 23-43.

[6] Pai, Ping-Feng, and Chih-Sheng Lin. "A hybrid ARIMA and support vector machines model in stock price forecasting." *Omega* 33.6 (2005): 497-505.

[7] M. Roondiwala, H. Patel and S. Varma, "Predicting stock prices using LSTM," International Journal of Science and Research (IJSR), vol. 6, no. 4, pp. 1754-1756, 2017.