Analyse News Using Machine Learning for Movement in Stock Prices

Mentored by- Aditya Anavkar

Stock Prices

WHAT ARE STOCK PRICES?

The stock market is a market that enables the seamless exchange of buying and selling of company stocks. Every Stock Exchange has its own Stock Index value. The index is the average value that is calculated by combining several stocks. This helps in representing the entire stock market and predicting the market's movement over time.

The stock price is the highest amount someone is willing to pay for the stock, or the lowest amount that it can be bought for.

NOTE:
The stock market works through a network of exchanges – like New York Stock Exchange or Sensex.

WHY ARE WE INTERESTED IN IT?

There are many factors involved in the prediction of stock prices – physical factors vs. physiological, rational and irrational behavior, etc. All these aspects combine to make share prices volatile.

Our interest in the value of company stocks arises from the fact that by predicting it ahead of time we can take certain decisions and hence gain significant profits.

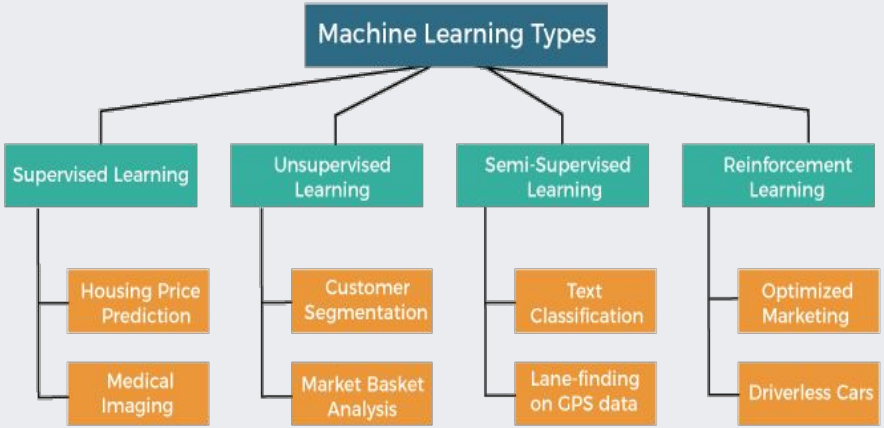
1 Machine Learning

and its application in stock markets

Introduction to ML

Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed. The basic steps involved in the process are listed below:

Get Data → Clean, Prepare & Manipulate Data → Train Model → Test & Validate → Improve



Machine learning allows the user to feed a computer algorithm an immense amount of data and have the computer analyze and make data-driven recommendations & decisions based on only the input data. It is used in various fields today.

“

A baby learns to crawl, walk and then run. We are in the crawling stage when it comes to applying machine learning.

-Dave Waters

”

ML in Stock Price prediction

Modeling turbulent structures requires machine learning algorithms which are capable of finding hidden structures within the data and predict how they will affect them in the future.

Machine learning helps by analyzing large chunks of data, spotting significant patterns and generating a single output that navigates traders towards a particular decision based on predicted asset prices.

Time Series Forecasting

A time series dataset is different as time series adds an explicit order dependence between observations; a time dimension. This additional dimension is both a constraint and a structure that provides a source of additional information.

Time series analysis provides a body of techniques to better understand a dataset. Perhaps the most useful of these is the decomposition of a time series into 4 constituent parts:

1. Trend

2. Seasonal variation

3. Cyclical variation

4. Irregular variation

Time series forecasting Models

NAIVE

ARIMA

LSTM

Naive forecasting

It uses the previous observation directly as the forecast (without changes). It is also called ‘persistence forecast’ as the prior observation is persisted.

Autoregressive Integrated Moving Averages

A model that uses the dependency between an observation & a residual error from a moving average model applied to lagged observations.

Long Short Term Memory

LSTM cells are used in recurrent neural networks that learn to predict the future from data sequences of variable lengths.

Naive forecasting

Naïve forecasting is the technique in which the last period's sales are used for the next period's forecast without predictions or adjusting the factors. Forecasts produced using a naïve approach are equal to the final observed value. The naïve forecasting method is the easiest of all methods and it is suitable for finance and sales departments because it ensures that these departments work to improve the company.

