A Comparative Study of SVM and LSTM Deep Learning Algorithms for Stock Market Prediction

By

Sai Krishna Lakshminarayanan - M.Sc. Computer Science (Data Analytics) 18230229

NUI Galway
OÉ Gaillimh

Supervisor

Dr John McCrae

Computer Science (Data Analytics)
College of Engineering and Informatics
National University of Ireland, Galway

August 2019

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification.

Sai Krishna Lakshminarayanan

August 2019

Acknowledgements

I would like to thank my research supervisor, Dr John McCrae, who has been an excellent research mentor and role model to me throughout this research work. Thank you very much for your continuous supervision, guidance, patience and endless motivation to research which has helped me to reach the destination. I am grateful to you for your availability at any time to listen and support me for my experimental plan, troubleshooting and technical discussions.

I would also like to thank my program director Dr Enda Howley for giving me this opportunity to work on such challenging research and prove my abilities. His constant support, guidance, motivation and suggestions throughout the course helped me work and achieve in the right path for pursuing my master's successfully.

Finally, my deepest gratitude to my family and friends for supporting me throughout my journey in completing my Masters.

Abstract

The research presents a comparative study of the performance of Long Short Term Memory (LSTM) neural network models with Support Vector Machine (SVM) regression models. The framework built as a part of this study comprises of eight models. In this, 4 models are built using LSTM and 4 models using SVM respectively. There are two major datasets which are used for this research. One is the base standard Dow Jones Index (DJI) stock price dataset and another is the combination of this stock price dataset along with external added input parameters of crude oil and gold prices. This comparative study shows the best model in combination with our input dataset. Two models each are built for the base standard DJI stock price datasets by using LSTM and SVM approaches. Another two models are built for this standard DJI dataset by combining LSTM and SVM along with moving averages (MA) to provide a smoothing effect to reduce the noise. Similarly, two models are built for the combination of DJI stock price, Crude oil and Gold price dataset by using LSTM and SVM approaches. Finally, two models using LSTM and SVM are built for this combination dataset by applying the moving average technique. This approach not only gives us the best model for stock price prediction but also helps us understand the effect of crude oil and gold price data on the performance of the forecasting models built. The performance of the models is measured in terms of their Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error, Mean Absolute Percentage Error (MAPE) and R squared (R2) score values. The methodologies and the results of the models are discussed and possible enhancements to this work are also provided.

Table of Content

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
BSE	Bombay Stock Exchange
CRISP	Cross-Industry Standard Process
CNN	Convolutional Neural Network
CSV	Comma Separated Value
DJIA	Dow Jones Industrial Average
DM	Data Mining
EIA	Energy Information Administration
EPTA	Error Propagation Training Algorithm
KNN	K-Nearest Neighbours
LSTM	Long Short-Term Memory
MA	Moving Average
MAE	Mean Absolute Error
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
NASDAQ	National Association of Securities Dealers

	Automated Quotations
NSE	National Stock Exchange
NYSE	New York Stock Exchange
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
S&P500	Standard and Poor's 500
SMA	Simple Moving Average
SVM	Support Vector Machine
SVR	Support Vector Regression
USA	United States of America
VWAP	Volume Weighted Average Price

Chapter 1 - Introduction

1.1 Motivation

Stock market prediction is considered to be one of the most painstaking tasks due to its volatility. The challenge of stock market prediction is so lucrative that even a small increase in prediction by the new model can bring about huge profits. Stock prices are considered to be an essential part of the prediction. The stock price is the price of the solitary stock from the entire stocks sold by a corporation. When these stocks are bought, the person can own the corresponding portion of the public company. These stocks are sold by the founders of the corporations at a regular interval to generate new funds. The stock prices are determined based on the supply and demand for the stocks of the corporation. The stock prices increase with a surge in demand to buy the stocks and decrease with a surge in demand to sell the stocks. This demand can be linked to several external factors that determine the nature of the stock market. The external factors include the socio-economic conditions, government policies and political conditions. Therefore, when the stock prediction has to be done by analysing the external factors, it can lead to a lot of noise and volatility. This made it difficult to apply stock market prediction using simple time series or regression methods. To overcome this hurdle, researchers have taken machine learning techniques and studied their feasibility in stock market prediction. This has become more popular now with the increased interest in the fields of big data and artificial intelligence coupled with boosted computational capabilities for supporting automated methods for forecasting the stock prices (Lin,2019). Deep learning, which is considered as a sub-discipline of machine learning is becoming more prevalent now in research communities to provide solutions for forecasting problems (Misal,2019). It can be seen that a lot of sophisticated research is being carried out in this problem for predicting the stock

prices correctly with the least error possible.

1.2 Data Used

The main dataset for this project is the stock price data. For this, the Dow Jones Industrial Average (DJIA) is considered. The Dow Jones index is a price-weighted index of 30 components traded on the New York Stock Exchange (NYSE). This is considered because it covers a wide range of companies from a variety of sectors. The Dow Jones was considered due to its relatively lower number of components when compared to the S & P 500 which has more than 500 components. When these many components are analysed, it takes a huge toll on the computational requirements. The mixture of availability of enormous data and lesser computational needs made Dow Jones an ideal pick.

The primary data for the stock price data was available on the Yahoo finance website. The data was collected by writing a Python script to perform web scraping. Through this web scraping, the data is collected and stored as a comma-separated value (CSV) file. The data is taken from January 1, 2014, to December 31, 2018. It is to be noted here that only the interday trading values are obtained. This refers to the trading conducted across various days and intraday refers to trading conducted within the day. This is because the intraday trading prices are not readily available like the interday prices and it also increases the computational need and complexity. One another key information that could not be obtained is the order book. The order book has a list of buy and sell details for the corresponding company stock. This refers to the midpoint between the largest buying and smallest selling price which is considered to be important in the prediction of the closing price. It can help provide a prediction of the price utilizing the weighted average of the orders.

For the external input parameters, the crude oil dataset and gold price dataset are considered. The crude oil dataset is available in the U.S.A's Energy

Information Administration (EIA) website. The gold price dataset is obtained from the gold price website namely goldprice.org and these data are obtained from the same period of January 1, 2014, to December 31, 2018. They are obtained by a similar web scraping method. They are stored as two separate CSVs respectively. Similar to the stock data, they only contain the inter-day stock prices and do not have the intraday prices of crude oil and gold respectively

1.3 Research Problem

The main intention of this research is to find various conditions that are used in the stock market prediction and to combine them and see whether they hold like when they used separately. Three main research problem questions are pondered upon in this thesis. They are as follows,

1. Comparison of stock market prediction by using the base dataset of stock price and then a combination of the stock price with additional external parameters of crude oil and gold prices. The goal is to find whether the addition of these external parameters helped in improving the effectiveness of the stock market prediction.
2. Comparison of stock market prediction by using machine learning algorithms such as Support Vector Machine (SVM) and deep learning algorithms such as Long Short Term Memory (LSTM). The goal is to find whether the conventional way of performing the regression task with SVM holds good for stock market prediction or whether the newer concepts like LSTM provide better effectiveness in prediction.
3. Comparison of stock market prediction by using Moving Average (MA) with the SVM and LSTM algorithms on the basic stock price and advanced dataset of stock price along with crude oil and gold prices. The goal is to find out whether the addition of MA to the present models

using SVM and LSTM improves the effect when applied on the base and advanced dataset respectively.

1.4 Research Methodology

According to researchers, there are various approaches to performing the forecast for stock market prediction. One of the initial approaches that were considered was statistical and time series based ones. These included methodologies like Auto-Regressive Integrated Moving Average (ARIMA) (Khashi, 2007). These models were built mainly to deal with temporal data. But the major disadvantage with these methods was that it wasn't able to analyse the external factors that influenced the stock price data. This has caused the researchers to focus on machine learning methodologies to overcome the hurdles. The two major subdivisions in machine learning are supervised learning and unsupervised learning. In supervised learning, training data consist of the correct output for the feature set that is present. The algorithm is made to learn the outputs for the given features in the training data and made to predict the output values for the corresponding features in the test data accordingly. In unsupervised learning, this output is not present and consist of unlabeled feature sets which are then clustered into distinct groups respectively.

In our dataset, the output data of the stock price is present and therefore the methodology of supervised learning is applied. The supervised learning is further subdivided into several methodologies like classification and regression methodologies. The classification scenario is used when the output is to be predicted as a labelled set. In the case of regression, the output is considered to be continuous values. Since the prediction of the stock prices has to be done every day, the regression scenario is considered. Therefore, regression is considered to be performed by using the Support Vector Machine. The important

advantage of SVM is that it allows error within the regression of training data so that the error in the test data is reduced significantly. In addition to SVMs, LSTM has also seen increased usage recently. The major advantage of LSTMs is that it could learn selectively and can remember or forget the required historical data. The stock price data can be highly volatile, therefore, to provide some smoothing effect, the moving average algorithm can be considered along with the SVM and LSTM algorithms.

1.5 Organization of the research

In the upcoming chapters, the two models for predictive modelling and the techniques used has been discussed in the following order.

- Chapter 2 – presents the literature survey
- Chapter 3 – The basics of each model in a stepwise manner as described in the CRISP-DM cycle is explained.
- Chapter 4 - This chapter presents the details of the experiments conducted with all the analysis, interpretation and results. The process of choosing the best model among the methodologies is explained.
- Chapter 5 - This is the last chapter which presents the conclusions based on the study conducted in this project along with recommendations and some prospects for the future work.

Chapter 2 - Literature Survey

2.1 Introduction

This section presents a brief discussion on the existing studies on time series models, machine learning models, deep learning models and a comparative study that has been conducted for stock market prediction and forecasting.

2.2 Summary of base papers

This paper (Xiaotao, Keung, 2016) presents a dynamic model for forecasting the intraday volume percentages by considering two methodologies. In the first one, the average part as the intraday volume pattern is used and in the second one, the residual terms such as abnormal changes are considered. The data is considered from the S&P 500 and an empirical test is done for half-year data. The SVM model is built by taking in the input pattern with two categories to predict the result. One of them is the previous day volume percentages in the same time interval and another in the current recent volume percentage. The results show that the rolling average of previous day volume percentage is helpful in the prediction and this dynamic SVM based prediction outperforms other statistical methods in prediction of the volume and improves tracking performance immensely.

In this paper (Hiransha, Gopal Krishnan,2018), algorithms are classified into linear and non-linear models for forecasting stock prices. The paper uses four different types of architectures namely Multi-Layer Perceptron (MLP), Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM) and Convolutional Neural Network (CNN) for prediction. Two stock market data are taken from the National Stock Exchange (NSE) and the New York Stock Exchange (NYSE) respectively. The network was trained with data from one stock market and tested on another. The results show that the model was able to predict for NYSE

using the NSE data. This was found out to be as a result of these stock markets sharing similar inner dynamics in functioning. Further, it is observed that CNN outperforms other models in the prediction.

The paper (Kale, Khanvilkar,2019) considers the task of predicting the next day's opening value in National Stock Exchange using an Artificial Neural Network (ANN). The neural network is trained using the Error Propagation Training Algorithm (EPTA). The multilayer feedforward network is trained using current day finishing value of input parameters like gold and foreign currency rates. The closing stock value is also incorporated using the Simple Moving Average (SMA) model. A Correlation technique is used to study and analyse the relationship between the closing stock value and the parameters. The results show that the input parameters were useful in predicting the end day stock prices using ANN.

The paper (Abe, Nakayama,2018) uses deep learning techniques to incorporate the prediction of stock returns for one month ahead in the Japanese Stock Market. The data is collected from the Japanese Stock Market monthly. The model is built with considering correlation coefficient and directional accuracy as performance measures over the loss function. The results of this paper imply that deep neural networks outperform shallows and in some cases outperform the representative learning models thereby giving promise as a skilful machine learning method to predict stock returns in cross-section.

A study (Liew, Kyung-Soo,2017) carried out using cutting edge machine learning algorithms are used to predict the returns. The paper showcases the prediction of the direction of the liquid ETFs using the supervised learning classification algorithms. The techniques include Support Vector Machine and deep neural networks. The information sets are divided into past returns, past volume and dummies for days and months or the combination of all three. The gain criterion is seen here to assist in the classifier's performance. The paper also emphasises the importance of cross-sectional and intertemporal volumes as powerful

the information set in prediction. The results show that the model works well only between the horizons of one to three months and should incorporate prior knowledge of markets and intuition on asset class behaviour.

The paper (Madge,2015) uses SVM in predicting the direction of the stock prices. In this study, it is seen that daily closing prices for 34 technology stocks are considered to calculate the price volatility and momentum for each stock and finally for the overall sector. The goal of the SVM model with these parameters is to predict whether the price of the stock will be higher or lower in the future than it is present today. The results from the paper tell us that the predictive ability is superior in the long run when compared to the shorter time frame for prediction.

The paper (Henrique et al.,2018) focuses on researching the predictions for stock prices for large and small scale capitalisations. The paper performs this across three markets differ from each other by comparing prices with both daily and minute by minute frequencies. The base model built by using SVM is compared with the standard random walk approach proposed in the Efficient Market Hypothesis (EMH). The EMH states that it is not quite possible to predict the market movements regularly and concludes that the market movements can be at best be predicted to the accuracy around the range of random walk approach. The prediction errors that are obtained from the model are measured. The results conclude that SVM models have strong predictive powers for the stock prices when updated regularly with solid precision during low volatility periods.

In this paper (Patel et al.,2014), the focus is mainly on predicting the direction of the movement of the stock price and index for Indian stock markets. Four machine learning algorithms namely SVM, ANN, Naive Bayes and Random Forest are used in this paper for prediction. The major emphasis is given on data preprocessing for increasing accuracy. Two inputs are considered for research purposes. They are the stock data of Reliance Industries and Infosys Ltd for 10

years from 2003 to 2012. The research work is done by considering the technical parameters like open, high, low and close prices as trend deterministic data and comparing it with the base data. The results show that when the technical parameters are kept as continuous values, random forest outperforms all the other prediction models in overall performance. It is seen that the overall performance of all the prediction model is improved when the technical parameters are portrayed as trend deterministic data when compared to the base data.

The paper (Gerlein et al.,2016) considers building a machine learning model for a six-year trading period in dollar currency exchange rates. The study is done by performing periodic retraining and the number of attributes and retraining set size are varied regularly for this purpose. The classification models are built by using neural networks and SVM respectively. The model considers the combination of external attributes along with technical indicators as inputs. The results show that this approach has helped in increasing the accuracy of the classification capabilities which directly impacts the final profitability for the model.

The paper (Xiao et al,2013) proposes a study of the three-stage nonlinear ensemble model. In this proposed model, three types of neural network models namely Elman network, generalized regression neural network (GRNN) and wavelet neural network (WNN) are considered. It is constructed using three non-overlapping training sets and then optimised using improved particle swarm optimization (IPSO). The major advantage in this method is the flexibility in the operations for complex non-linear relationships. The results show that the ensemble model significantly improves prediction performance when compared with other individual linear models.

The paper (Nayak et al. ,2015) proposes a hybrid framework of the combination of SVM with K- Nearest Neighbours(KNN) for the Indian Stock Market prediction. The paper considers two indices namely BSE and NSE. The hybrid model makes

use of SVM with a variety of kernel functions to predict profit or loss. The output from the SVM helps to calculate the best nearest neighbour from the given training data to predict future stock prices for the requested period. The performance of this hybrid model is measured using MSE. The results show us that the hybrid model scales well to high dimensional data and the trade-off present between the given classifier complexity and error can be controlled and thereby getting good prediction results.

The paper (Reddy, 2018) compares the working of the SVM model with other linear models for stock price prediction. The goal of the paper is to perform prediction for the large and small capitalizations for the SVM model and compare it with other linear models. The results show us that the SVM model that is built upon well-trained predictor has high efficiency in generating profit when compared to other selected benchmarks.

In the paper (Qian et al. ,2019), the focus is on predicting the stock prices under different stability. The research is based on conducting a stationary analysis of the time series data and then applying LSTM to predict stock prices. The results from the LSTM model is then compared to the standard ARIMA model. The results show that LSTM is insensitive to stability response and has better prediction accuracy than ARIMA model.

In the paper (Wang et al. ,2018), the study is based on the application of LSTM in stock market domain. The research work involves the comparison of a backpropagation neural network model with the LSTM model. The results show us that the LSTM model has better accuracy and this is achieved by the optimization of dropout rates to get better prediction results.

