Financial Time Series Analysis with Machine Learning

Submitted as Research Report / Honours in SIT723

06-June-2022

T1-2021

Satvik Sharma

STUDENT ID 218595095

COURSE - Bachelor of Software Engineering Honours (S464)

Supervised by: Prof. Maia Angelova

Abstract

This article surveys the publications available on the financial time series analysis with machine learning and aims to answer what machine learning models are most commonly used, how accurate the models can be and what is the future like for the field. Different machine learning techniques have been discussed. A number of papers have been read and a technical analysis on the basis of each paper has been done and according to which a conclusion has been drawn, which is, each of the machine learning model has its own pros and cons, but still is not precise to rely on the forecast. The paper follows the workings of the statistical machine learning models, more closely on the machine learning models from the ARIMA family. the dataset for the stock has been chosen from the top five stocks from the Nifty-50, and their working from 2011 to 2020 is discussed. the models have been evaluated using Mean Absolute Error, Root Mean Squared Error and R2 score after analysing all the algorithms, it was found that ARIMAX model's trend was as close to the actual trend, even though in some of the cases SARIMAX was also performing better. Both of these models followed the trend as close to the actual values and when predicting stocks for the future these two would be optimal choice for the data analysts. More future work is required, like working on the longer, shorter and mid term duration of the training dataset and hybridisation of these models, which could be continued in the next iteration of the unit.

Contents

List of Figures

List of Tables

1 Introduction

Time Series can be defined as the study of dynamic consequences over a period of time. This can be represented with a simple first order linear equation, for example in the equation
yt=Θyt-1+c
yt represents the value to be taken out at time t in respect to the changes (Θ), external factors(c) and previous value at time t-1 [12]. This article focuses on the financial time series, where the study is more concerned with the financial assets like stocks, shares, currency evaluation, etc. Even though the study of financial time series is a part of time series, but it is highly logical area, where the uncertainty is extremely high, let's say for example the asset returns in the stocks' time series cannot be observed directly. The addition of uncertainty, statistical theory, methods and high volatile market makes financial time series analysis different from regular time series analysis [28].

1.1 Introduction to Financial Time Series

Financial time series has always been of interest of business and financial analysts and when machine learning started to gain popularity, more and more publications have kept adding to the finance literature as determining more effective ways of prediction is important for right investments. This paper will survey the publications available used for financial time series analysis with machine learning and will try to answer the following questions

- What machine learning models are used for financial time series analysis?
- Difference between the most commonly used machine learning algorithms
- Future of machine learning in the field of financial time series analysis

1.2 Objectives

After answering those questions in the literature review, we can start working on the machine learning models and learn how the statistical models work. For the project, I am choosing the data set from the stock from the Nifty-50 for the time period of ten

years from 2011 to 2020 and working on the rate of return value for each of the stock chosen. The rate of return for the whole research is defined as
Rate of return= ((closing value-previous close value)/previous close value) x 100
The models that have been compared for the research analysis belong from the ARIMA family, and see how they are working and how accurate its prediction is. The accuracy will be determined with the evaluation criteria based on root mean squared error, absolute mean error and R2 score and will try to answer the following questions.

- Different patterns in each of the algorithms
- Possibility of hybridisation
- Creation of customer portfolios
- Future research work

1.3 Resources used for the research project

- Google Scholar: most of the literature reviewed has been found on the google scholar and reading different published articles.
- Kaggle: the dataset has been chosen from the Kaggle website.
- Jupyter Notebook/Anaconda: Jupyter notebook available on anaconda is used for all the machine learning algorithms
- Trello: Trello board is used for managing all the tasks of the project.
- Teams' folder: Microsoft teams is used as a cloud folder for all the tasks, documents et cetera.

2 Literature Review

There are hundreds of publications on stock market, trading systems, forex, et cetera that use the machine learning technologies like Artificial Neural Network, Evolutionary Computations, Genetic programming, Hybrid techniques or some other [25].

Some of the papers also studied the traditional time series models and compared them with the machine learning techniques, for example in one of the papers written by Xin-Yao Qian, ARIMA was used in comparison with the Logical regression, SVM, and Denoising Autoencoders and it was found that the other machine learning models performed better than ARIMA because the other machine learning models took externals factors into account as well [23]. One of the papers, compares the performance of LSTM and ARIMA, and in conclusion it was found that LSTM based models run much better than the ARIMA based models even if the dataset was for one month [26].

2.1 Literature Review Analysis

There are number of papers which have been mentioned in the table below, that have been reviewed and thus analysed what machine learning algorithms outperforms others. The table will show a small summary of what machine learning models were used and what was the conclusion of the papers.

Table 1: Table reviewing literature.

