# Intraday Stock Market Prediction

**Pradeep Kumar**
**pradeep17174@iiitd.ac.in**

**Himanshu**
**himanshu17153@iiitd.ac.in**

**Vikas Agnihorti**
**vikas17207@iiitd.ac.in**

## 1 Introduction

Stock exchanges are the institutions that facilitate the trading of different financial instruments (i.e., equity, stocks/shares, derivatives, bonds, options, futures) between market participants. Average Daily Volume of SP 500 index is 100 Billion Dollars (no of shares traded* price of the asset). This attracts the best minds around the world from every field to predict which stocks will provide the best profits in future. In this project, we are also doing something similar but we are using previous data of that particular stock. Predicting the direction of stocks is one of the widely studied fields in Economics, Mathematics and Computer Science. The investors want to predict the right time for buying and selling stock to maximize their profit or return on investment. The volatile nature of the stock market gives the opportunity for traders to invest for short and long term. The prices of stocks change every second which makes it very difficult to predict the price accurately. Stock Market is a zero-sum game, for every winner, there is a loser. That is why researchers try to develop models strategies to help navigate the rough waters of the stock market and get out of it with profits. The goal of investors is to maximize profits and reduce the risk of loss for every position by leveraging various Machine Learning Models.

This project attempts to estimate the direction of movement of stock prices of TESLA stock for a small time interval (i.e every five minutes) by using various features such as open, close, low high, trading volume. We have implemented SMA (Simple Moving Average) and Crossover strategy based on Fibonacci lookback periods due to their confluence with financial markets. We will implement various machine learning models such as Linear Regression, Logistic Regression, Support Vector Machine (SVM) with different kernels, RNN such as Long short-term memory(LSTM) and convolutional neural network (CNN) to predict the stock price movement.

## 2 Project Importance

- Helping investors to maximize profit while minimising their risk involved.
- Better informs the trader community through results obtained from Machine Learning models
- Hedge Funds that manage billions of dollars of investors money can use the predictions to improve their fund's ROI (Return on Investment).
- Machine learning will help to generate a short term strategy for HFT (High-Frequency Trading) bots developed by quants.
- The difficult part of the project is that we are predicting future stock prices using previous data available publicly to everyone.
- For the trader and investors predicting price is very important but due to volatility in the stock market makes it difficult to predict correctly.

## 3 Literature Survey

The amount of data available presently is very high that is why much research is being done on applying machine learning to aid the stock price trading by predicting the stock price movement with good accuracy.

Logistic regression can also be used to predict the stock price movement and, [1] Logistic regression has a particular advantage over other models in stock market prediction. Before the advancement of Deep Learning popular research topics were using SVM in making predictions in the financial area. In [2] L. J. Cao and Francis E. H. Tay used SVM for financial time series stock price prediction also its result signify that SVM can give good price predictions. Also in [3], SVM is used to predict the direction of movement of the stock price in the short term.

Apart from SVM, other machine learning models are also used in the financial domain. Sine stock price data is time sequence data hence the use of RNN models such as LSTM make more sense than using CNN. In [4] Convolutional Neural Network is used to predict the stock price movement. From the results of [4], we got that using above mentioned models give good accuracy on different stock price datasets.

Many researchers have also focussed on not only predicting the price movements but also the profit which can be generated by Machine Learning based trading.