2.3 Research Gaps addressed

From the literature review, it can be inferred that several methods have been developed in recent times for forecasting and predicting the stock prices. Many of the papers have given machine learning methods based results

and few of the research studies have also shown a comparison of these machine learning methods along with time series analysis methods. Almost all the papers considered the official stock price data alone for their research and some of them have considered other external input parameters such as crude oil price and gold price along with these stock price data. The overview of all the papers as part of the literature review is briefed in the below table.

Table 2.1 Literature study overview table

Paper	Data	Model	Comments
Xiaotao, Keung(2016)	S&P500	SVM	SVM outperforms common statistical methods and enhances tracking performance of VWAP strategy.
Hiransha, Gopal Krishnan(2018)	NSE	MLP,RNN,CNN,LSTM	CNN outperformed all others. Didn't consider hybrid networks and used a simple model. Able to predict NYSE using NSE.
Kale, Khanvilkar(2019)	NSE	ANN, SMA	Correlation technique used. Additional Input parameters like gold, forex rates used.
Abe, Nakayama(2018)	Japanese Stock Market	Deep Neural Network, Shallow Neural Network	Deep neural network model outperforms. Uses correlation coefficient as a performance measure.
Liew, Kyung-Soo(2017)	NASDAQ	Support Vector Machine, Deep Neural Networks	Works well between the horizons of 1 to 3 month only
Madge (2018)	NASDAQ	SVM	Prediction for the long term is more efficient than a short term result.
Henrique et al. (2018)	Brazilian, American, Chinese stocks	SVM	Strong predictive power when the model is updated regularly with solid precision during low volatility periods.

Patel et al. (2014)	BSE	SVM, Naive Bayes, ANN, Random Forest	Performance is improved for all prediction models when technical parameters are represented as deterministic data.
Gerlein et al. (2016)	FOREX	Neural Networks, SVM	Combination of external attributes along with input technical indicators as inputs have improved the classification accuracy to increase profitability.
Xiao et al. (2013)	Chinese Stocks	Ensemble model of Elman Network, GRNN, WNN	The prediction performance of the an ensemble model is significantly greater than other individual linear models.
Nayak et al. (2015)	BSE, NSE	Hybrid Model of SVM and kNN	Scales well to high dimensional data. Improves prediction capability.
Reddy(2018)	BSE	SVM	Model built by using SVM generates a higher profit when compared to other linear models in the selected benchmarks.
Qian et al. (2019)	Chinese Stocks	ARIMA, LSTM	LSTM is insensitive to stability responses and performs better than ARIMA giving higher prediction accuracy.
Wang et al. (2018)	Chinese Stocks	LSTM	LSTM provides higher accuracy than backpropagation neural networks through effective optimization of dropout rates.

After analysing every research study based on the above Table 2.1, it can be seen that predominant of the works have been done using SVM for stock prediction. It can also be that only a few papers have considered using external

input parameters to check whether it improves the prediction model. It is also seen that only a handful of research study compared the performance of SVM and LSTM models for stock prediction. It is also noted that only a couple of papers used moving averages to smooth the data before feeding into the model. Therefore, the research work is designed by addressing these gaps in the previous studies. The major focus of this research is to compare the performance of LSTM model and SVM model. The SVM models are built based on following the design decisions taken in papers like Xiaotao, Keung(2016), Madge (2018) and Henrique et al. (2018). The LSTM models are built based on following the design decisions taken in papers like Abe, Nakayama(2018) and Wang et al. (2018). The original data for stock prediction is considered as DJI in S & P 500 based on the paper Xiaotao, Keung(2016) and external input parameters like crude oil prices and gold prices are added to this data based on the papers Kale, Khanvilkar(2019) and Gerlein et al. (2016). The moving averages are applied to these datasets based on the paper Kale, Khanvilkar(2019). The performance of SVM models and LSTM models is evaluated on the original data and the original data with external input parameters and moving averages based on the metrics used in the papers like Henrique et al.(2018),Patel et al.(2014),Gerlein et al.(2016) and Xiao et al. (2013). The performances are then compared to understand whether the additional external input parameters and moving averages provide enhancement to the performance of the base SVM and LSTM models on the original data and improves it overall.

Chapter 3 - Methods and Models

In this chapter, the methodologies and models developed are discussed in detail. SVM and LSTM algorithms are used to build models as discussed in the previous chapter. Brief introductions are given to both of them and the 8 models are explained. The cross-industry standard process for data mining (CRISP-DM) methodology is used to structure this research. The steps involved in CRISP-DM are explained and an architecture diagram is provided to provide a graphical representation of the concepts involved in the research.

3.1 SVM Model

3.1.1 Introduction

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. SVMs are considered to be a set of related supervised learning methods used for classification and regression. In SVM, support vectors are the individual observation co-ordinates and support vector machine is a border where hyper-plane and line are best separated (Gandhi 2018). This can be seen clearly in figure 3.1. These support vectors are instrumental in determining the position of the hyperplane. Therefore, adding or removing support vectors would modify the position and orientation of hyperplanes**.**

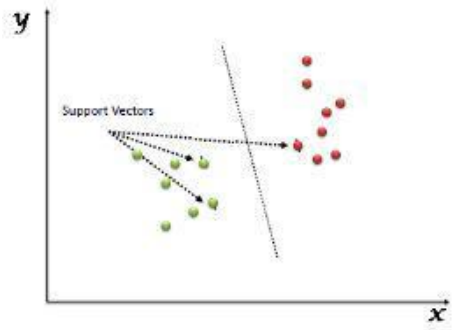

Figure 3.1 SVM Data points. Showing the data points represented in a 2D space and the support vector (Gandhi,2018).

3.1.2 Process Involved

To understand the SVM process, some of the common terminologies should be known (Bhattacharrya,2018). They are,

Kernel: It is used to map low dimensional data into a high dimensional data.

Hyper Plane: It is the separation line between the data classes. In SVR, it is the line that will help in predicting the continuous value or target value.

Boundary line: In SVR, there are two lines other than Hyper Plane that creates a margin. The support vectors can be present on the boundary or outside it. This boundary line separates the two classes present.

Support vectors: These are data points that are closest to the boundary. The distance of the points is considered to be minimum.

3.1.3 SVM Process Flow Diagram

The process flow diagram for the Support Vector Machine for the regression process is as follows,

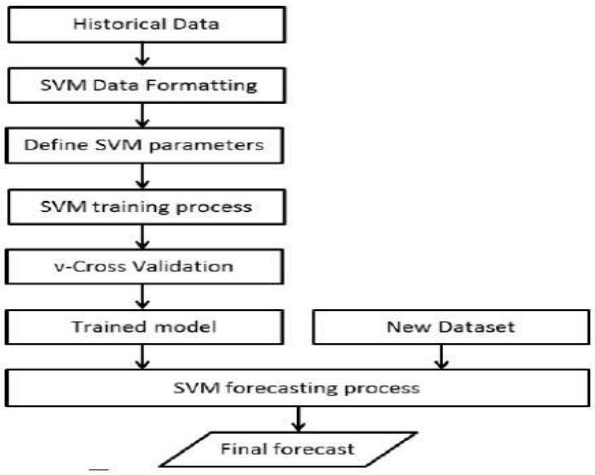

Figure 3.2 Process Flow of SVR. This shows the basic flow diagram of the support vector regression (Cipcigan,2013).

Data is provided as input to the model. These inputs should be having an appropriate structure to be read properly. In the next step, SVM parameters and kernel function are defined. Error cost term C and maximal margin ε are to be selected suitably. Once this is done, the training process begins. Data is subdivided into v parts. One subset is used for validation and rest are used for training the model. In this way, overfitting is avoided and good generalisation of the performance is achieved (Cipcigan, 2013). Once the model is created, test data is provided to the model and SVM produces a prediction output for this test data based on the trained model.

3.2 LSTM Model

3.2.1 Introduction

Traditional neural networks have a disadvantage of taking into consideration the impact of past events when making predictions for future outcomes. This is rectified by Recurrent Neural Networks (RNN). RNNs are networks with loops that allow the information to be passed from one stage to another in the whole network. RNNs take their input not only from the current sample but also from the past. The decision of the recurrent networks at time t-1 affects the decision which reaches one moment later at time step t. Such networks have two types of inputs, the present and the recent past, which combine to determine the future results (Olah,2015).

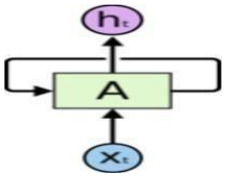

Figure 3.3 Basic Recurrent Neural Network. The basic flow of RNN with $X_t$as input and a feedback loop is shown (Olah,2015).

LSTM units were proposed by the German researcher Schmidhuber in the 1990s

to solve the problem of vanishing gradient. LSTMs help in preserving the error that can be backpropagation through time and layers. The error rate is maintained constant over time so that the network continues to learn over multiple time steps, thus providing an opportunity to link the causes and effects remotely.

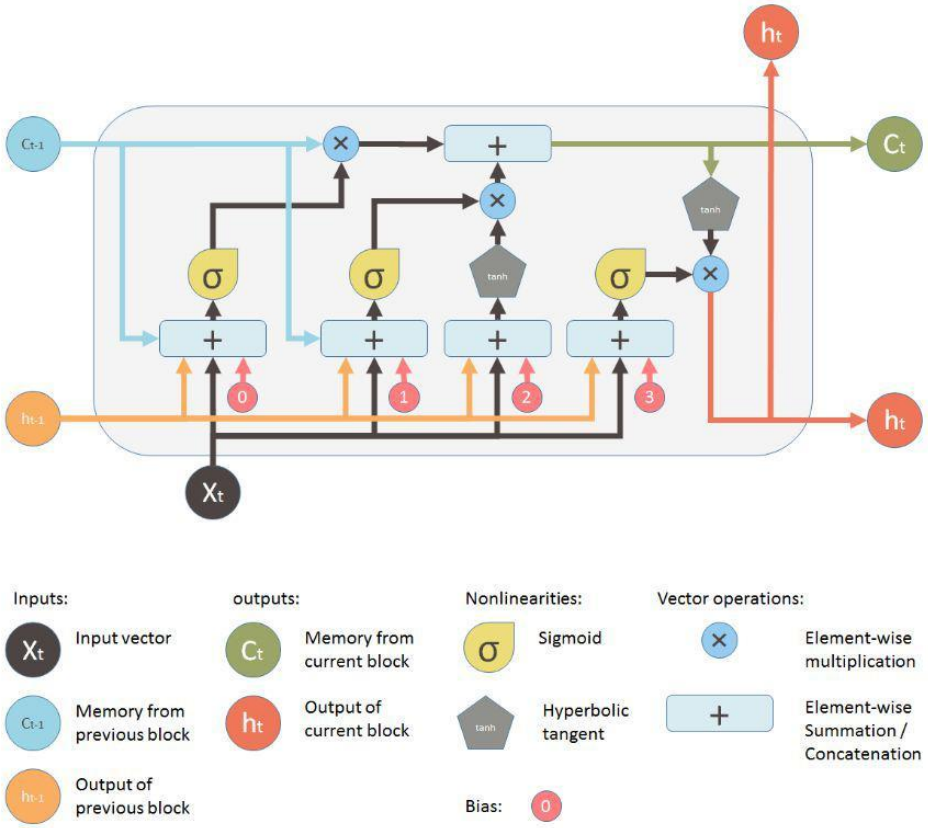

Figure 3.4 LSTM cell diagram. The gated cell diagram of an LSTM neural network is shown with forget and remember gate. At each stage the forget gate decides what information to be passed to the output gate from the input gate (Yan, 2016).

LSTMs contain information in a gated cell which is the key idea of these networks. The LSTMs can add or delete information to the cell through the gates. These gates are composed of a sigmoid neural network layer and a pointwise multiplicative operator. Overall, an LSTM has three of these gates to control and protect the cell state information.

3.2.2 Process Involved

Keras is used in the process of building an LSTM network. Keras is high-level neural network APIs written in Python. It can run on top of TensorFlow library with a focus on the faster experimentation. The major objective of Keras is to build a model that can organize the layers in the network (Thomas,2018). The below architecture shows the Keras LSTM network flow.

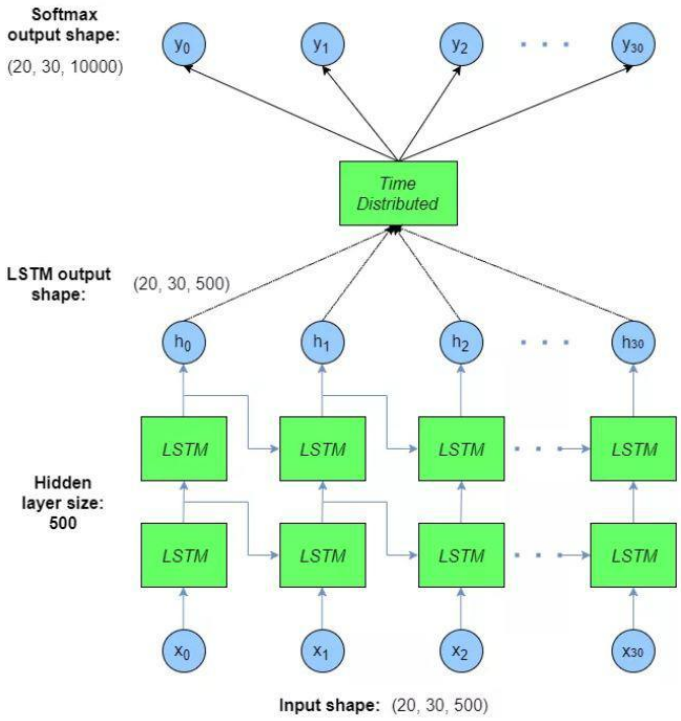

Figure 3.5 Keras LSTM architecture. The input function with several hidden layers, batch size and time steps are defined and inputted into hidden layers. The SoftMax function is applied to this output to determine the final output (Thomas,2018).

The input shape is the input to the model which has the input parameters ordered as batch size, hidden layers and number of time steps. The input data is fed into the stacked LSTM cell layers. The output layer has a SoftMax activation function which calculates the probabilities of each target class over all possible target classes. The output of the network is then compared with the training data and the respective error and gradient backpropagation are performed (Thomas,2018). The stepwise process is defined as follows:

Step 1: Preparing the Data

LSTMs have the in-built capability to account for any variations or discrepancies in the data. The first step is data preparation which involves converting the raw input data into the right shape required for inputting into the Keras LSTM model. LSTMs are very sensitive to the scale of the input data, especially when used with sigmoid or tanh activation function. Hence it is essential to normalize the data and then use it in the model. The input data is then split into training and test data (Ahmed et al., 2010).

Step 2: Defining the input parameters

The input parameters like the number of time steps, epoch size, batch size and hidden layers are defined before building the model (Ahmed et al., 2010).

Step 3: Defining the model and compilation

In this step, the model is organised into layers. Sequential model is the most frequently used model type which builds a linear stack of network layers. Once the model is defined and structured with the parameters defined in step 2, the compilation of the model is built using the Keras functions. This compiled model is then fit on the training data to make the model learn the data patterns (Ahmed et al., 2010).

Step 4: Model Training and Evaluation

The model is fitted on the training data and evaluated against both the training and test data. Finally, the predicted value is compared with the actual output data to estimate the model's accuracy and performance. Evaluation metrics like RMSE values and accuracy are used to evaluate the performance of the model.

3.2.3 LSTM Process Flow Diagram

The diagram below shows the basic architecture of the LSTM network and explains the data flow through the memory cells in the LSTM networks.

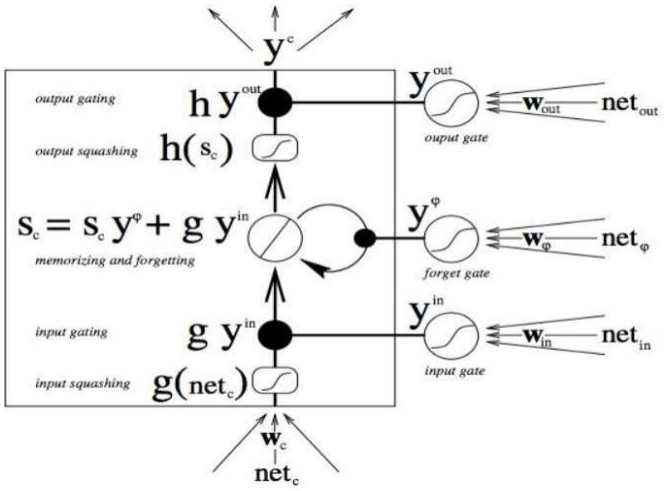

Figure 3.6 LSTM data flow structure. The information flow cycle of the overall LSTM network from input flow till output extraction is shown. Input information along with gated cell for forgetting and memorizing the data is shown in detail (Skymind,2017).

