Big Data, Big Money

Abhishek Singh, Jay Kachhadia, Shloak Gupta, Tanya Shrivastava

Problems and Objectives

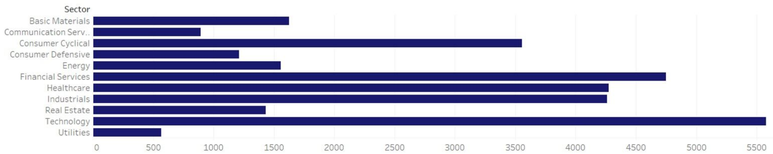
Stock Markets have a turn-over of thousands of billions of dollars which make it the prefered choice for people to invest their money in. There are 15,000 stocks traded in the U.S. on a daily basis which makes the problem a voluminous one. Along with volume, the volatile nature of the stock prices can make it difficult for a layman to weigh in all the factors contributing towards the rise and fall of a stock price.

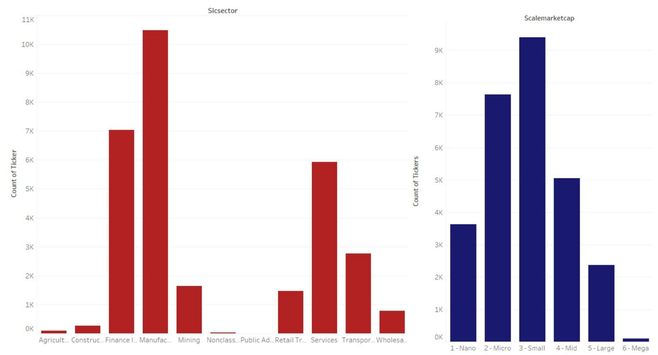
The objective is to help an investor make a decision supported by historical data and current trends from the stock market by predicting the closing prices of a day along with stock recommendations for financial investing.

Data Description

The dataset currently contains stock market data for more than 6,500 active tickers, with historical data since the year 1998. It has 7 million observations for 49 different variables. It was acquired from Quandl and was collected by an independent research firm, Sharadar.

Data Exploration





Benefits of Algorithmic Trading

Algo-trading provides the following benefits:

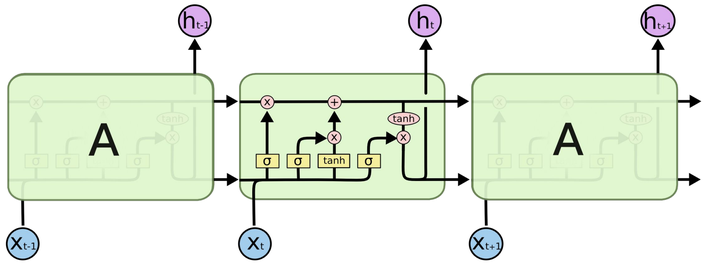
- Trades are executed at the best possible prices.
- Trade order placement is instant and accurate (there is a high chance of execution at the desired levels).
- Trades are timed correctly and instantly to avoid significant price changes.
- Reduced transaction costs.
- Simultaneous automated checks on multiple market conditions.
- Reduced risk of manual errors when placing trades.
- Algo-trading can be backtested using available historical and real-time data to see if it is a viable trading strategy.
- Reduced possibility of mistakes by human traders based on emotional and psychological factors.

Model Description

Linear Regression - Linear regression is used for finding linear relationship between a target and one or more predictors. For the scope of stock market data, we use Closing Price Lag(1) ,Closing Price Lag (2), 3 day Moving Averages, 7 day Moving Averages, Volume Lag(1), Momentum, and Rate of Change with respect to each date, as described below.