## 4 Dataset and its Pre-Proccessing

Our aim is to predict intraday price movements of Tesla stock. A dataset with Daily price quotes of all stocks and indexes are readily available on BSE, CBOE, ICE, NYSE websites. We require TSLA stock price quotes for every minute from opening to closing bell of any trading day. This involves a significant amount of computational and storage power to gather data from exchanges in order to store it offline for data analysis purposes. Intraday Stock Data for 1min, 5min, 15min timeframes is available at a cost of 200 USD per Gigabyte of data. We collected our data from AlphaVantage API (free plan with a limit of 3 requests per second, 500 requests per day) which collects its data from IEX exchange. It includes 5 minute time interval trading data of TSLA stock (both spot futures market) from 01-04-2020 to 31-07-2020. The data consists of Date, Time, Open, Close, Low, High, Volume attributes. Two types of analysis used to predict the direction of stock price movement: Fundamental and Technical Analysis. Fundamental Analysis leverages the information of Revenue, Earnings, P/E ratio, underlying value, FED rate, financial statements of a stock. Technical analysis believes that individuals can make investment decisions based on historical price data of an asset since the market follows expansion and contraction patterns. Various technical indicators exist in the financial industry to analyse the price direction of a stock, i.e., Simple Moving Averages(SMA), Exponential Moving Averages(EMA), Volume, Relative Strength Index(RSI). A significant amount of time was invested to learn the economics and working of these indicators to best predict the stock prices in future. Pre-processing of raw data collected from API is done by merging data of different months into a single unit. The average price for every time interval is calculated from high and low price attributes. SMA(i) involves average prices with a lookback window of i entries, i.e SMA (100) is the average price of the last 100 entries in the dataset. Features SMA for 5, 8, 13, 21, 55, 89, 150, 200 time periods are added into our dataset. SMA/EMA Crossover indicates the momentum of a trend is bullish (upwards) and bearish (downwards) direction of an asset. When a low period time frame moving average crosses above a higher period moving average, it constitutes a buy signal (LONG) and if low period time frame moving average crosses below the higher period moving average, it indicates a sell signal (SHORT). All combinations of SMAs crossovers have been added to the dataset.

Labelling of data is done to maximise the profits obtained from trading on signals predicted by our Machine Learning Models. Label 0 indicates selling the stock and Label 1 indicates buying the stock at a particular time period. We calculate labels by looking at the next entry/row of data if it is greater than the current price, we wish to buy (Label=1) and if it is lower than the current price, we wish to sell (Label=0). Our strategy involves both Long Buying and Short selling of the stock rather than the traditional only Long Buy setup.

## 5 Data Analysis

After Pre-processing, we analyse the data to help understand the nature of the Dataset. The price of Tesla Stock from April 2020 to July 2020 has been plotted in Fig1. This indicates that the stock is in a strong bullish trend (upwards). The Volume of stock traded is given in Fig2 showing the number of shares/stocks traded in a time interval.

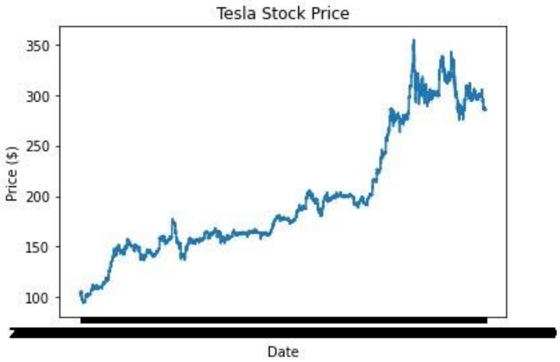

Figure 1: Price of Tesla Stock from April 2020 to July 2020

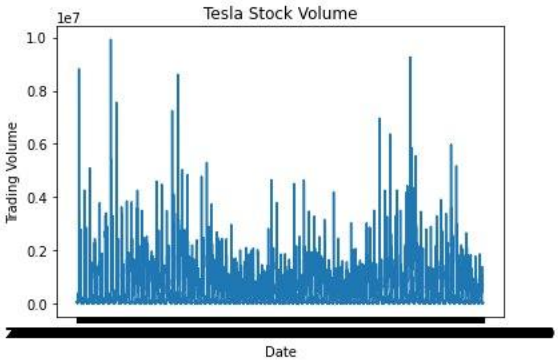

Figure 2: Volume of TESLA stock.

After Preprocessing, In these scatter plot (figure 3.4), we can see that both classes are highly overlapping and we expect the prediction to be vague and low. The possible classifiers that we can use are Logistic regression, SVM. RNN such as Long short-term memory(LSTM) and convolutional neural network (CNN)