In the above diagram, the arrows present at the bottom show where information flows into the cell at various points. The present input and the previous state are fed into the cell and each of the three gates. The small circles determine whether to allow the new input, delete the present cell state and/or allow the state that impacts the network's output at the given time step. S_c represents the current state of the cell and gy^{in} represents the current input in that state. The large bold letters represent the result of each operation (Skymind,2017).

3.3 Moving Average

3.3.1 Introduction

Moving average is defined as the calculation to analyse data points by creating a series of averages of various subsets of the entire dataset. It is regularly used with time series data to smooth out fluctuations and provide insights on long term trends. The threshold between short-term and long-term depends on the given task and the parameters of moving average are set based on it (Hayes,2019).

3.3.2 Process Involved

In the finance sector, a simple moving average (SMA) is defined as the unweighted mean of previous n data. For example, a simple equally weighted running mean for an n-day sample of the closing price is the mean of the previous n days' closing prices (Hayes,2019). When there is a series of numbers and a fixed 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 excluding the first number of the series and including the next value in the subset. This process is repeated until the final value in the subset is reached (Hayes,2019).

3.4 Modelling and Evaluation – CRISP-DM Cycle

This cross-industry standard process for data mining is the data mining process model that describes the stages or approaches involved in solving any data analytics problem. It involves a list of processes required to be performed in a stepwise manner for approaching any data mining project.

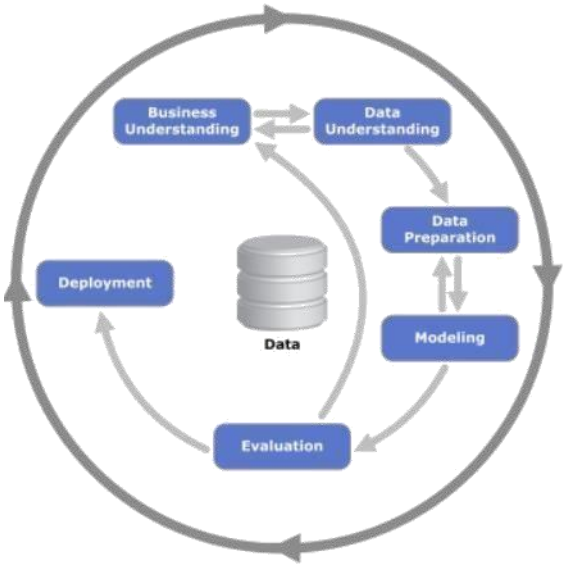

Figure 3.7 CRISP-DM architecture. The data mining process with stage-wise implementation flow (Vorhies, 2016).

The CRISP-DM architecture of this research is shown below

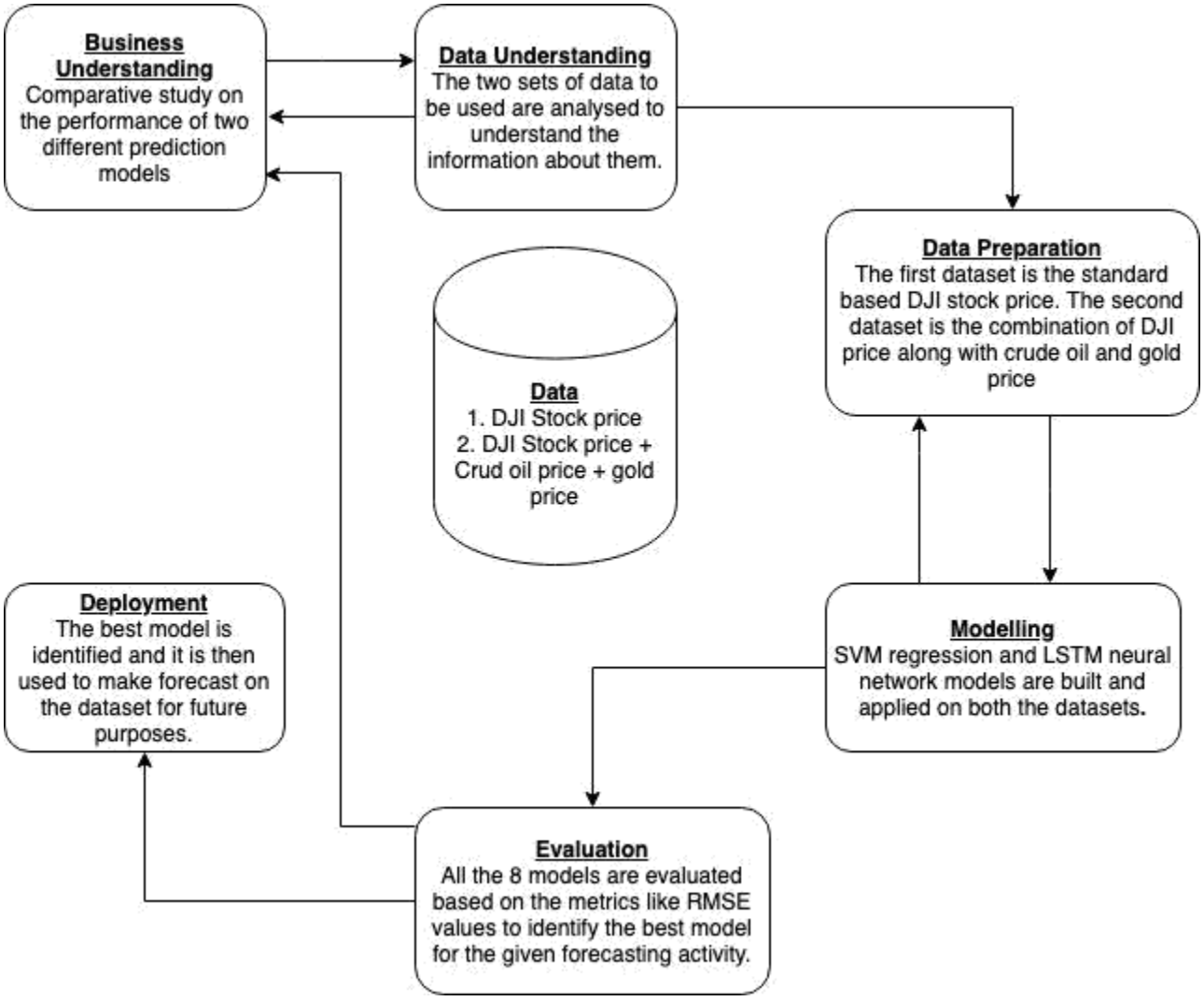

Figure 3.8 CRISP-DM architecture-Comparative Study. This architecture gives the workflow of the research study in terms of the CRISP-DM process.

The six major steps involved as defined above in the perspective of our data analytics problem is explained below (Vorhies, 2016).

Stage 1: Business Understanding

The overall goal of this research is to build eight different types of prediction models and conduct a comparative study on the performance of the models developed to identify which among them will give better forecasting on stock price prediction for DJI stock. The eight models used in this study are,

1. SVM Base Model on DJI stock data
2. SVM Base Model with moving averages on DJI stock data
3. SVM Advanced Model on combined data of DJI stock, crude oil and gold prices
4. SVM Advanced Model with moving averages on combined data of DJI stock, crude oil and gold prices
5. LSTM Base Model on DJI stock data
6. LSTM Base Model with moving averages on DJI stock data
7. LSTM Advanced Model on combined data of DJI stock, crude oil and gold prices
8. LSTM Advanced Model with moving averages on combined data of DJI stock, crude oil and gold prices

Stage 2: Data Understanding

For this research study, 3 datasets are considered. The base dataset is the DJI stock dataset from 2015 to 2018. It contains five values detailing movements in the price over each day. The values are as follows,

1. Open: Starting price for the stock in a given trading day
2. Close: Final price for the stock on that day
3. High: Highest price at which the stock traded on that day
4. Low: Lowest price at which the stock traded on that day
5. Volume: Total number of shares traded in the market on that day

The other two datasets are crude oil and gold price dataset. They are also considered from 2015 to 2018 to maintain the uniformity. Both of them contain the closing day prices respectively.

Stage 3: Data Preparation

Data preparation is done according to the models that are to be developed. Before this, data handling issues are identified and tackled. The major issues are handling of empty values and combining of several datasets properly without any duplications. When there are null values in the dataset, it is removed because it is not wise to consider the mean or median for missing values of prices. This is because the prices are highly volatile and therefore it can lead to misleading conclusions when the null values are replaced by the mean or median of the prices. Similarly, while combining the stock price, gold price and crude oil price dataset, there are issues with joining them correctly. This is because for some of the days when the stock market was opened, either the gold exchange rate or crude oil exchange rate markets were closed or vice versa. Therefore, it is decided to perform an inner join with the date as primary keys while combining these three datasets. In this way, the combined data is obtained properly without any mismatching of data while joining. For the SVM and LSTM base models, the base DJI stock price dataset is used in the original format. For the SVM and LSTM advanced models, the combined dataset of the DJI stock price and crude oil and gold price dataset is used in the original format. For the next 4 models, a substantial amount of data preparation is carried out. The concept of feature engineering is used for this purpose. Feature engineering can be said as the process of generating new features based on existing features in the given domain to improve the performance of the algorithm used. Moving averages are used in generating new features based on the available ones. For the SVM base

model with moving averages and LSTM base model with moving averages, the following features are generated (Liu,2017),

$AvgPrice_5$	The average close price over the past five days
$AvgPrice_{30}$	The average close price over the past month
$AvgPrice_{365}$	The average close price over the past year
$\frac{AvgPrice_5}{AvgPrice_{30}}$	The ratio between the average price over the past week and that over the past month
$\frac{AvgPrice_5}{AvgPrice_{365}}$	The ratio between the average price over the past week and that over the past year
$\frac{AvgPrice_{30}}{AvgPrice_{365}}$	The ratio between the average price over the past month and that over the past year
$AvgVolume_5$	The average volume over the past five days
$AvgVolume_{30}$	The average volume over the past month
$AvgVolume_{365}$	The average volume over the past year
$\frac{AvgVolume_5}{AvgVolume_{30}}$	The ratio between the average volume over the past week and that over the past month
$\frac{AvgVolume_5}{AvgVolume_{365}}$	The ratio between the average volume over the past week and that over the past year
$\frac{AvgVolume_{30}}{AvgVolume_{365}}$	The ratio between the average volume over the past month and that over the past year
$StdPrice_5$	The standard deviation of the close prices over the past five days
$StdPrice_{30}$	The standard deviation of the close prices over the past month
$StdPrice_{365}$	The standard deviation of the close prices over the past year

$\frac{\text{StdPrice}_5}{\text{StdPrice}_{30}}$	The ratio between the standard deviation of the prices over the past week and that over the past month
$\frac{\text{StdPrice}_5}{\text{StdPrice}_{365}}$	The ratio between the standard deviation of the prices over the past week and that over the past year
$\frac{\text{StdPrice}_{30}}{\text{StdPrice}_{365}}$	The ratio between the standard deviation of the prices over the past month and that over the past year
StdVolume_5	The standard deviation of the volumes over the past five days
StdVolume_{30}	The standard deviation of the volumes over the past month
StdVolume_{365}	The standard deviation of the volumes over the past year
$\frac{\text{StdVolume}_5}{\text{StdVolume}_{30}}$	The ratio between the standard deviation of the volumes over the past week and that over the past month
$\frac{\text{StdVolume}_5}{\text{StdVolume}_{365}}$	The ratio between the standard deviation of the volumes over the past week and that over the past year
$\frac{\text{StdVolume}_{30}}{\text{StdVolume}_{365}}$	The ratio between the standard deviation of the volumes over the past month and that over the past year
$\text{return}_{i:i-1}$	Daily return of the past day
$\text{return}_{i:i-5}$	Weekly return of the past week
$\text{return}_{i:i-30}$	Monthly return of the past month
$\text{return}_{i:i-365}$	Yearly return of the past year
MovingAvg_{i_5}	Moving average of the daily returns over the past week
MovingAvg_{i_30}	Moving average of the daily returns over the past month
MovingAvg_{i_365}	Moving average of the daily returns over the past year

Figure 3.9 Feature generation. This gives a basic understanding of the various features that can be generated based on the stock price data (Liu, 2017).

For the SVM advance model with moving averages and LSTM advance model with moving averages, the following features are generated in extra when compared with the base model with moving averages. The datasets are highly volatile and have a huge range. Therefore they are made to undergo normalisation using the min-max scaler here. Once the predictions are done, the inverse scaler is used to bring back the original stock prices (Brownlee,2016).

Stage 4: Modeling

Model 1- SVM Base Model

In this model, the input data is considered as the DJI stock price data. The input data is then split into training data and test data. This is done as 75% and 25% respectively. Scalar transformation is then done for training data and required input parameters are given to the SVM base model and trained. The model is then fitted based on these values and stored for future use. The fitted model is then used to predict the test data and finally, the predicted test data is then compared with the actual test data. The evaluation measures like MSE, RMSE, MAE, MAPE and R2 scores are obtained based on this. Finally, a visual output is given to display the predicted data vs expected data to understand the performance of the model.

Model 2- SVM Base Model with moving averages

In this model, change is performed in the DJI stock price data by applying moving averages to the input parameters present. Rest of the steps are similar to the first SVM model and evaluation measures are carried out on the results.

Model 3- SVM Advanced Model

In SVM advanced model, the input data is taken as the combination of DJI stock price, crude oil price and gold price data. The combined data is split similar to 75 % training and 25% test data. The SVM model is fitted based on the training data with the necessary input parameters.

Model 4- SVM Advanced Model with moving averages

In this model, moving averages are applied to the input parameters present in the

combined data used in SVM advanced model. Rest of the steps are similar to the previous model and results are produced based on it.

Model 5- LSTM Base Model

The input data is considered as the DJI stock price data. The input data is then split into training data 75% and test data 25%. The data is then normalised by using a min-max scaler to fit the scaling of the model. Keras model tensors are created and the input and output dataset is tuned accordingly. The activation function is used to normalise the input layers. It helps in enhancing the process of learning. Batch normalization is done by LSTM to reduce the amount of shifting around of hidden unit values. This is called a covariance shift (Ioffe et al. ,2015). Keras model is capable of performing this normalisation as part of model building using its in-built function(Keras Documentation, 2015). Sequential model in Keras is built with all the required parameters and compiled for further analysis.

The compiled model is fitted on the input data for the defined number of epochs and the predictions on the test data are made. The predicted output is evaluated in terms of MSE, RMSE, MAE, MAPE and R2 values. A graphical plot of the predicted outcome vs the expected outcome gives us a visual understanding of the model's performance.

Model 6- LSTM Base Model with moving averages

In this model, moving averages are applied to the data used in LSTM base model. Rest of the steps are identical to the base model and results are obtained based on it.

Model 7- LSTM Advanced Model

In the LSTM advanced model, input data is combined data of DJI stock price, crude oil price and gold price. Data is normalised using min-maxæ scaler similar to the base model approach and then split to 75 % training data and 25 % testing data. LSTM model is then fitted with training data and used to predict test data based on it. Evaluation measures are performed on predicted output and results are stored. The graphical plot is used to show the difference between the predicted data and actual test data present.

Model 8- LSTM Advanced Model with moving averages

In the final model, moving averages are applied to combined data and then similar steps are followed like in LSTM advanced model.

Stage 5: Evaluation

The evaluation metrics that are suitable for regression are identified first. It is found out that evaluation metrics like MSE, RMSE, MAE, MAPE and R2 scores are suitable for checking the performance of the models (Swalin,2018).

Mean squared error (MSE) is used to measure the average of the squares of the errors. It is calculated as the average squared difference between the predicted value and the actual value. It is mathematically represented as,

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2$$

Where,

n - total number of observations

Pi - Predicted outcome

Oi - Expected outcome

Root mean squared error (RMSE) is used to represent the sample standard deviation of differences between predicted and actual values. It can be represented by a mathematical formula as,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}$$

Where,

n - total number of observations

Pi - Predicted outcome

Oi - Expected outcome

Mean absolute error (MAE) is the average of the absolute differences between actual and predicted values. It is a linear score and therefore individual differences are weighted evenly in the averages. It is mathematically represented as,

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i|$$

Where,

n - total number of observations

Pi - Predicted outcome

Oi - Expected outcome

In all these three-evaluation metrics MSE, RMSE and MAE, lower values indicated better performance of the model.