In naïve forecasting there are no calculations or formulas, only an assertion of the actual sales numbers. In some cases, naïve forecasting can accurately predict situations, while others can be problematic because it considers only the previous period to forecast the next period. Thus, historical sales data is the foremost requirement for naïve forecasting and factors such as seasonality are not considered. Oftentimes, salespeople will use naïve forecasting to make goals as a way to make sure they are always either improving or maintaining their contribution to the company.

ARIMA

ARIMA stands for AutoRegressive Integrated Moving Average, a model which is a class of statistical models for analyzing and forecasting time series data.

It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

A pure Auto Regressive (AR only) model is one where Yt depends only on its own lags. That is, Yt is a function of the 'lags of Yt'.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

where, Yt-p is the lag p of the series, βp is the coefficient of lag p that the model estimates and α is the intercept term, also estimated by the model.

A pure Moving Average (MA only) model is one where Yt depends only on the lagged forecast errors.

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

where the error terms are the errors of the autoregressive models of the respective lags. The errors ϵt is error from the following equation -

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_0 Y_0 + \epsilon_t$$

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms.

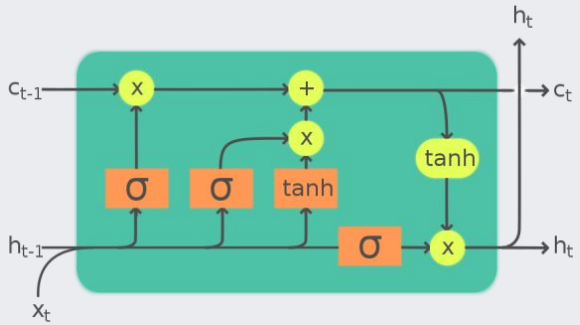
$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

Predicted Yt = Constant + Linear combination Lags of Y (upto p lags) + linear combination of Lagged forecast errors (upto q lags)

LSTM

The main idea behind LSTM cells is to learn the important parts of the sequence seen so far and forget the less important ones. This is achieved by the so-called gates, i.e., functions that have different learning objectives such as:

1. A compact representation of the time series seen so far
2. How to combine new input with the past representation of the series?



3. What to forget about the series?
4. What to output as a prediction for the next time step?

LSTM

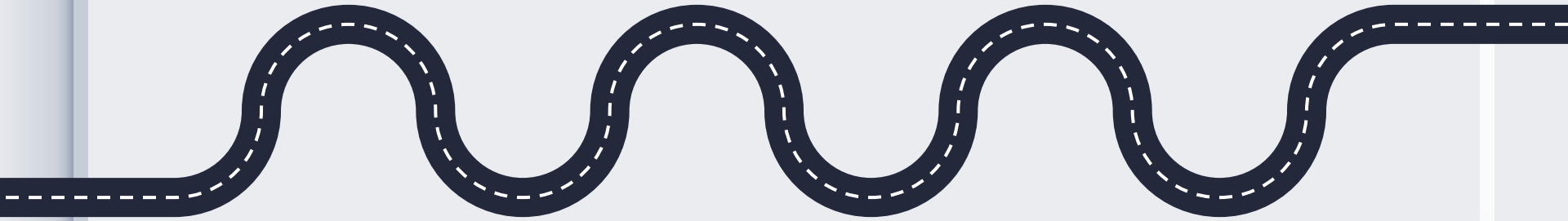
Designing an optimal LSTM based model can be a difficult task that requires careful hyperparameter tuning. Here is the list of the most important parameters an LSTM based model needs to consider:

- How many LSTM cells are to use in order to represent the sequence? Each LSTM cell will focus on specific aspects of the time series processed so far. A few LSTM cells are unlikely to capture the structure of the sequence while too many LSTM cells might lead to overfitting.
- It is typical that first, we convert the input sequence into another sequence, i.e. the values ht. This yields a new representation as the ht states capture the structure of the series processed so far. But at some point, we won't need all htvalues but rather only the last ht. This will allow us to feed the different ht's into a fully connected layer as each ht corresponds to the final output of an individual LSTM cell. Designing the exact architecture might require careful fine tuning and many trials.