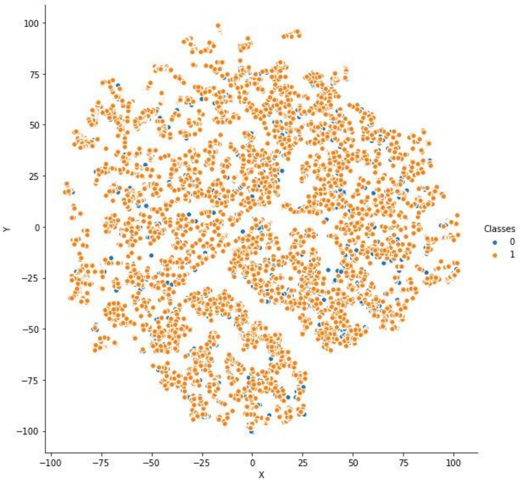

Figure 3: 2-D TNSE Scatter Plot

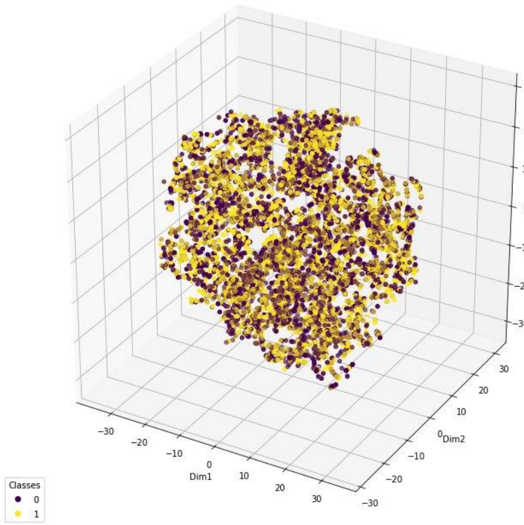

Figure 4: 3-D TNSE Scatter Plot

## 6 Methods

### Features Selection

Feature Selection is an essential step in classification using Machine Learning algorithms to achieve better results. There are various reasons to reduce the features of dataset if possible, i.e., Curse of Dimensionality- as dimensionality of data features increases; the model complexity, running time and errors increase along with them. Also, when we input poor quality features into our model, the efficiency of the model is negatively affected. Simple models that achieve the task in hand are preferred over high order and complex models. We have used lasso regularization method and Extra Tree Classifier(Extremely Randomized Trees) to reduce our feature space of the dataset.

- Original Features : Date, Time, Open Price , Close Price , Low , High, Trading volume, Average Price, M.A.s and E.M.A.s (total 79 features)

After applying Features Selection Algorithm, the list of important features are given below:

- Open Price, Close Price, Trading Volume, Time, Average Price.
- Exponential Moving average: 5-E.M.A., 8-E.M.A., 13-E.M.A., 21-E.M.A., 55-E.M.A. 89-E.M.A, 150-E.M.A, 200-E.M.A.
- Moving Average: 5-MA, 8-MA, 13-MA, 21-MA, 55-MA 89-MA, 150-MA, 200-MA.
- Crossovers: 21/55 MA-Cross, 8/13 MA-Cross, 55/89 EMA-Cross, 21/150 EMA-Cross, 89/200 MA-Cross.

**Baseline Model**

The main objective of this project is to predict the next 5 min price movement. This is a classification problem since we need to predict whether the stock price will go up or down in the next timeframe (5-min) . Hence, we will use Logistic Regression without regularisation as our baseline model. To further increase the accuracy we have used the following machine learning models:

### 6.1 Logistic Regression

Logistic Regression is one of the most sought model after linear regression machine learning model, applicable for both regression and classification problems. We have used logistic regression as our baseline model and we have used l1 and l2 Regularisation.

### 6.2 SVM

Support vector machine is one of the best binary classifier which is a supervised machine learning model. It divides data into two classes for classification. The task of SVM to find best possible boundary which divides data into two classes. This boundary line is known as **hyperplane**. Distance between two classes or ”margin”. Margin are the distance from hyperplane to the closest data point from each class. In SVM we tries to maximize the margin.

In this project we have use linear, polynomial SVM as figure 3,4 can see that data is very randomly distributed. We have used three kernel linear, rbf (Radial Basis Function Kernel), polynomial(Sigmoid). After training simple model. We have performed grid search over C for the optimisation of the problem.

$$min_{w\xi b}\frac{1}{2} \times \|W|^2 + C \times \sum_{i=1}^{n} \xi^{(i)}]$$

$$\text{such that} \quad y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi^{(i)}; \quad \forall i \in \{1,\ldots,n\}$$

### 6.3 Navies Bayes

Naive Bayes (NB) Classifier assumes that all features are independent of each other. Since, the features are independent, Bayes Theorem can be applied and is used to predict instances unknown to the model.