R squared (R2) is used for studying how well the selected independent variables explain the variability in the dependent variable. When MSE is high, R2 becomes small and it means that the performance of the model is poor. It can be mathematically represented as,

R2= $1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$

Where,

Numerator part is MSE

Denominator part is variance in Y values

n - total observations

Mean absolute percentage error (MAPE) is calculated as the average of percentage errors. It can be mathematically represented as,

MAPE=$\left(\frac{1}{n}\sum\frac{|Actual - Forecast|}{|Actual|}\right) \times 100$

In this stage the machine learning models built in the above steps are then evaluated based on the metrics like RMSE, R2 and MAPE values and the best one is chosen (Swalin,2018). The RMSE value is obtained from the MSE value and the MAPE is obtained from the MAE value. In the case of RMSE and MAPE, lower the value better the performance of the model. In the case of R2, models whose values are closer to 1 have better performances.

Stage 6: Deployment

This stage is the final deployment of the code into a system to work on unknown or new emerging sets of data. This is not within the scope of this project as it is limited only to the prediction of stock prices and analysing the best model based on results. But, this work can be extended to different sets of data from various other stock markets for stock prediction. It can also be tested to see its prediction activity for disease forecasting and weather forecasting based on the various

analysis done in the models. The experimental analysis and results of each of the models are discussed in detail in Chapter 4.

3.5 Tools and Techniques Used

Python is chosen as the programming language for this project. This is because of the following reasons (Rayome,2018),

1. Python has huge community support behind it. Most of the issues can be solved by looking at StackOverflow solutions. This is because of the emergence of python as one of the most popular languages on the site. Therefore, this makes the work very easy to find direct answers for programming queries.

2. Python contains several useful tools for data science. Packages like Numpy, SciPy, Pandas, Scikit-learn and Keras etc are readily available for free and are also well documented. These python packages help in reducing the complexity of the code, thereby helping in improving the code readability and reusability.

3. Python is considered to be a flexible language and allows for programs that are similar to pseudo-code. It is beneficial in implementing and testing the pseudo-codes given in academic papers.

There are also some disadvantages to python. It is a language which is typed dynamically and packages are prone to be duck typed. This means that some of the methods return a dictionary like output instead of being an actual dictionary. Therefore a lot of trials and error testing has to be done to confirm the return type of the method when it is not mentioned properly in the document. This

can cause inconvenience while trying to learn and implement new packages in python. Now, the packages that are mainly used in this research are briefly explained as follows,

Numpy

Numpy is a python package that provides high level mathematical and scientific abstractions wrapped in python. Most of the early programming languages didn't provide support for mathematical abstractions because it affected the semantics and syntaxes of the code. This issue is solved in Python by using numpy to use such functions. Numpy is also used for providing basic numerical routines like providing tools for finding eigenvectors (Klein,2018).

Scikit-learn

Scikit-learn is a machine learning package available for python programming language. It contains a vast variety of classification, regression and clustering algorithms like k-means, linear regression, logistic regression and support vector machine etc. It is mainly used to interoperate with other numerical and scientific python packages like numpy and scipy. Most of the package is written in python with few core algorithms written in Cython for better performance. (Nnamdi,2019).

Keras

Keras is a python package which can run on top of TensorFlow, Microsoft cognitive toolkit (CNTK) and Theano. It is developed to enable fast experimentation. Keras supports modularity and extensibility and provides fast prototyping. It supports both convolution neural networks and recurrent neural networks and also the combination of them. It contains various implementations of regularly used neural network building blocks like activation functions, optimizers, dropout rate and layers etc to ensure that working is easy with various image and text data. Keras package code is also available on GitHub and has various community support forums including a slack channel,

gitter channel and GitHub issues page (Nnamdi,2019).

3.6 Overall Project Architecture

The architecture of the overall project given below helps us understand the workflow of the research study.

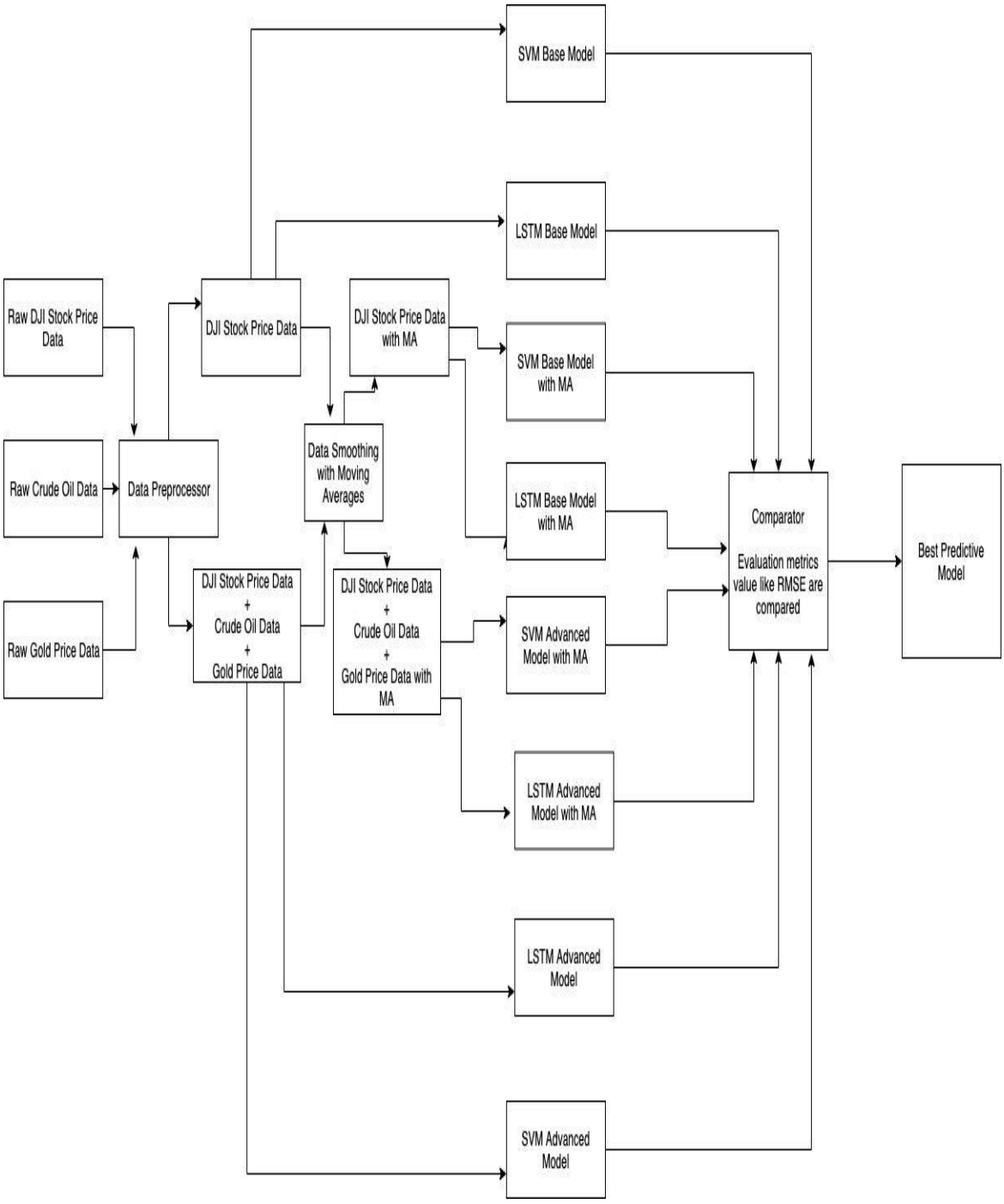

Figure 3.10 Project Architecture. This architecture gives the understanding of the entire research work in a stepwise manner at each stage.

The raw datasets of DJI stock price, crude oil price and gold price are obtained by web scraping. The raw data is then preprocessed and maintained as DJI stock price data separately and combined data of DJI stock price, crude oil price and gold price. Moving averages are applied to both these and those data are also stored. Now, 4 SVM based models and 4 LSTM based models are built using the available data. The results are then compared by using the evaluation metrics like RMSE, MAPE and R2 score. Finally, the best model is chosen based on these results.

Chapter 4 - Experimental Analysis and Results

In this section, data analysis is done on stock price, crude oil price and gold price data to gain deeper insight. Then, experimental analysis is done on 4 SVM models and 4 LSTM models respectively. Finally, the results of each model are interpreted and compared to choose the best result. A t-test is done to confirm that the selected best model provides results which are statistically significant from the rest of the models.

4.1 Data Analysis

4.1.1 DJI stock price

The DJI stock price data is obtained from yahoo finance through means of web scraping (Liu, 2017). This is done by providing the index as DJI to search the stock price data between the years 2015 to 2018. Once it is obtained, descriptive statistics is done on the data as follows,

Table 4.1 Descriptive statistics of DJI stock price

Values	Open	High	Low	Volume	Close
count	1,249.00	1,249.00	1,249.00	1,249.00	1,249.00
mean	19,812.14	19,902.27	19,714.45	198,226,933.55	19,813.17
std	3,260.32	3,275.08	3,242.70	134,382,321.23	3,256.38
min	15,372.93	15,478.21	15,340.69	40,350,000.00	15,372.80
25%	17,277.11	17,409.72	17,163.73	88,460,000.00	17,280.83
50%	18,135.72	18,203.37	18,062.49	123,870,000.00	18,128.66
75%	22,349.70	22,405.63	22,299.58	299,670,000.00	22,381.20
max	26,833.47	26,951.81	26,789.08	900,510,000.00	26,828.39

Now, a histogram is plotted to know at what range of closing values were the stock at the considered time interval. It is seen that for most of the days, the closing day stock price was between 17000 to 19000.

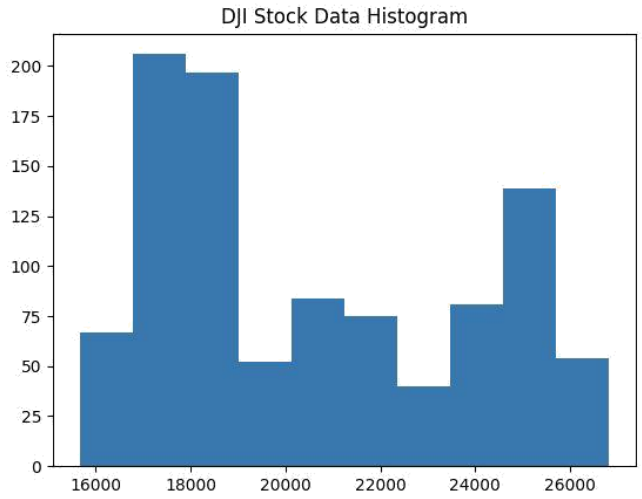

Figure 4.1 Histogram of DJI stock price

Next, a line plot is built to understand the pattern of the closing stock prices per day. It is seen that there has been a huge surge in stock prices for the last quarter of data.

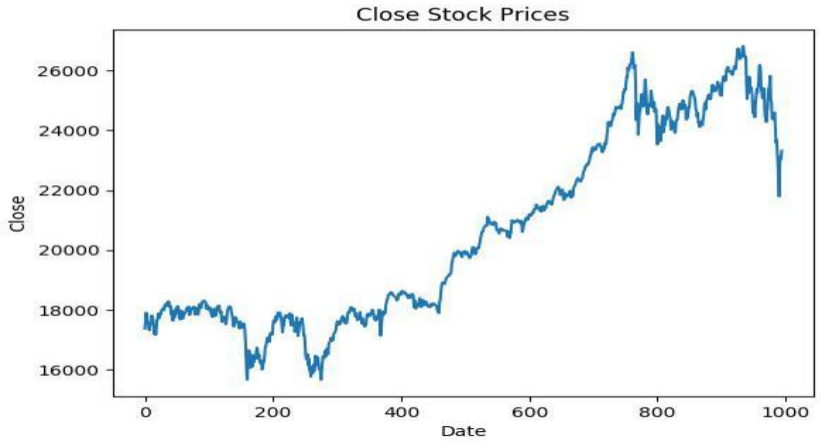

Figure 4.2 DJI closing stock price

Similarly, a line plot is built for stock volumes and it is also found to be following a similar pattern.

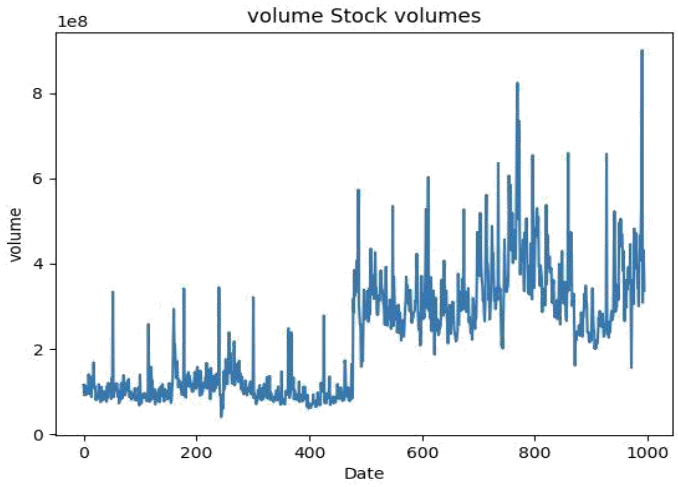

Figure 4.3 Volumes of DJI stock per day

Daily returns are the difference in opening and closing prices per day. When it is positive, it means that closing prices were greater than opening prices and when it is negative, it means that closing prices were lesser than opening prices. From the daily returns plot, it can be seen that there are huge changes in daily returns in the final quarter.

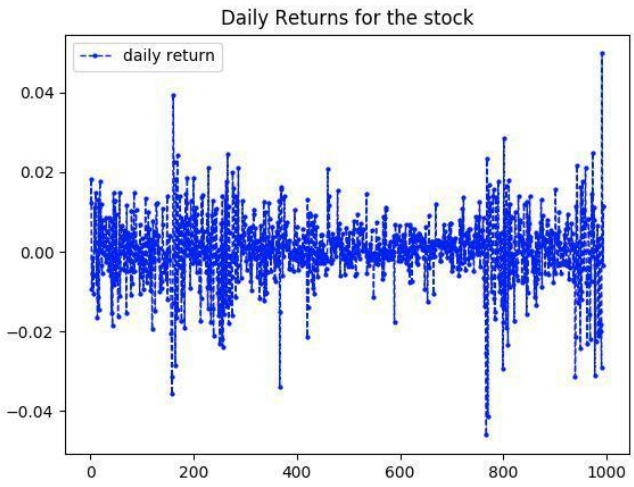

Figure 4.4 Daily returns of DJI stock

Next, moving averages are applied to the stock price data and its descriptive statistics is as follows,

Table 4.2 Descriptive statistics of combined data with moving averages

value	Average price of stock with moving averages for 5 days	Average price of stock with moving averages for 30 days	Average price of stock with moving averages for 365 days
count	995.00	995.00	995.00
mean	20,570.29	20,523.56	19,567.85
std	3,205.00	3,191.21	2,691.42
min	15,918.04	16,077.50	16,787.31
0.25	17,809.35	17,792.05	17,448.52
0.50	19,868.14	19,745.27	17,921.01
0.75	24,128.88	24,230.06	21,731.53
max	26,667.87	26,427.10	25,122.93

value	Average volumes of stock with moving averages for 5 days	Average volumes of stock with moving averages for 30 days	Average volumes of stock with moving averages for 365 days
count	995.00	995.00	995.00
mean	224,326,514.57	221,432,329.74	193,470,644.37
std	127,185,323.22	120,205,155.87	104,097,999.53
min	63,654,000.00	72,096,666.67	91,874,404.76
0.25	102,038,000.00	103,197,619.05	106,312,162.70
0.50	229,224,000.00	244,658,095.24	118,247,103.17
0.75	323,445,000.00	321,404,285.71	311,841,964.29
max	687,108,000.00	526,674,761.90	358,670,714.29

Next, the effect of moving averages is seen on both closing stock prices and volume. It is seen that it reduces the volatility in the values and brings about a

smoothness in the curve.