2 Our Implementation

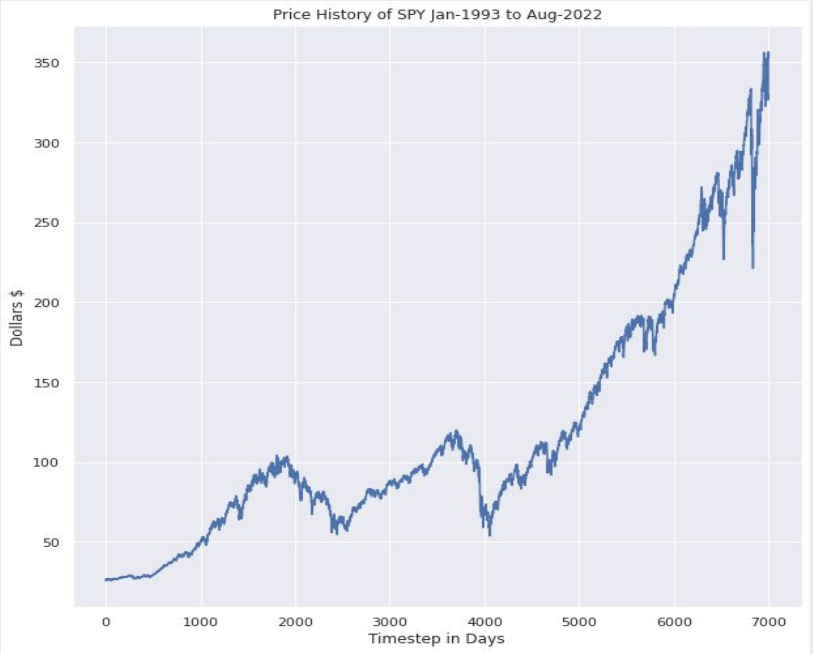
Procedure of Implementation

1. Picking of a stock market index with a long history of time series data
2. Pre-processing and splitting dataset for model
3. Reshaping of data & Implementing Naive Forecast & Simple Moving Average Models
4. Converting non stationary data to stationary data and then checking for autocorrelation for testing different ARIMA Model
5. Building & Implementing a deep LSTM model
6. Evaluating & comparing each model while minimizing prediction error



SPY

The SPDR S&P 500 trust is an exchange-traded fund which trades on the NYSE Arca under the symbol (NYSE Arca: SPY). SPDR is an acronym for the Standard & Poor's Depositary Receipts, the former name of the ETF. This fund is the largest and oldest ETF in the world. The trust seeks to provide investment results that, before expenses, correspond generally to the price and yield performance of the S&P 500 index.



SPY

The data was pulled using the yfinance API. The time period of the data is the entire existence of SPY ETF in January of 1993 until today's date of 20/08/2022.

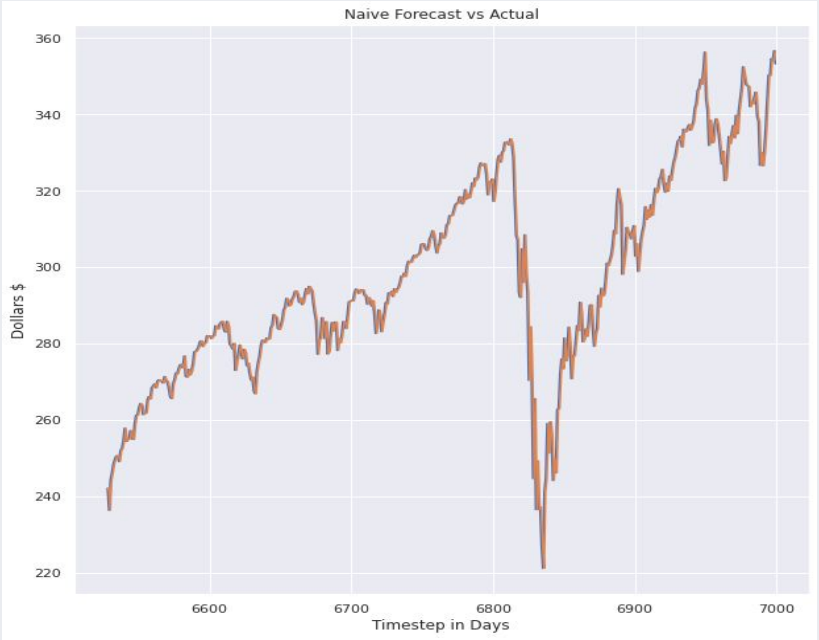
This project was a univariate time series focused on predicting the close price of the next day through various methods. The data is in daily time steps. As we can see the data is just shy of 7,000 data points. The graph shows the entire data range and how it is broken up into the train,validate,test segments.

