Machine Learning Engineering Nanodegree

Capstone Project

Stock Price Prediction using LSTM

*Narrendran Ramesh*

*September 25, 2019*

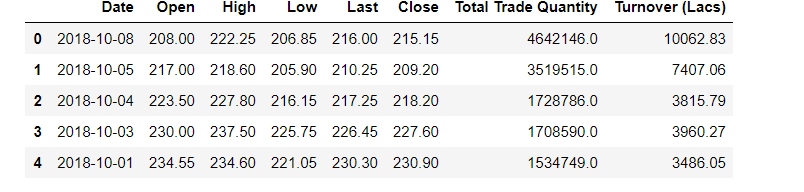
**Introduction**

Time series analysis is a specialized branch of statistics used extensively in fields such as Econometrics & Operation Research.

Time Series is being widely used in analytics & data science. This is specifically designed time series problem for you and the challenge is to forecast traffic.

**Dataset:**

Dataset used is the data for ‘Tata Global Beverages’from Quandl



There are multiple variables in the dataset – date, open, high, low, last, close, total\_trade\_quantity, and turnover.

* The columns Open and Close represent the starting and final price at which the stock is traded on a particular day.
* High, Low and Last represent the maximum, minimum, and last price of the share for the day.
* Total Trade Quantity is the number of shares bought or sold in the day and Turnover (Lacs) is the turnover of the particular company on a given date.

The profit or loss calculation is usually determined by the closing price of a stock for the day, hence we will consider the closing price as the target variable

**Techniques Applied are**

1. Moving Average
2. Linear Regression
3. k-Nearest Neighbors
4. Auto ARIMA
5. Prophet
6. Long Short Term Memory (LSTM)
7. **Moving Average**

The predicted closing price for each day will be the average of a set of previously observed values. Instead of using the simple average, we will be using the moving average technique which uses the latest set of values for each prediction.

1. **Linear Regression**

The most basic machine learning algorithm that can be implemented on this data is linear regression. The linear regression model returns an equation that determines the relationship between the independent variables and the dependent variable.

The equation for linear regression can be written as:

https://i0.wp.com/s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2018/10/stock11.png?zoom=1.25&resize=234%2C41&ssl=1

Here, x1, x2,….xn represent the independent variables while the coefficients θ1, θ2, …. θn represent the weights.

1. **Linear Regression**

Linear regression is a simple technique and quite easy to interpret, but there are a few obvious disadvantages. One problem in using regression algorithms is that the model overfits to the date and month column. Instead of taking into account the previous values from the point of prediction, the model will consider the value from the same date a month ago, or the same date/month a year ago.

1. **k-Nearest Neighbors**

Based on the independent variables, kNN finds the similarity between new data points and old data points

1. **Auto ARIMA**

Parameter tuning for ARIMA consumes a lot of time. So auto ARIMA which automatically selects the best combination of (p,q,d) that provides the least error is used.

1. **Prophet**

Prophet, designed and pioneered by Facebook, is a time series forecasting library that requires no data preprocessing and is extremely simple to implement. The input for Prophet is a dataframe with two columns: date and target (ds and y).

Prophet tries to capture the seasonality in the past data and works well when the dataset is large.

1. **Long Short Term Memory (LSTM)**

LSTMs are widely used for sequence prediction problems and have proven to be extremely effective. The reason they work so well is because LSTM is able to store past information that is important, and forget the information that is not. LSTM has three gates:

The input gate: The input gate adds information to the cell state

The forget gate: It removes the information that is no longer required by the model

The output gate: Output Gate at LSTM selects the information to be shown as output