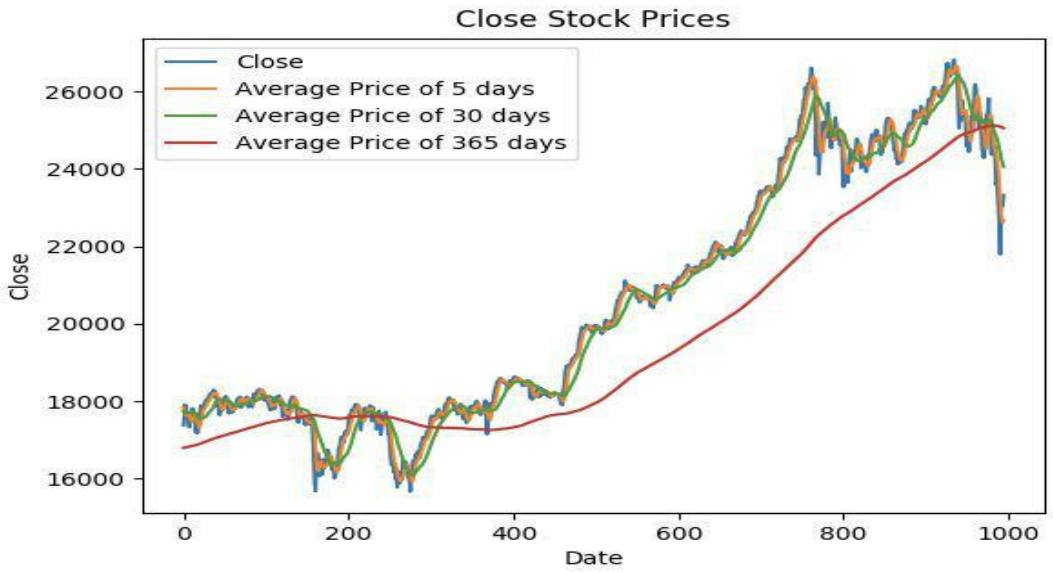

Figure 4.5 Closing stock prices of DJI per day with moving averages applied.

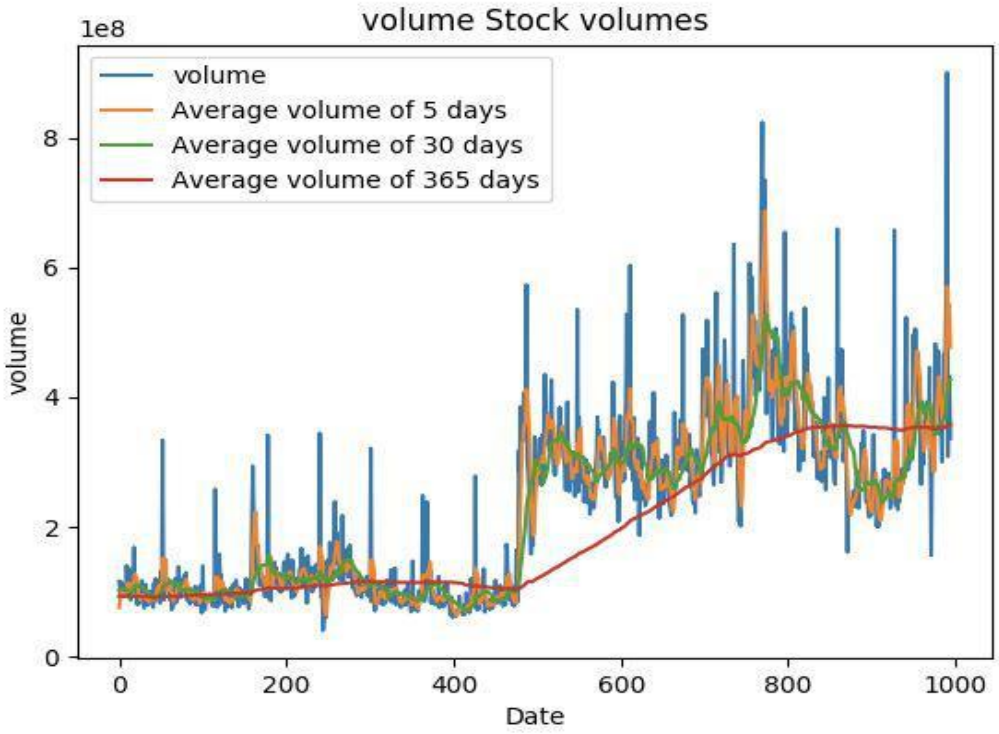

Figure 4.6 Volumes of DJI stocks per day with moving averages applied

4.1.2 Gold price

Next, Gold prices are also obtained using web scraping from the goldprice.org website. Moving averages are then applied to these gold prices. Descriptive statistics for gold prices are as follows,

Table 4.3 Descriptive statistics of gold price

value	gold
count	1,249.00
mean	1,240.20
std	68.68
min	1,049.40
25%	1,199.00
50%	1,246.40
75%	1,291.50
max	1,385.00

value	Average gold price with moving averages for 5 days	Average gold price with moving averages for 30 days	Average gold price with moving averages for 365 days
count	995.00	995.00	995.00
mean	1,233.67	1,233.23	1,234.25
std	70.63	68.65	44.71
min	1,059.11	1,067.73	1,144.98
0.25	1,193.96	1,195.62	1,197.34
0.50	1,240.76	1,239.11	1,246.55
0.75	1,283.40	1,279.78	1,265.43
max	1,357.01	1,345.20	1,297.10

Similar to the stock price, line plots are built to understand if there is any pattern in the gold prices. It is seen that gold prices fluctuated throughout the period unevenly and therefore no pattern can be seen. Moving averages are then applied to the gold price which then sees a smoothing effect on the rough curves in the plots.

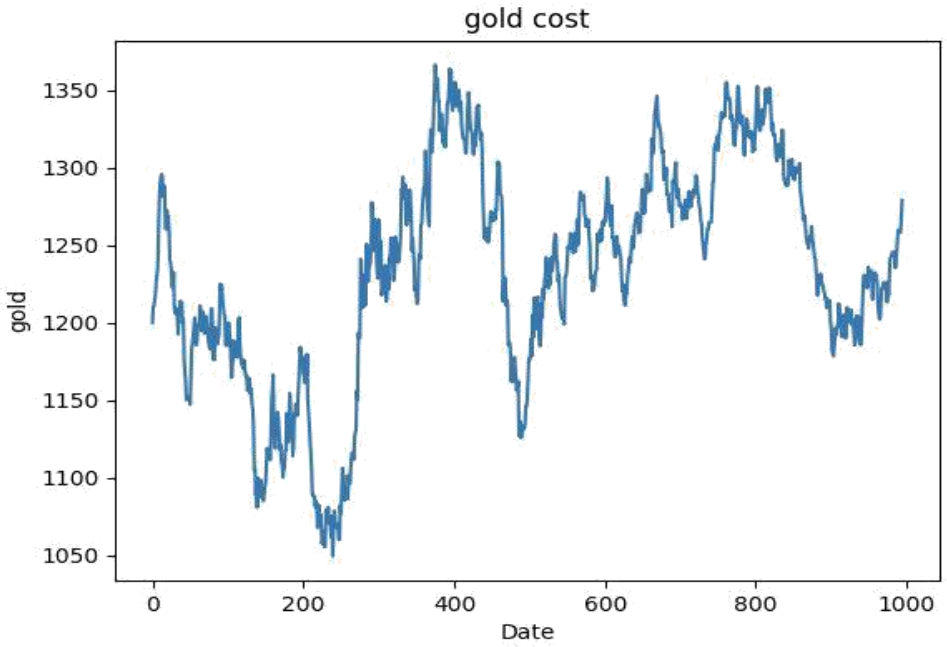

Figure 4.7 Daily gold prices

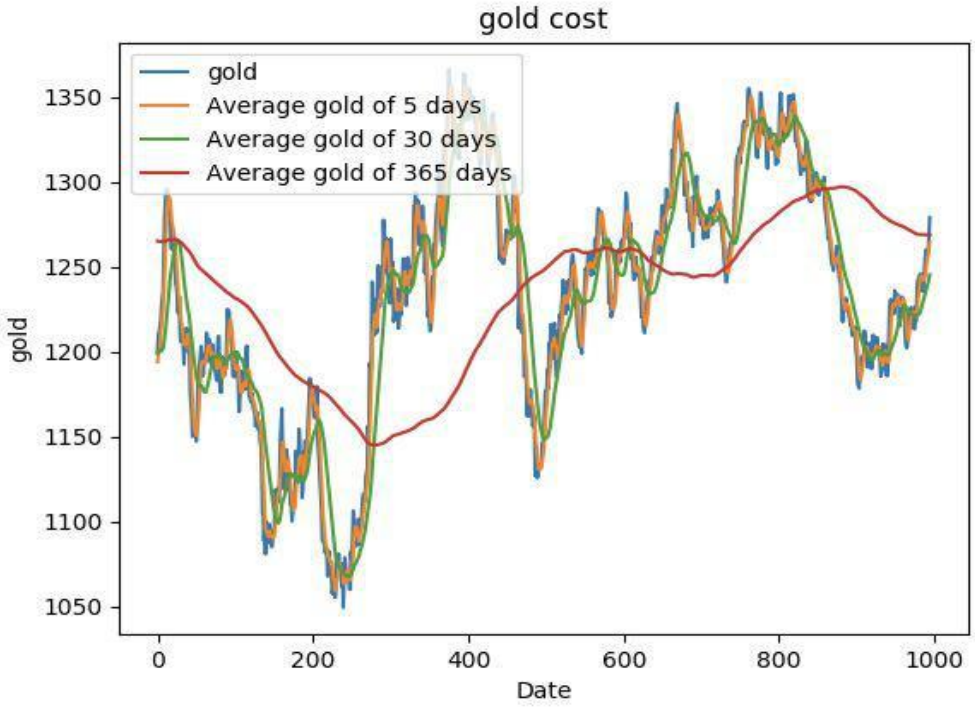

Figure 4.8 Daily gold prices with moving averages applied

4.1.3 Crude oil price

Crude oil prices are also obtained by web scraping from the United States of America's energy information website. Descriptive statistics of crude oil prices are as follows,

Table 4.4 Descriptive statistics of crude oil price

	crude oil
count	1,249.00
mean	60.27
std	19.65
min	26.21
25%	46.77
50%	52.81
75%	68.47
max	107.26

value	Average crude oil price with moving averages for 5 days	Average crude oil price with moving averages for 30 days	Average crude oil price with moving averages for 365 days
count	995.00	995.00	995.00
mean	52.04	52.10	54.90
std	10.02	9.84	12.23
min	28.02	30.15	41.40
25%	45.67	46.12	46.33

50%	50.19	49.79	50.43
75%	59.36	59.26	61.63
max	75.12	72.76	92.56

Now, the line plot is built for crude oil prices. Similar to gold prices, it is found that it is difficult to find any pattern in the crude oil price. When moving averages are applied, it is seen here also that rough edges are smoothed out.

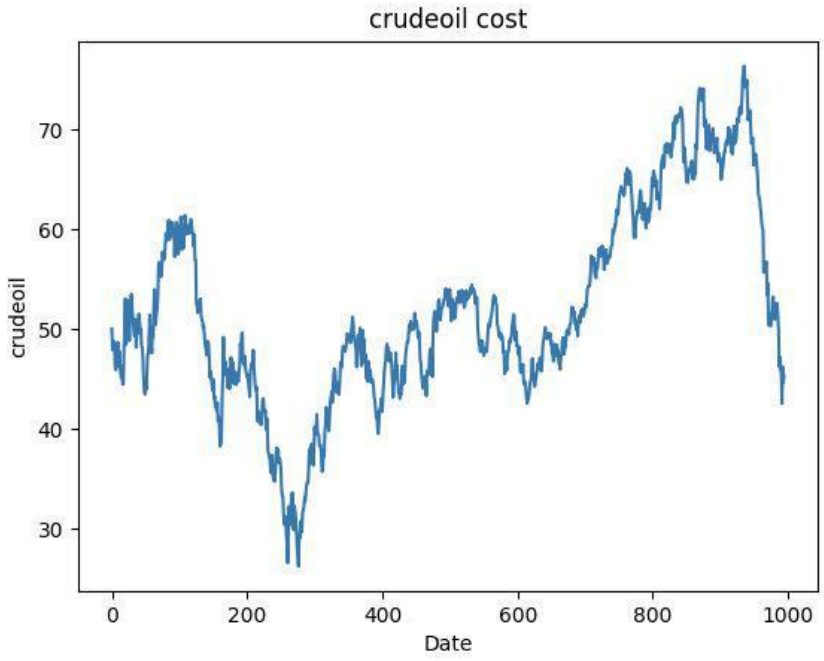

Figure 4.9 Daily crude oil prices with moving averages applied

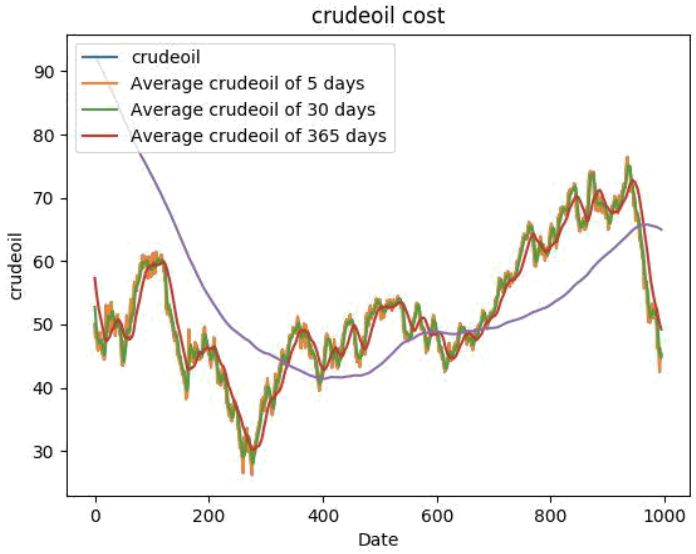

Figure 4.10 Daily crude oil prices with moving averages applied

4.2 SVM Model analysis

4.2.1 SVM Base Model

For all the 4 SVM models, the parameters to be used are decided based on the design decisions followed from (Keung,2016), (Madge,2017) and (Henrique et al., 2018). The kernel is set as linear, C is set as 10, tolerance for stopping criterion is set as 1e-3 and epsilon is set as 0.1 based on these research papers.

Now, the model is built based on stock price data alone. The data is normalised and scaled using the min-max scaler (Liu,2017). This is done to reduce range such that the range is now between 0 and 1. The data is then split as 75 % training data and 25 % test data. The SVM model is fitted based on the training data and the test data values are predicted based on this. A line plot is built to see the difference in the actual test data and predicted test data. It is seen that most of the values that are predicted are well below the actual test data present. Now, the train data, test data and predicted test data are plotted together to get the complete picture.

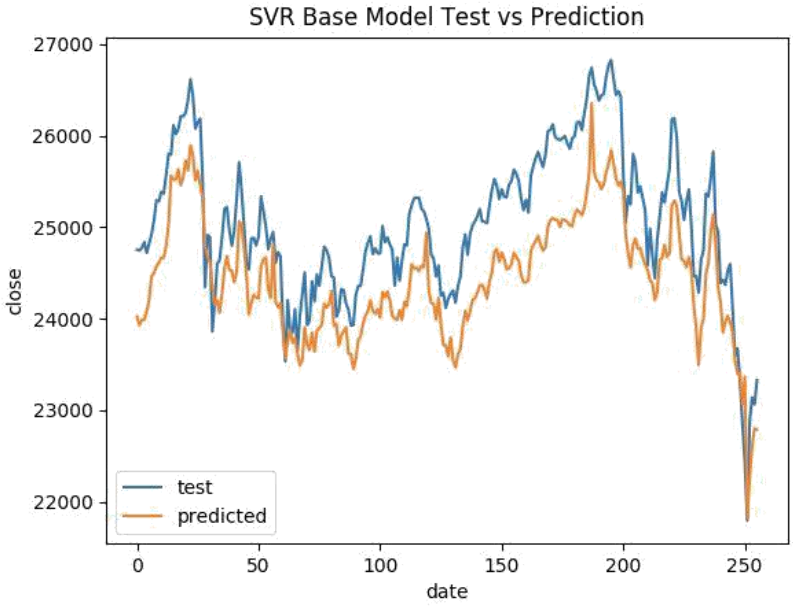

Figure 4.11 SVR Base Model Test vs Prediction.