	Date	Open	High	Low	Close	Volume	Dividends	Stock Splits
6900	2020-06-24	308.60	309.27	300.90	302.88	132813500	0.0	0
6901	2020-06-25	302.26	306.41	300.08	306.12	89468000	0.0	0
6902	2020-06-26	304.94	305.17	298.23	298.85	127961000	0.0	0
6903	2020-06-29	300.21	303.40	297.74	303.25	79773300	0.0	0
6904	2020-06-30	302.78	308.96	302.61	307.13	113394800	0.0	0

3 Model Results

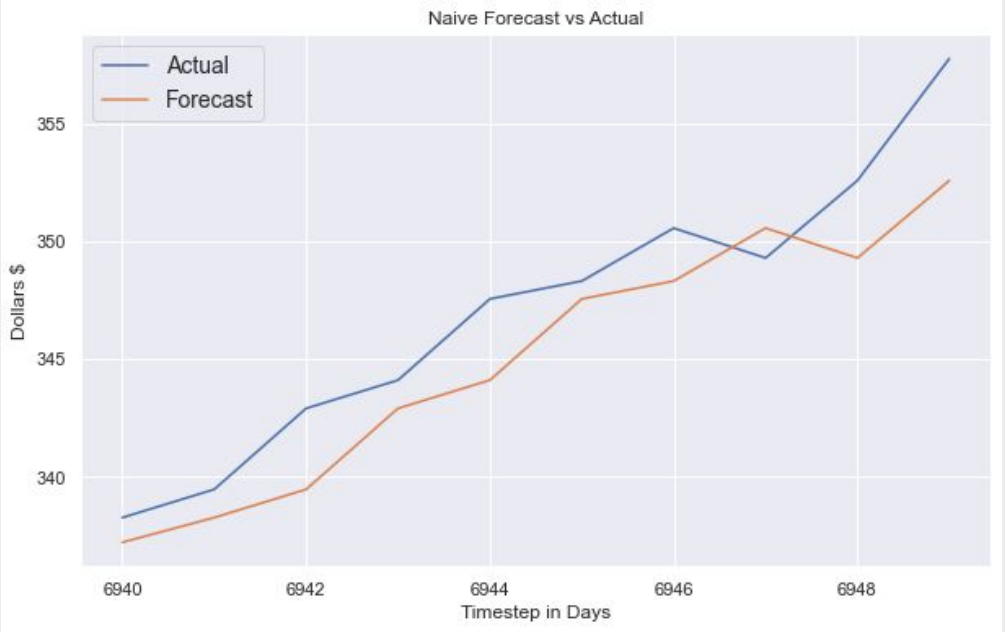
Naive Forecast Model

Naive models are naive due to the fact they don't actually "predict". The naive model uses the price from day before as it's prediction as tomorrow's price. Since there is not a large change from day to day (usually) in the stock market, this model performs really well.



Naive Forecast Model

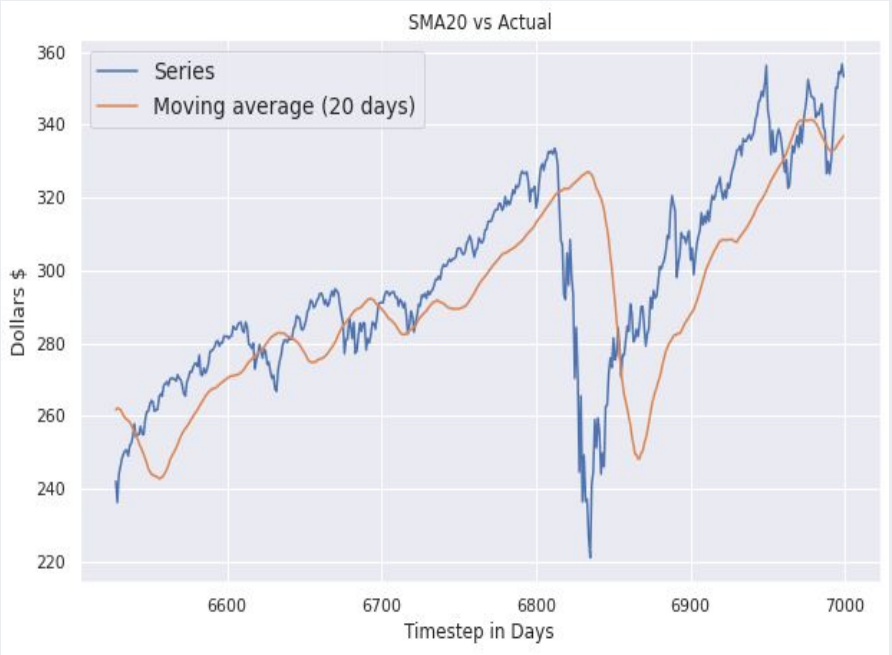
This shows a full view of the entire training period.However, since the predicted and the actual prices are so close, it is really hard to see the differences here.