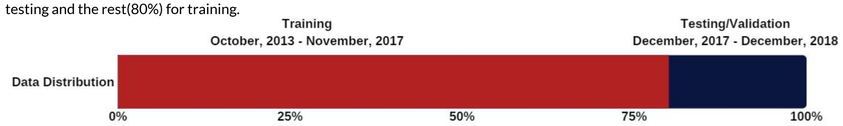
Dependent Variable	Definition/Formula
Closing Price Lag(1)	Closing Price from Yesterday.
Closing Price Lag (2)	Closing Price from day before yesterday.
3 Day Moving Average	Average of Closing Price for the past three days.
7 Day Moving Average	Average of Closing Price for the past week.
Volume Lag(1)	Volume of stock sold from yesterday.
Momentum	Difference between the Closing Price of yesterday and five days ago. [Closing Price Lag(1)-Closing Price Lag(5)]
Rate of Change (ROC)	Ratio of Closing Price from Yesterday and Closing Price from five days ago. [Closing Price Lag(1)/Closing Price Lag(5)]

LSTM - Long Short-term Memory is a of Recurrent Neural Network(RNN) with memory. By definition, Recurrent networks, take as their input the current data and what they have perceived previously in time. In case of LSTM, the key is the cell state, the horizontal line running through the top of the diagram. All the changes are made to this cell state. The first layer of LSTM decides which data to keep from the previous cell state, the second layer selectively adds current input to the cell state, and the last selects the output to be carried to the next state. We use previous closing price to predict the closing price of the present day.



Model Comparison Metrics

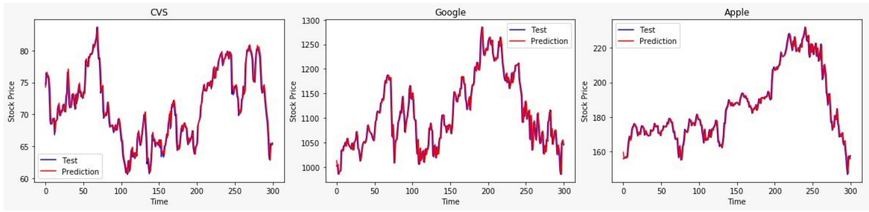
As the data is in time series format,we cannot have a random split. We have used the data of the last 6 months(20%) for validation and testing and the rest(80%) for training.

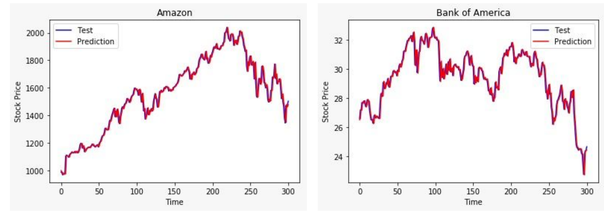


The generalization performance is measured in both the models by their RMSE.

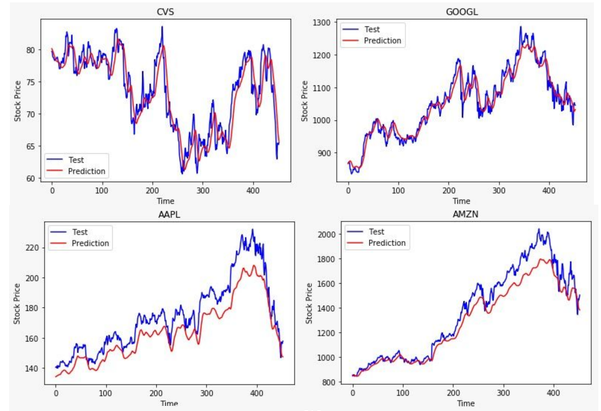
Result - Prediction Performance

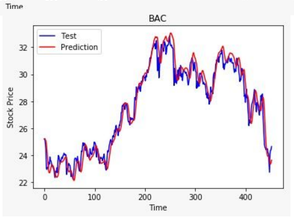
Linear Regression Performance on Test data





LSTM Performance on Test data





Result and Comparison

Ticker	Linear Regression-RMSE	LSTM-RMSE
CVS	1.365716623	8.106236054
GOOGL	18.33497695	139.7331102
AAPL	3.19028808	32.34350611
AMZN	34.55183514	470.9848764
BAC	0.4369531	4.539236938

Conclusion

Through this project we were able to predict the stock prices using its historical closing prices up to a great extent. We were also able to recommend stocks by using our models for different stocks and finding the best performing stock by looking at the returns. Lower RMSE in Linear Regression clearly indicates that adding features like Momentum, Moving Average helps in reducing the prediction error. Going forward we would include these features in the LSTM and add more features like type of industry, market cap, the volume traded and overall index performance. Dream of making millions is still a little far away but through this project, we took a big leap towards it.