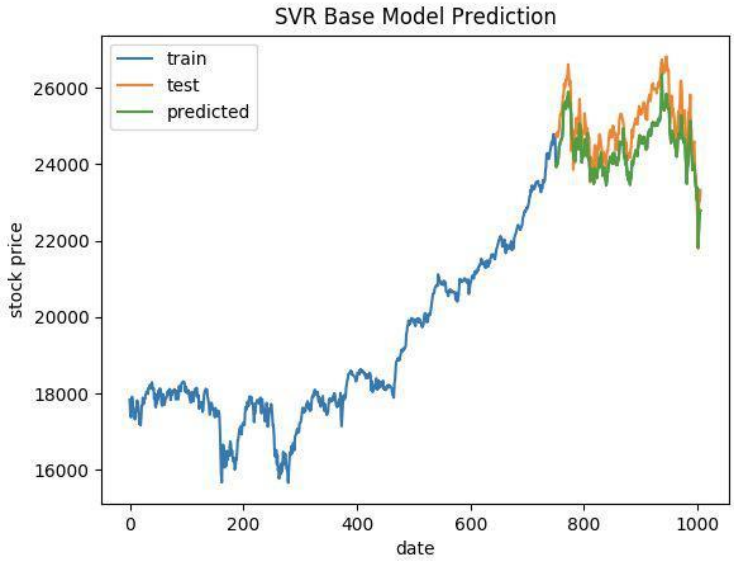

Figure 4.12 SVR Base Model Full Prediction.

Finally, evaluation metrics are carried out based on the actual test data and predicted test data and the results are obtained The results are as follows,

- MSE - 467214.38
- RMSE - 682.63

- MAE - 597.96
- MAPE - 0.30
- R2 - 2.47

It can be seen from the results that the model has produced only a decent approach for predicting stock prices. These similar strategies are repeated for the rest of the SVM models and finally, some insights are obtained based on all the SVM model results.

4.2.2 SVM Base Model with Moving Average

In this model, moving averages are applied to the DJI stock price data. Rest of the steps are followed similar to the SVM base model and the predictions are made. From the actual test data vs prediction test data plot, it can be seen that the prediction is still not up to the desired standards and it greatly under predicts the stock prices.

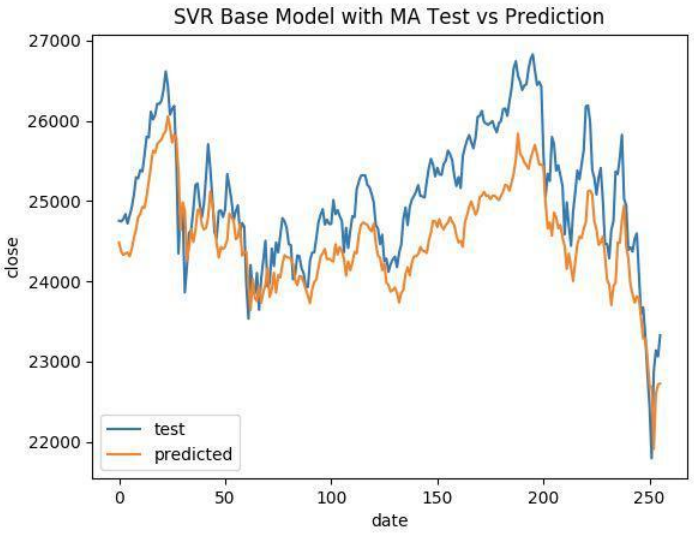

Figure 4.13 SVR Base Model Full Prediction.

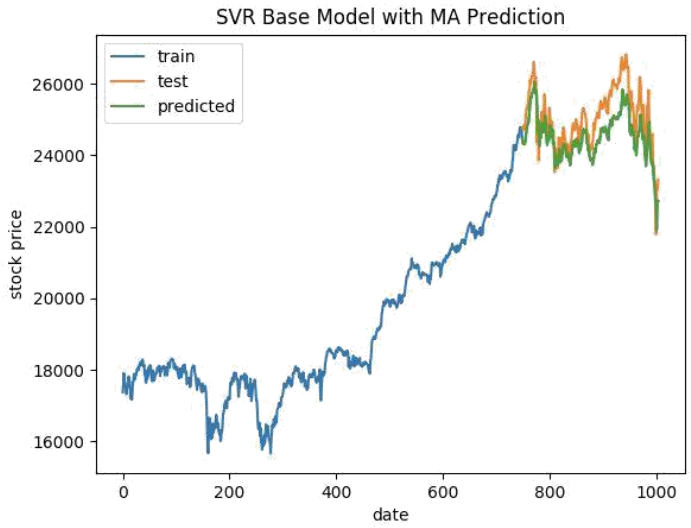

Figure 4.14 SVR Base Model with MA Full Prediction.

The evaluation metrics are carried out and the results are as follows,

- MSE - 424163.43
- RMSE - 651.27
- MAE - 570.54
- MAPE - 2.25
- R2 - 0.36

From the results, it is seen that applying moving averages has improved the prediction of the closing stock prices. But, it is also seen that the improvements are only marginal and still the error rates are high.

4.2.3 SVM Advanced Model

In the advanced model, the DJI stock price data is combined with gold price and crude oil price data. Rest of the steps are similar to the base model. Now, it is seen from the line plot of test data vs predicted test data that the predictions seem to be far worse when compared to the previous two models as it drastically under predicts the test data.

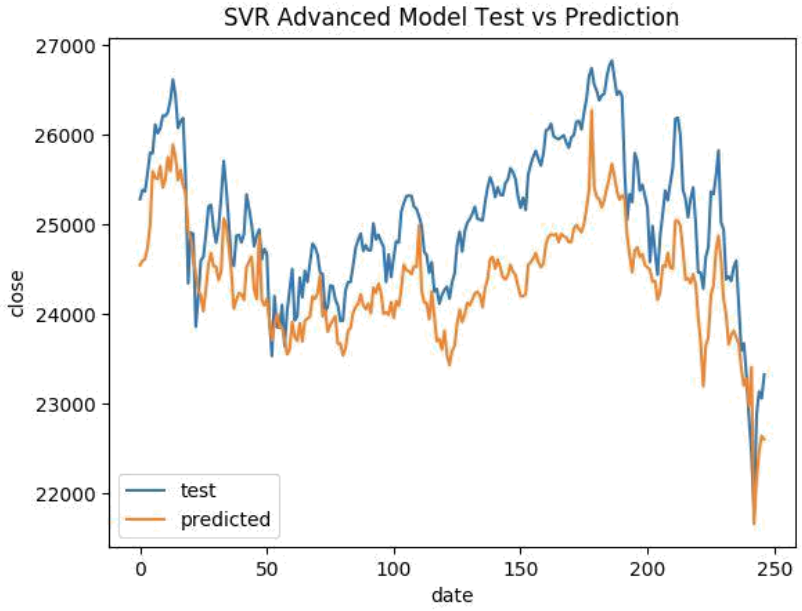

Figure 4.15 SVR Advanced Model Test vs Prediction

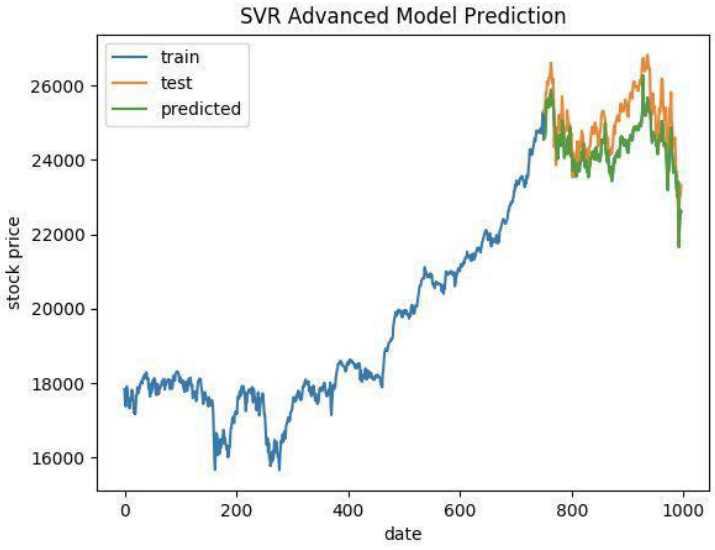

Figure 4.16 SVR Advanced Model Full Prediction.

The evaluation metrics are carried out and the results are as follows,

- MSE - 590974.85
- RMSE - 768.74
- MAE - 700.34

- MAPE - 2.77
- R2 - 2.77

From the results, it can be seen that advanced model results are very poor. This can be attributed to the noise caused by the merging of the datasets without performing any smoothing effects to reduce the noise (Liu, 2017).

4.2.4 SVM Advanced Model with Moving Average

In this, moving averages are applied to the previous combined data and the rest of the steps are followed similarly to build the prediction model. From the actual test data vs predicted test data plot, it can be seen that for initial days, the prediction is comparatively very accurate and goes on gradually to give lower values in prediction as to the increase in days.

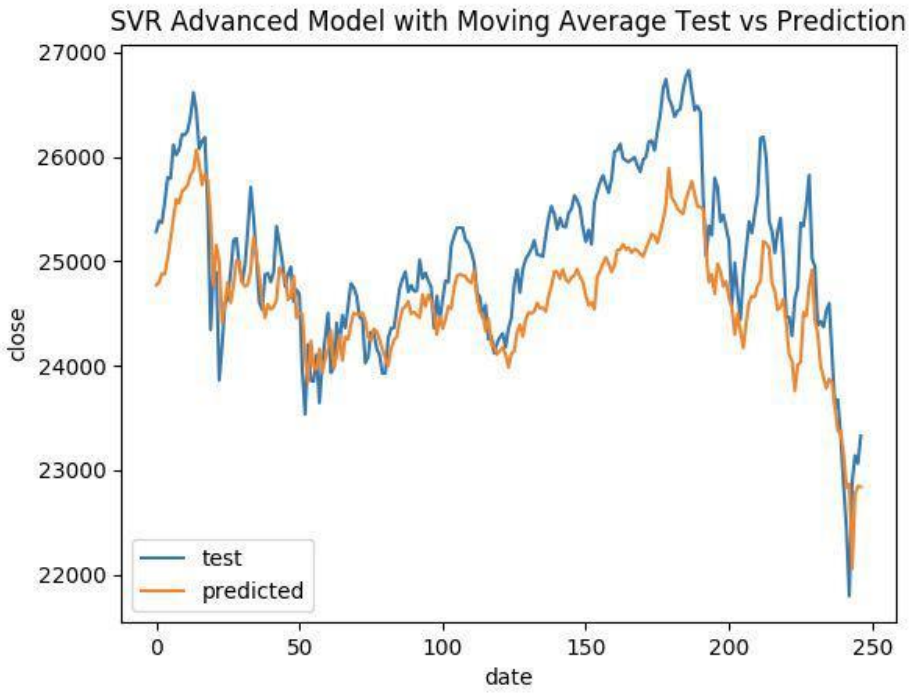

Figure 4.17 SVR Advanced Model with MA Test vs Prediction

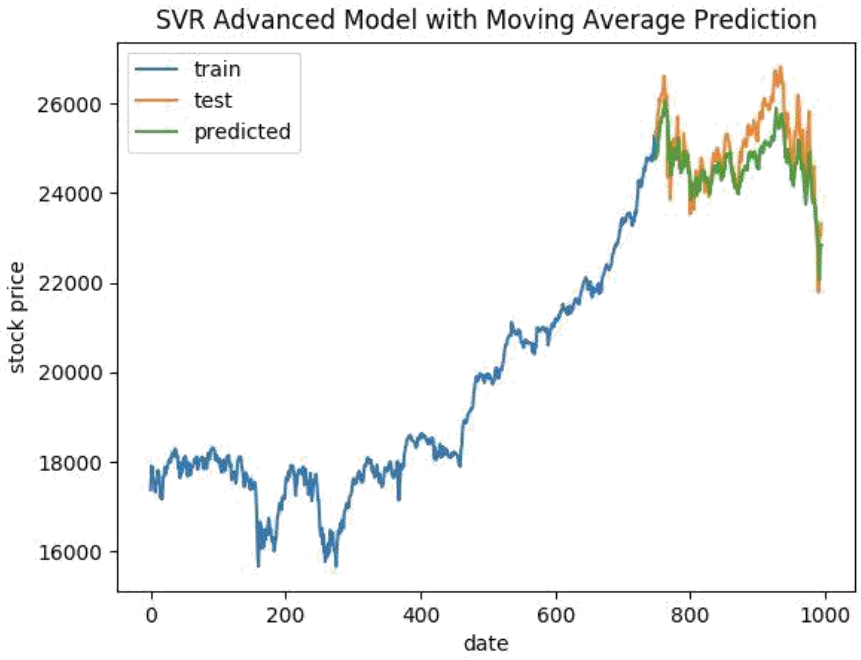

Figure 4.18 SVR Advanced Model with MA Full Prediction

The evaluation metrics are carried out and the results are as follows,

- MSE - 339067.29
- RMSE - 582.29
- MAE - 486.09
- MAPE - 1.91
- R2 - 0.51

From the results, it can be concluded that when moving averages are applied to the combined data, they give substantially good results when compared to all others. It also avoids giving worse results like the SVM advanced model because of the smoothing effect caused by the moving averages to reduce the noise present in the combined data. The error values are also found to be much lower than the other SVM models.

4.2.5 SVM Model Results

Model	RMSE	MSE	MAE	R2	MAPE
SVR Base	682.63	467214.38	597.96	0.30	2.47
SVR Base+ MV	651.27	424163.43	570.54	0.36	2.25
SVR Advanced	768.74	590974.85	700.34	0.15	2.77
SVR Advanced +MV	582.29	339067.29	486.09	0.51	1.91

- From the base model results, it is found out that predicting the stock price values without any additional external input parameters will give only poor results.
- From the SVR advanced result, it is understood that prediction results become only worse when external input parameters are added without incorporating any mechanism to reduce the noise present in the combining of data.
- Moving averages help in improving the prediction performance of the model and SVR advanced model with moving averages gives the best prediction among the 4 SVM models.

4.3 LSTM Model analysis

4.3.1 LSTM Base Model

For all the 4 LSTM models, the parameters to be used are decided based on the design decisions followed from (Nakayama,2018) and (Wang et al, 2018). The batch size is set as 16 and epochs is set as 200 which can be reduced to 100 when there is no substantial improvement in the loss values. The activation function is set as relu and the units are set as 256. Dropout is set as 0.3 to avoid over-fitting. The dense layer is added with 1 unit

Now, the model is built based on stock price data alone. The data is normalised and scaled using the min-max scaler (Liu,2017). This is done to reduce range such that the range is now between 0 and 1. The data

is then split as 75 % training data and 25 % test data. The loss function is then initiated when the LSTM model is compiled. It can be seen that the test and train loss improves over the epochs. Here it is stopped with 100 epochs itself as the loss didn't improve for 40 epochs.

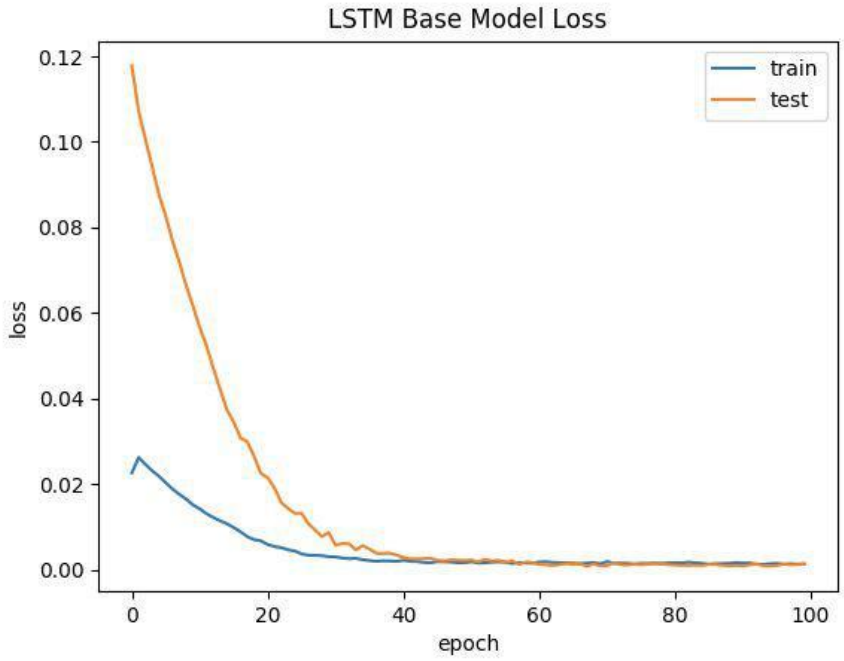

Figure 4.19 LSTM Base Model Loss

The LSTM model is fitted based on the training data and the test data values are predicted based on this. A line plot is built to see the difference in the actual test data and predicted test data. It is seen that most of the values that are predicted are well below the actual test data present. Now, the train data, test data and predicted test data are plotted together to get the complete picture. From the plot, it can be seen that the LSTM base model performs prediction much better than any of the SVM prediction models.