Moving Average Models

Simple Moving Averages (SMA) are a way of smoothing out the noise in the data to get a better idea of which way the signal is trending. These are not good predictive models, but we wanted to showcase these models as they are often used in conjunction with other models to generate trading signals. The Naive Forecast is actually the same thing as a 1 day moving average.

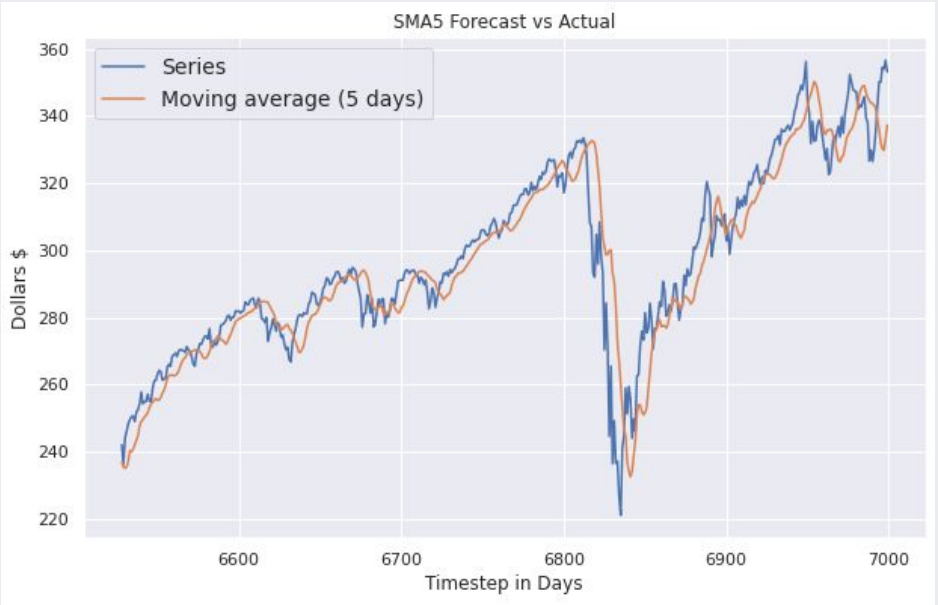
Here we can see that the 20 day moving average is not a good predictor but it is indicative of a trend. 20 Days may sound like an arbitrary number, but it is important to remember there are only 5 trading days in a week and not 7. This means 20 days is a full trading month.



Moving Average Models
(Extended version of Naive Forecasting)

There are a variety of moving average types most commonly simple moving averages or exponential. Simple moving averages take the average of the price over a certain span of time while exponential applies a weight factor to the average that decreases over time.

The 5 SMA follows the actually values much more closely than the 20 SMA as expected. As 20 days is a trading month, 5 days is a full trading week.