### 6.4 Random Forest Classifier

RF Classifier is a supervised Machine Learning algorithm based on Ensemble learning. It combines multiple Decision Trees leading to “forest” of trees and can be used for both regression and classification problems.

### 6.5 Recurrent neural network (RNN) model

We will be using different layered LSTM and will experiment with different activation functions. Different kinds of regularization will be used to find the best value of hyperparameter. We have tuned hyperparameter to get maximum possible accuracy.We have performed grid search over various parameter and hyperparameter We have used three types of LSTM single layer, multilayer and GRU. 150 hidden units in our Neural Network.

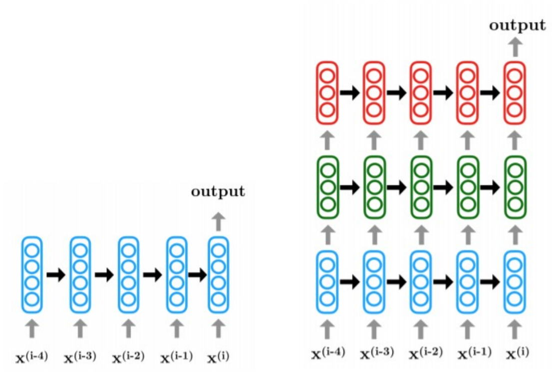

Figure 5: Left: Single Layer RNN, Right: Multilayer RNN

### 6.6 CNN Model

Finally we have used convolutional neural networks to classify and different regularization will be used to tune the model. Also for tuning all the hyperparameters we have using Gridsearch.

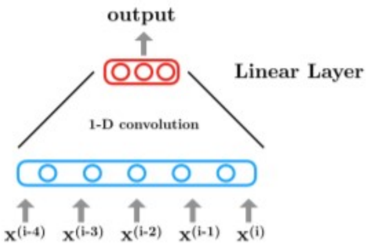

Figure 6: 1-D Convolution Neural Network

Currently we plan to experiment with the above mentioned models, but we may add some more models to the list if time permits.

## 7 Results

### 7.1 Portifolio Performance

To measure the success of any trading strategy, the performance of portfolio is considered, i.e., how much profits investors can cash-in executing this trading strategy over a period of time. Our focus is to execute fast buy and sell orders at intra-day timeframe which can be done using APIs given by most trading exchanges in the world. The predictions made by our Machine Learning models are purely directional (buy/sell) and it's hard to quantify it in terms of real profits. Hence, we need to build an algorithm that takes model predictions as input and generate value of our portfolio over time. The concept of LONG and SHORT is fundamental to proceed further in our portfolio analysis. When we LONG an asset (TSLA stock in this scenario), we buy the asset at current market price to create a new position or closing an earlier position. While we SHORT an asset, we wish to sell the asset at current market price to close an existing position or create a new position altogether. The algorithm will LONG (buy) the Tesla stock when predicted value of our model is towards bullish direction (Label=1) and will SHORT (sell) the stock when predicted direction is bearish (Label=0). It will continue to hold the position until the direction flips to the other side, i.e., 1 to 0 or 0 to 1. After the direction is inverted, our position will be squared off to cash in FIAT currency. Trading Exchanges provide leverage services to its customers who are willing to take more risk. If leverage is enabled, the results after a trade is completed are amplified both ways, negative and positive. We assume there is no leverage in our portfolio analysis. Also, we go all in for every trade executed through our algorithm, i.e., when signal is bullish, we spend all of the cash/share reserves into buying TSLA stock and when signal is bearish, we sell all our cash/share reserves into selling TSLA stock at the given price. Also, we assume that transaction cost for every position taken in the market is zero. For every ML model, the principal amount of cash in the beginning is Rs 1,00,000 (no leverage) and portfolio performance is evaluated over testing set. We have given the results for various Machine Learning models along with their training and validation performance in the table.