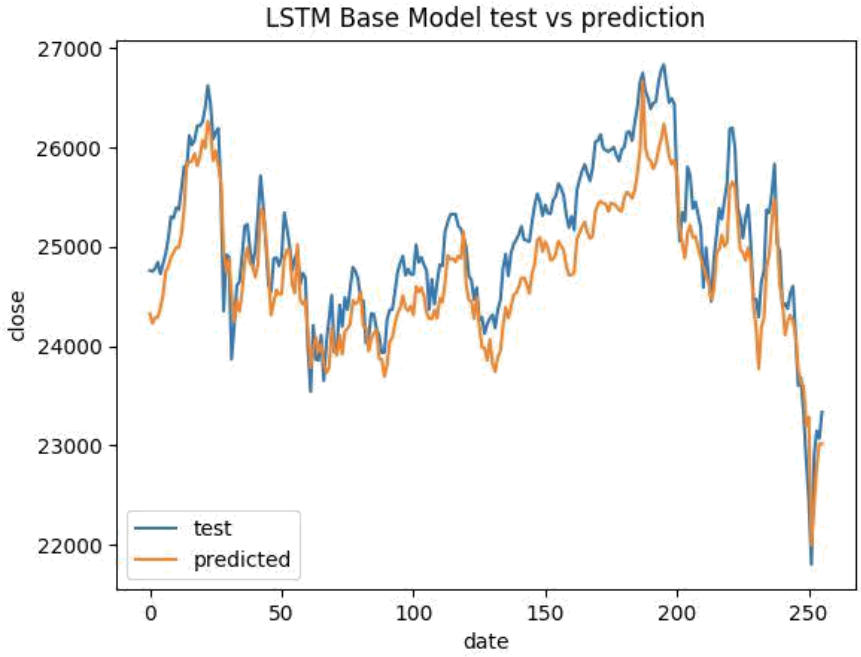

Figure 4.20 LSTM Base Model Test vs Prediction

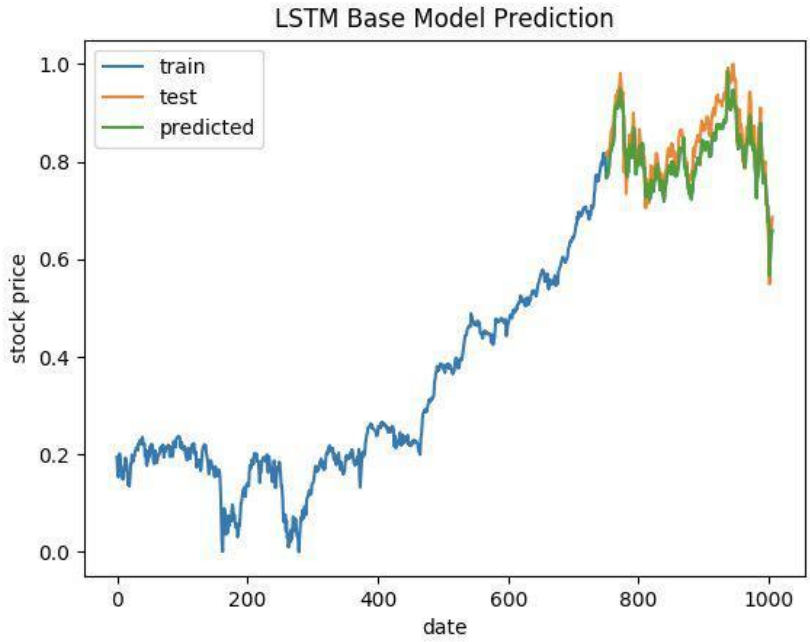

Figure 4.21 LSTM Base Model Full Prediction

The evaluation metrics are carried out and the results are as follows,

- MSE - 159519.52
- RMSE - 399.39
- MAE - 356.04
- MAPE - 1.41
- R2 - 0.76

From the results, it is seen that all the error values are nearly half of the values found in the base model in SVM. This clearly shows the superiority of the LSTM model in the prediction of stock prices.

4.3.2 LSTM Base Model with Moving Average

In this, moving averages are applied to the DJI stock price data and the rest of the steps are followed similar to the base model. From the loss plot, it is seen that the training and testing loss have improved over the epochs.

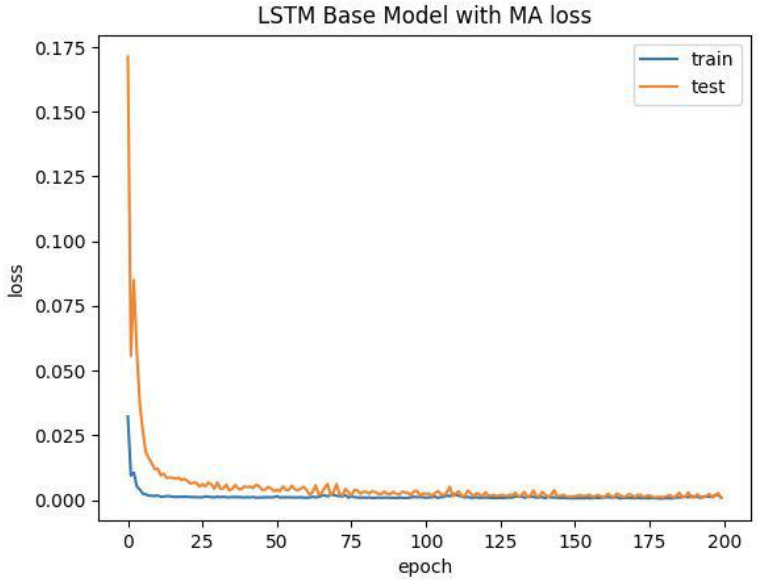

Figure 4.22 LSTM Base Model with MA Loss

From the actual test data vs predicted test data plot, it is seen that the prediction has improved when compared to the base model.

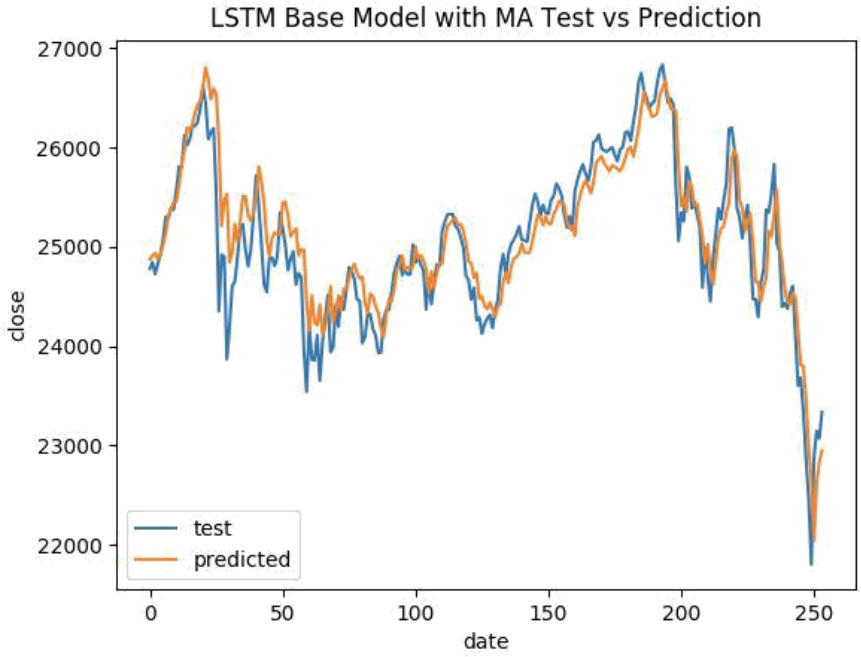

Figure 4.23 LSTM Base Model with MA Test vs Prediction

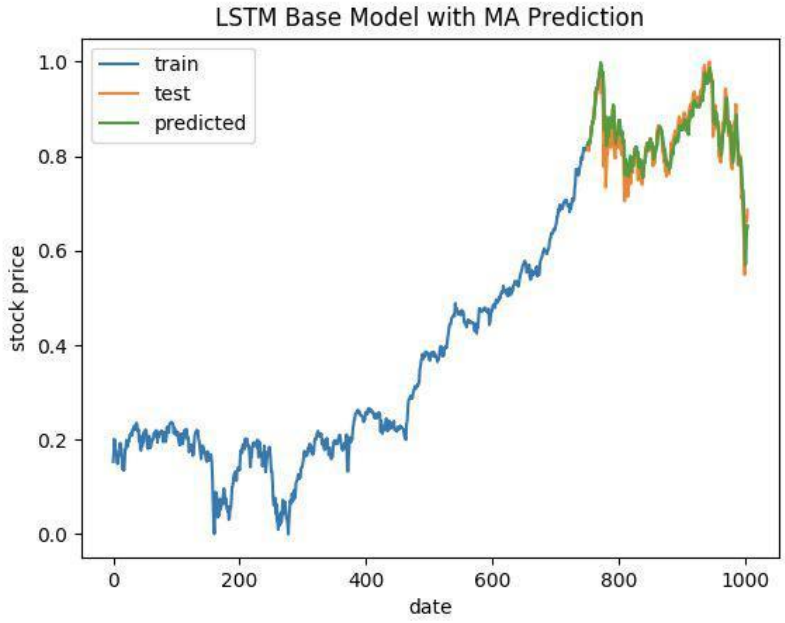

Figure 4.24 LSTM Base Model with MA Full Prediction

The evaluation metrics are carried out and the results are as follows,

- MSE - 122183.47
- RMSE - 349.54
- MAE - 255.73
- MAPE - 1.05
- R2 - 0.81

From the results, it is seen that error values are lower than the base model. This means that moving averages has helped in improving the prediction performance of the model

4.3.3 LSTM Advanced Model

In this, the DJI stock price data, crude oil price and gold price data are combined and the rest of the steps are followed similar to the base model. From the loss plot, it is seen that the training and testing loss have improved over the epochs.

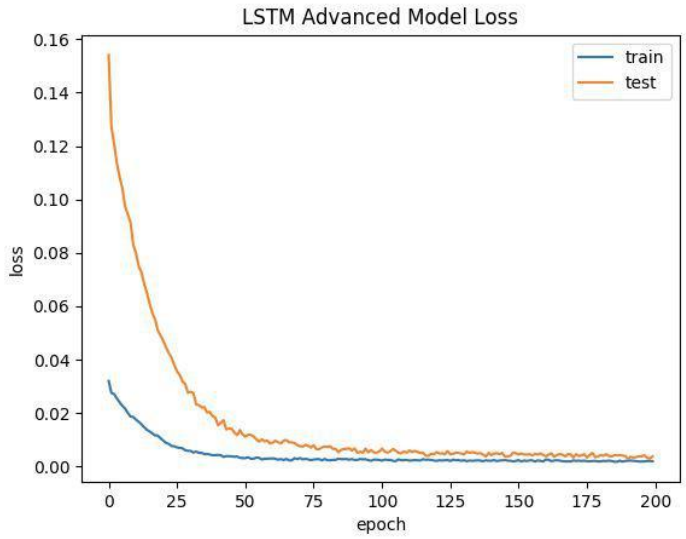

Figure 4.25 LSTM Advanced Model Loss

From the actual test data vs predicted test data plot, it is seen that the prediction has gotten worse when compared to the base model. This is maybe due to the noise present in the combined data.

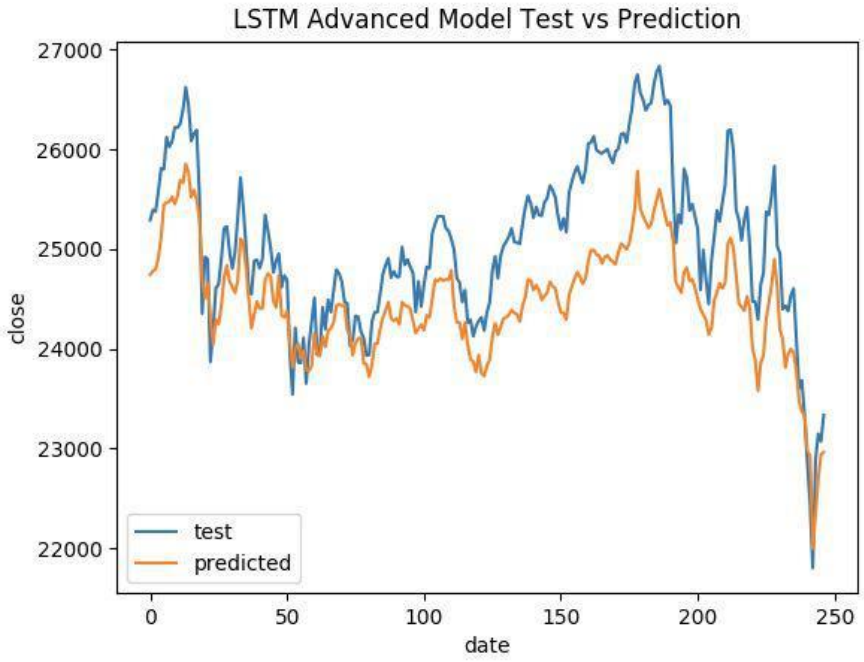

Figure 4.26 LSTM Advanced Model Test vs Prediction

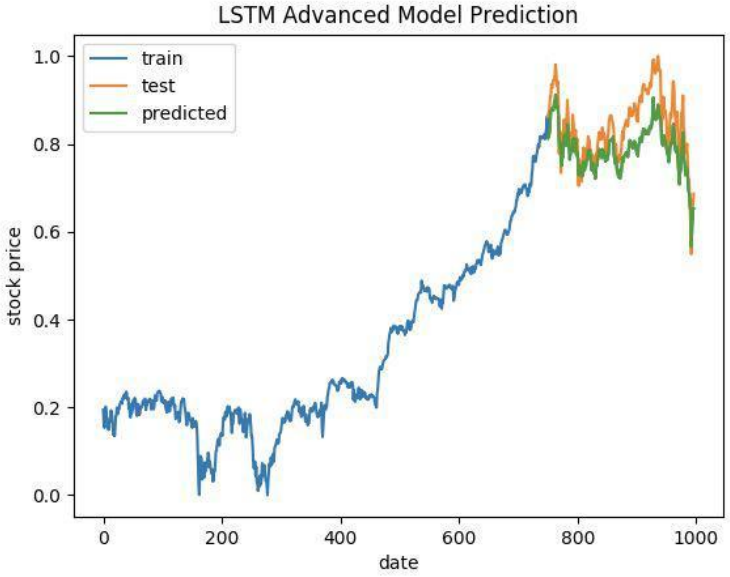

Figure 4.27 LSTM Advanced Model Full Prediction

The evaluation metrics are carried out and the results are as follows,

- MSE - 467426.27
- RMSE - 683.68
- MAE - 603.93
- MAPE - 2.52
- R2 - 0.32

From the results, it is seen that all the error values are roughly twice the values found in the base model in SVM. This clearly shows the inferiority of the LSTM advanced model in the prediction of stock prices.

4.3.4 LSTM Advanced Model with Moving Average

In this, moving averages are applied to the previous combined data and the rest of the steps are followed similarly to build the prediction model. From the loss plot, it is seen that the training and test losses are improved over the epochs.

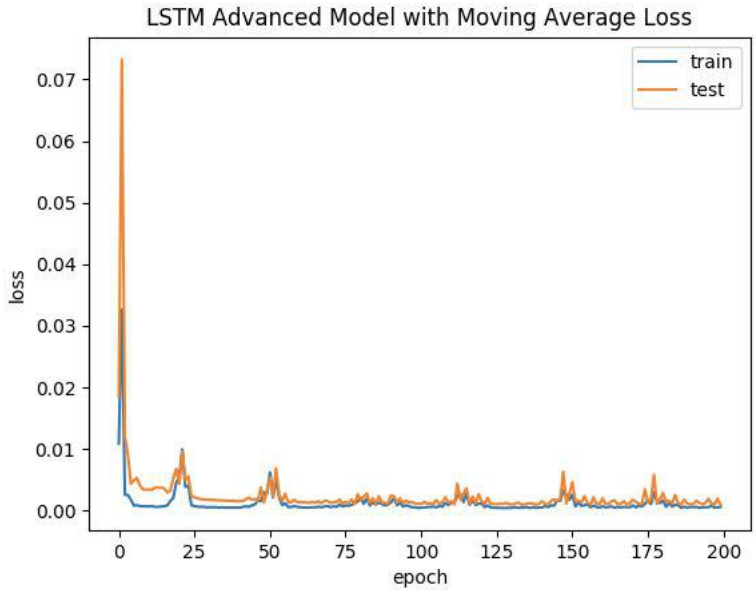

Figure 4.28 LSTM Advanced Model with MA Loss

From the actual test data vs predicted test data plot, it can be seen that prediction is most accurate among all the models developed.