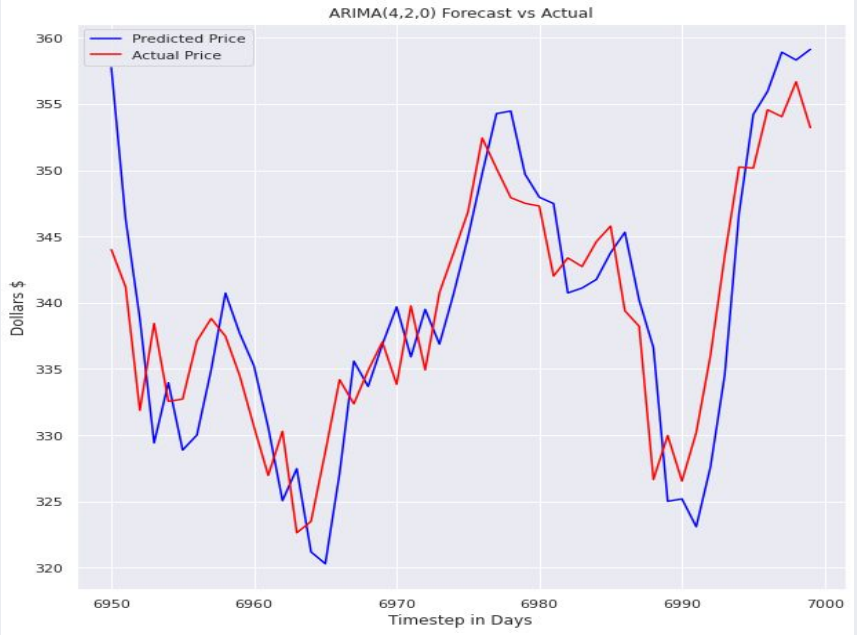
ARIMA Model

1st input:(p) uses the dependent relationship between an observation and some number of lagged observations.

2nd input:(d) stands for the differencing required to get the data to become stationary. Stationary is just a fancy way of saying the mean of the data does not change over time. The difference is simply Day(T) - Day(t-1).

3rd input:(q) is the size of the moving average window

The ARIMA model used was a (p=4,d=2,q=0) model as this was the perfectly fitting while taking care of running time.

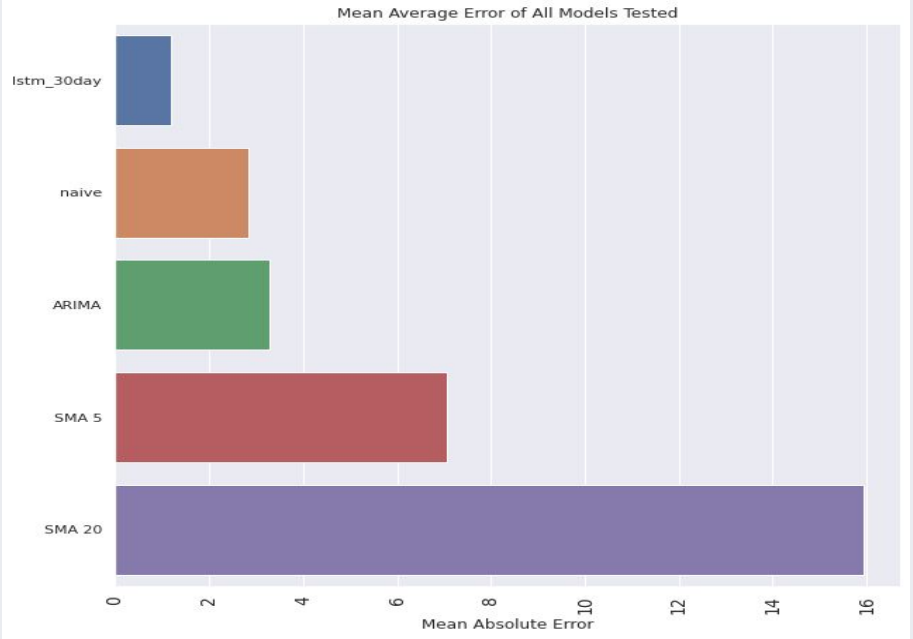


LSTM Model

LSTM stands for Long Short-Term Memory. This means the cell actually has a memory and is therefore much better at remembering patterns and making predictions based on patterns. It has emerged out to be the best performing model we have built in this project

Summary

The resulting error from all models built and tested are shown in the bar graph. As we can see the most simple model, a Naive forecast outperformed many algorithms in terms of error and the LSTM model with a 30 day rolling window outperformed all models.



Future prospects of project

Preprocessing with CNN

Augmentation techniques are applied when size of data is insufficient. In the pre-processing step such as cropping, scaling, translation, or mirroring to increase the variance, are done to increase the size of the data. These pre-processing techniques are quite effective to improve the performance of deep learning

Full CNN - WaveNet

WaveNet is a type of feedforward neural network known as a deep CNN. In WaveNet, the CNN takes a raw signal as an input and synthesises an output one sample at a time. It does so by sampling from a categorical distribution of a signal value that is encoded using μ-law companding transformation and quantized to 256 possible values.

Thank You!

Vishal Kumar
200110118

Nidhi Shaw
200260033

Priyanshu Niranjan
200110085

Atishay Jain
200110118