| Performance of various model | | |
|---|---|---|
| Model Used | Train | Test |
| Logistic Regression | 0.591 | 0.579 |
| SVM (R.B.F) | 0.588 | 0.515 |
| SVM (Linear) | 0.618 | 0.600 |
| SVM (Poly.) | 0.593 | 0.494 |
| LSTM(Single Layer) | 0.512 | 0.512 |
| LSTM(Multi Layer) | 0.513 | 0.511 |
| LSTM(GRU) | 0.526 | 0.501 |
| CNN | 0.511 | 0.506 |

### 7.2 Statistical Performance

From the above table, we observe that accuracy of the training set is higher than validation set as model tends to learn training data.

In all above models, SVM and Logistic Regression in performing better than various Deep Learning Neural Networks. Since, data is randomly distributed, Linear SVM is performing better than various Neural Networks like LSTM (Single Layer, Multilayer, GRU) as the dataset consists of around 15000 data points, we can expect this kind of performance from Neural Network.

## 8 Conclusion

In this project, we have implemented various machine learning models to predict the future direction of price of Tesla (TSLA) stock. We have collected data from the Vantage API to get the recent data of TSLA Intraday stock of every 5-minutes. We converted stock market price prediction problem from regression to a classification problem. Further, we have used various machine learning techniques such as SVM, Logistic Regression, LSTM(Single layer, Multilayer, GRU), CNN. Out of the models we experimented with, SVM (Linear) is performing best on our available dataset. We had limited amount of data available.

## 9 Future Work

Intuitively, we expect to see a positive relation between actual profits and accuracy of the model, but on the contrary we observed that the higher accuracy of the model doesn't correspond to more profits. This is due to the direction of price predicted by model for some time periods can lead to more volatility in the portfolio than the others, affecting it both positive and negative ways. We can increase the size of data to feed into Neural Networks which can lead to better results. We can also try to predict the price of upcoming time frame.

## 10 Contributions

- Pradeep Kumar : Literature Survey, SVM(Linear), LSTM(Multilayer), CNN
- Himanshu : Literature Survey, Logistic Regression, SVM, LSTM
- Vikas Agnihotri : Literature Survey, Data Preprocessing, Feature Selection, SVM(rbf), LSTM(GRU)

## References

[1] J. Gong and S. Sun, ”A New Approach of Stock Price Prediction Based on Logistic Regression Model,” 2009 International Conference on New Trends in Information and Service Science, Beijing, 2009, pp. 1366-1371, doi: 10.1109/NISS.2009.267

[2] L. J. Cao and F. E. H. Tay, ”Support vector machine with adaptive parameters in financial time series forecasting,” in IEEE Transactions on Neural Networks, vol. 14, no. 6, pp. 1506-1518, Nov. 2003.

[3] Kercheval, Alec N., and Y. Zhang. “Modelling high-frequency limit order book dynamics with support vector machines.” Quantitative Finance 15.8(2015):1-15

[4] J. F. Chen, W. L. Chen, C. P. Huang, S. H. Huang and A. P. Chen, “Financial Time-Series Data Analysis Using Deep Convolutional Neural Networks,” 2016 7th International Conference on Cloud Computing and Big Data (CCBD), Macau, 2016, pp. 87-92

[5] Hegazy, Osman, Soliman, Omar S. Salam, Mustafa A (2013) Machine Learning

Model for Stock Market Prediction. Faculty of Computers and Informatics, Cairo University, and Higher Technological Institute (H.T.I), 10th of Ramadan City, Egypt

[6] Scik-learn,tensorflow and various python documentation

### *Appendix*

**S&P 500**: containing the market capitalization-weighted index of 500 largest publicly traded US companies.

**Trading Volume**: Total number of contracts/shares traded between buyers sellers en period of time.

**SMA**: Simple Moving Average is the average prices with a lookback window of n time periods from current entry in the data.

$$SMA_{(i)} = \frac{(P_t + P_{t-1} + \ldots\ldots + P_{t-i})}{i}$$

**EMA**: Exponential Moving Average

$$EMA_n = (P_t - EMA_{i-1}).\frac{2}{n+1} + EMA_{n-1}$$

Link to Code : Click Here