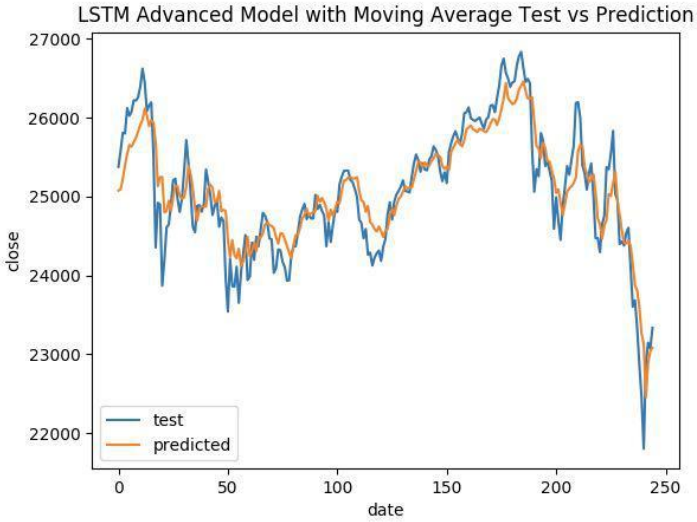

Figure 4.29 LSTM Advanced Model with MA Test vs Prediction

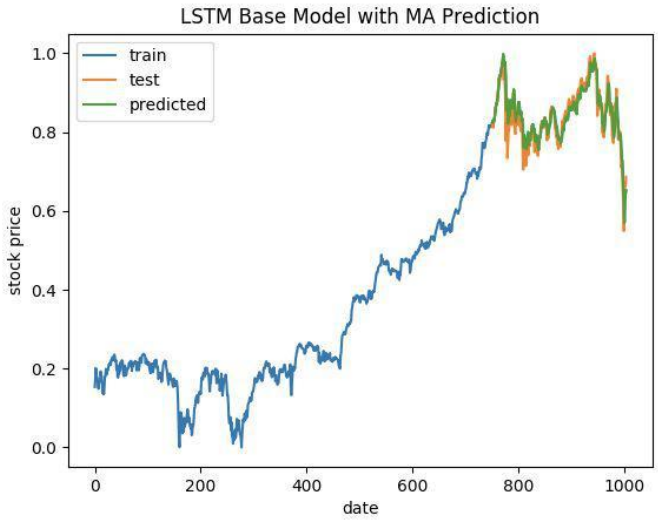

Figure 4.30 LSTM Advanced Model with MA Full Prediction

The evaluation metrics are carried out and the results are as follows,

- MSE - **120731.41**
- RMSE - **347.46**
- MAE - **262.42**
- MAPE - **1.03**
- R2 - **0.83**

From the results, it is seen that all the error values are lowest among the LSTM models. Therefore, it can be considered to be the effect of applying moving averages to the combined data to provide a smoothing effect to the noisy data.

4.3.5 LSTM Model Results

Table 4.6 LSTM model results table

LSTM Base	399.39	159519.52	356.04	0.76	1.41
LSTM Base + MV	349.54	122183.47	255.73	0.81	1.05
LSTM Advanced	683.68	467426.27	603.93	0.32	2.52
LSTM Advanced + MV	**347.46**	**120731.41**	**262.42**	**0.83**	**1.03**

- From the table, it is seen that applying moving averages helps in improving the prediction performance of the LSTM models.
- When the LSTM model is applied to the combined data, it gives the worst result due to the noise present in the data and its nature of high volatility.
- Just like in SVM models, LSTM advanced model with moving averages provide the best result for prediction of closing stock prices.

4.4 Comparison of Results

Table 4.7 Evaluation Metrics Values for the 8 models

Model	RMSE	MSE	MAE	R2	MAPE
SVR Base	682.63	467214.38	597.96	0.30	2.47
SVR Base+ MV	651.27	424163.43	570.54	0.36	2.25
SVR Advanced	768.74	590974.85	700.34	0.15	2.77
SVR Advanced +MV	582.29	339067.29	486.09	0.51	1.91
LSTM Base	399.39	159519.52	356.04	0.76	1.41
LSTM Base + MV	349.54	122183.47	255.73	0.81	1.05
LSTM Advanced	683.68	467426.27	603.93	0.32	2.52
LSTM Advanced + MV	**347.46**	**120731.41**	**262.42**	**0.83**	**1.03**

- Lower MSE score indicates better performance of the model. Here it is seen that LSTM advanced model with moving averages to be having the lowest MSE among the 8 models with a score of **120731.41.**
- Lower RMSE score indicates better performance of the model. Here, it is seen that LSTM advanced model with moving averages to be having the lowest RMSE among the 8 models with a score of **347.46**.
- Lower MAE score indicates better performance of the model. Here, it is seen that LSTM advanced model with moving averages to be having the lowest MAE among the 8 models with a score of **262.42**.
- Lower MAPE percentage indicates better performance of the model. Here, it is seen that LSTM advanced model with moving averages to be having the lowest MAPE among the 8 models with a percentage of **1.03**.
- Higher R2 score indicates better performance of the model. Here, it is seen that LSTM advanced model with moving averages to be having the highest R2

score among the 8 models with a score of **0.83**.

From the results, it is seen that LSTM advanced model to be the one with the best result for the prediction of stock prices. Now, to confirm that this is a statistically significant change from the rest of the models, a t-test is conducted with the LSTM advanced model with moving averages and other 7 models. The null hypothesis or H0 considers that there is no statistically significant change between the two models and alternate hypothesis or H1 considers that there is a statistically significant change between the two models. When the p-value is less than 0.05, the null hypothesis is rejected.

Table 4.8 t-test table

Models	P-value
SVR Base - LSTM Advanced + MV	9.38e-25
SVR Base + MV - LSTM Advanced + MV	1.25e-20
SVR Advanced - LSTM Advanced + MV	2.17e-29
SVR Advanced + MV - LSTM Advanced + MV	1.97e-13
LSTM Base - LSTM Advanced + MV	3.53e-58
LSTM Base + MV - LSTM Advanced + MV	1.57e-4
LSTM Advanced - LSTM Advanced + MV	4.82e-97

From the table, it is seen that p-value is less than 0.05 for all the cases. Therefore, the null hypothesis is rejected and it is concluded that there is a statistically significant change in predicting the stock prices by using LSTM advanced model with moving averages. Therefore, it is concluded that the LSTM advanced model with moving averages has the best result for predicting stock prices.

Chapter 4- Conclusion and Future Works

5.1 Overall Discussion

In this research, a comparative study is done between the SVM models and the LSTM based recurrent neural network models for DJI stock price prediction. It is seen that all the LSTM models outperform the SVM models. This is because the LSTM neural network models used in this study have an in-built capability to forget or remember the relevant information (Venna *et al.*, 2018). This functionality takes into consideration any type of patterns in the dataset without any need for pre-analysis or adjustments. The gated cell in the LSTM network passes only the relevant information into the next stage and ignores the irrelevant information. Thus, the models learn the input information with all the patterns and trends in the dataset and forecast future outcomes more effectively. The forecasting results of these models are hence more reliable than any other prediction models. The LSTM neural networks perform better when trained on a larger dataset. Greater the quality of the training data, higher the accuracy of the model's forecasting results. It is seen that in the advanced models where the external input parameters are added without applying moving averages, they performed worse than the base models. This is seen due to the presence of noise in the combined data. It is also seen that when moving averages are applied, the models perform better than the base models because of the smoothing of the data by moving averages. My contribution to this research work has been to provide a novel method to combine the SVM models built based on (Keung,2016), (Madge,2018) and (Henrique et al. , 2018), the LSTM models built based on (Nakayama,2018) and (Wang,2018), addition of external input parameters like crude oil and gold prices based on (Khanvikar,2019) and (Gerlein et al. , 2016) and moving averages based on (Khanvikar,2019). Based on these, 8 models are developed. I then contributed by building the evaluation metrics to select the best model for prediction based on combining the metrics used in papers like (Henrique et al., 2018), (Patel et al.,

2018) and (Xiao et al., 2013). It is seen from this discussion that this novel method shows significant improvement over the state-of-the-art by integrating knowledge from as aforementioned sources.

5.2 Conclusion

The overall conclusions from this research work are as follows,

- Without moving averages, the SVM and LSTM models on base stock price dataset perform better individually without the addition of external parameters like crude oil and gold price. This is because of the noise present as a result of the merging of the data without adequate preprocessing and smoothing by moving averages.

- Overall, LSTM performs better than SVM in all the scenarios. This is because of its ability to remember or forget the data in an efficient manner than SVM.

- With moving averages, the SVM and LSTM models both perform significantly better on the combined dataset over the standard base dataset. This is because of the smoothing effect of the moving averages on the data which helps in learning the influence of the external parameters on the base stock price in a much better manner.

- Overall, the LSTM model with moving averages applied over the combined dataset was evaluated to be the most efficient model in predicting the stock prices for the future. It is also confirmed that there is a statistically significant change in predicting the stock prices by using LSTM advanced model with moving averages. Therefore, it is concluded that the LSTM advanced model with moving averages has the best result for predicting stock prices.

The comparative study shown in this research provides a better scope in improving the performance of real-time stock price prediction. But there are also drawbacks present in this research. It doesn't consider external disaster events like cyclones, tsunami and terrorist attacks etc. It considers those situations to be normal and therefore can't predict how the stock prices will be during such calamity. Similarly, it is not capable of handling situations like prediction during stock market crashes as no adequate training is providing for such cases.

5.3 Future Works

In this study, the implementation of SVM and LSTM using moving averages are done separately. For future works, intraday prices can also be used to compare the values and to understand the volatility of the stock, crude oil and gold prices in a better manner. The stock sell and buy data can also be used to understand how the stock price and external factors surge and dip have influenced the buying and selling pattern. This will help in developing a more accurate prediction. The models can also be extended to provide live interactive predictions based on the user given data and subsequently can be used for other forecasting problems like weather forecasting, disease forecasting and house price forecasting etc.

References

1. Karen Lin (2019) *Role of Data Science in Artificial Intelligence.* Available at:https://towardsdatascience.com/role-of-data-science-in-artificial-intelligence-95 0efedd2579

2. Disa Mishal (2019) *4 Reasons Why You Should Use Deep Learning For Time Series Forecasting.*Available at: https://www.analyticsindiamag.com/4-reasons-why-you-should-use-deep-learning-for-time-series-forecasting/

3. Andy (2018a) *Keras LSTM tutorial - A powerful deep learning language model- Adventures in Machine Learning, adventuresinmachinelearning.com.* Available at: http://adventuresinmachinelearning.com/keras-lstm-tutorial/.

4. Andy (2018b) *The vanishing gradient problem and ReLUs - a TensorFlow investigation - Adventures in Machine Learning, adventures in machine learning blog.*Available at: http://adventuresinmachinelearning.com/vanishing-gradient-problem-TensorFlow/

5. Gandhi (2018) *Support Vector Machine — Introduction to Machine Learning Algorithms.*Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-l earning-algorithms-934a444fca47

6. Bhattacharyya (2018) *Support Vector Regression Or SVR.* Available at: https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff

7. Sayad (2014) *Support Vector Regression.* Available at: http://www.saedsayad.com/support_vector_machine_reg.htm?source=post_page-----8eb3acf6d0ff----------

8. Liana Cipcigan (2013)*'Forecasting Electric Vehicle charging demand using Support Vector Machines'.* doi: 10.1109/UPEC.2013.6714942

9. Christopher Olah (2015) *Understanding LSTM Networks, Colah Blogs.*

Available at: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

10. Artificial Intelligence - Skymind (2017) *A Beginner's Guide to LSTMs, Skymind*. Available at: https://skymind.ai/wiki/lstm.html

11. Ahmed, N. K. et al. (2010) *'An Empirical Comparison of Machine Learning Models for Time Series Forecasting'*, Econometric Reviews, 29(5–6). doi: 10.1080/07474938.2010.481556.

12. Shi Yan (2016) *'Understanding LSTM and its diagrams'.* Available at:

https://medium.com/mlreview/understanding-lstm-and-its-diagrams-37e2f46f1714

13. Andrew Thomas(2018) *'Keras LSTM tutorial – How to easily build a powerful deep learning language model'*. Available at: https://adventuresinmachinelearning.com/keras-lstm-tutorial/

14. Hayes (2019) *Simple Moving Average - SMA Definition.* Available at: https://www.investopedia.com/terms/s/sma.asp

15. Rayome (2018) *Why Python is so popular with developers: 3 reasons the language has exploded.* Available at: https://www.techrepublic.com/article/why-python-is-so-popular-with-developers-3-r easons-the-language-has-exploded/

16. Klein (2018) *Numpy Tutorial.* Available at: https://www.python-course.eu/numpy.php

17. Nnamdi (2019) *Top 5 Machine Learning Libraries.*Available at:

https://blog.bitsrc.io/top-5-javascript-machine-learning-libraries-604e52acb548

18. Xiaotao Liu, Kin Keung Lai (2016) *'Intraday volume percentages*

forecasting using a dynamic SVM based approach'.

19. Hiransha Ma, Gopalakrishnan E.Ab, Vijay Krishna Menonab, Soman K.P (2018) *'NSE Stock Market Prediction Using Deep-Learning Models'.*

20. Ashutosh Kale, Omkaar Khanvilkar (2018) *'Forecasting Indian Stock Market Using Artificial Neural Networks'.*

21. Masaya Abe, Hideki Nakayama (2018)*'Deep Learning for Forecasting Stock Returns in the Cross-Section'*.

22. Liew, Jim Kyung-Soo and Mayster, Boris(2017), *'Forecasting ETFs with Machine Learning Algorithms'.*

23. Saahil Madge (2018), *'Predicting Stock Price Direction using Support VectorMachines'*. Available at: https://www.cs.princeton.edu/sites/default/files/uploads/saahil_madge.pdf

24. Bruno Miranda Henrique, Vinicius Amorim Sobreiro, Herbert Kimura(2018), *'Stock price prediction using support vector regression on daily and up to the minute prices'*. https://doi.org/10.1016/j.jfds.2018.04.003

25. Jigar Patel et al. (2014) *'Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques'.* https://doi.org/10.1016/j.eswa.2014.07.040

26. Gerlein et al. (2016) *'Evaluating machine learning classification for financial trading: An empirical approach'*. https://doi.org/10.1016/j.eswa.2016.01.018

27. Xiao et al. (2013) *'Ensemble ANNs-PSO-GA Approach for Day-ahead Stock E-exchange Prices Forecasting'.* International Journal of Computational Intelligence SystemsVolume 6, Issue 1, February 2013, Pages 96-114. doi: 10.1080/18756891.2013.756227

28. Nayak et al. (2015) *'A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices'.* https://doi.org/10.1016/j.asoc.2015.06.040

29. Barak et al. (2014) *'Developing an approach to evaluate stocks by forecasting effective features with data mining methods'.* https://doi.org/10.1016/j.eswa.2014.09.026

30. Reddy (2018) *'Stock Market Prediction Using Machine Learning'.* doi: 10.13140/RG.2.2.12300.77448

31. Qian et al. (2019) *'Stock Prediction Based on LSTM under Different Stability'.* doi: 10.1109/ICCCBDA.2019.8725709

32. Wang et al. (2018) *'LSTM Model Optimization on Stock Price Forecasting'.* doi: 10.1109/DCABES.2018.00052

33. Vorhies (2016) *CRISP-DM – a Standard Methodology to Ensure Good Outcome - Data Science Central, Data Science Central Blogs.* Available at: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome

34. Ioffe et al. (2015). Available at:https://arxiv.org/pdf/1502.03167v3.pdf

35. Alvira Swalin (2018) *"Choosing the Right Metric for Evaluating Machine Learning Models"* Available at: https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-mod els-part-1-a99d7d7414e4

36. Jason Brownlee (2016) "*How to Normalize and Standardize Time Series Data in Python"* Available at: https://machinelearningmastery.com/normalize-standardize-time-series-data-python

37. Yuxi Liu (2017) *'Python Machine Learning By Example'.* Available at: https://www.oreilly.com/library/view/python-machine-learning/9781783553112/

38. Keras Documentation (2015) *Normalization Layers - Keras document* Available at: https://keras.io/layers/normalization