Model Name	Paper	Performance Criteria	feature data-set	time period	environment	conclusion
ANN, LSTM, MLR	[16]	MSE, r, RMSE	Prediction of irrigation groundwater quality	-	Python	ANN and MLR model have highest accuracy in multiple scenarios
ARIMA, LSTM	[26]	percentage Reduction in RMSE	Dow Jones Index	1-month	Python with Keras and Theano	LSTM based models work better than ARIMA

Continuation of Table 1						
Model Name	Paper	Performance Criteria	feature data-set	time period	environment	conclusion
ARIMA, LSTM	[10]	RMSE	DAX30, S and P500	10-years (2004-2014)	-	KNN outperforms other models
KNN, SVM, Gaussian progress	[1]	-	M3 time series competition data	Monthly- for thousand time series	-	Best two models are MLP and gaussian progression
ANN, ARMAX, 3-D Hydrody - namic model	[30]	MAE, RMSE, R	Water level fluctuation in alpine lake	1-year (2009-2010)	Python Statsmodels	The best working models are ANN and ARIMAX
ARIMA, ANN, LSTM	[19]	P values, graphical observations	Dell's stock price	1 year (2010)	-	ANN model is better than ARIMA model and LSTM, but the hybrid of ARIMA-GARCH model can be used for more accuracy.

Continuation of Table 1						
Model Name	Paper	Performance Criteria	feature data-set	time period	environment	conclusion
ANN, LSTM	[29]	nRMSE, MAPE, R2	Prediction of Solar irradiance	Hourly	Google Colab, Python scikit-learn, Keras	Due to small dataset, the ANN model works better than the LSTM, but both works much better than the persistence model
ARIMA, GARCH, regime-switching	[8]	RMSE	Home price indices by the OFHEO	20-years (1980-2000)	-	Regime switching models perform better than ARIMA and GARCH

Continuation of Table 1						
Model Name	Paper	Performance Criteria	feature data-set	time period	environment	conclusion
ARIMA-ANN, ARIMA-Kalman	[18]	MAE, MAPE, MSE	Wind speed prediction	-	-	Both hybrid models have good forecasting accuracy and suitable for wind samplings.
ARIMA-SVM, ARIMA-ANN, ARIMA-Random Forest	[17]	-	indian stock trend	5-years (2004-2009)	MATLAB6.1, SPSS13.0	The hybrid model ANN-ARIMA, was able to predict great values than other models

Continuation of Table 1						
Model Name	Paper	Performance Criteria	feature data-set	time period	environment	conclusion
AR, ARDL, KNN, SVR, Naïve, VAR, MLP	[31]	RMSE, R2	Inflation forecasting	30-years (1984-2014)	-	SVR and ARDL outperforms other models and machine learning models work best with more volatile and irregular series.
Hybrid ARIMA models	[9]	-	Canadian lynx time Series, sunspot time series, airline and star data	Different time periods	-	The hybrid system leads to a higher accuracy in prediction
ANN, KNN	[22]	-	Recorded EEG signals	Patient data set	-	ANN classifier's accuracy and sensitivity was higher than that of KNN classifier.

Continuation of Table 1						
Model Name	Paper	Performance Criteria	feature data-set	time period	environment	conclusion
ANN, LSTM, MLR	[16]	MSE, r, RMSE	Prediction of irrigation groundwater quality	-	Python	ANN and MLR model have highest accuracy in multiple scenarios
LR,LSTM	[15]	RMSE, MAPE	For hire vehicles and yellow taxi	-	-	LR is used to select the important variables and LSTM helps to improve the accuracy.
ARIMA	[3]	-	New York Stock exchange and Nigeria stock exchange	-	python	ARIMA has strong potential for short term prediction

Continuation of Table 1						
Model Name	Paper	Performance Criteria	feature data-set	time period	environment	conclusion
MLR, KNN, ANN, ANFIS	[14]	Mash-Sutcliff Coefficient	Stream flow prediction	Monthly	-	The accuracy of each of the model depends on the condition, but the hybridisation was effective.
End of Table						

2.2 Common methodologies used

After observing the table above, it can be clearly seen that the ARIMA is not a good option for the analysis of financial time series even though it is one of the most common methods used. Multiple machine learning methods like LSTM and ANN are also most commonly used as well as have good prediction accuracy as well. hybrid models have the highest accuracy among the machine learning models. The table below enlists the popular models from the journals and books published.

Table 2: Methodologies used in the reviewed literature.

Model Name	Total Papers	Paper
ARIMA	5	[23], [26], [19], [8], [3]
ANN	7	[10], [30], [19], [29], [22], [16], [14]
LSTM	5	[26], [19], [29], [16], [15]
LR	2	[23], [15]

Continuation of Table 2		
Model Name	Total Papers	Paper
SVM	3	[23], [10], [1]
KNN	4	[1], [31], [22], [14]
Hybrid Models	3	[18], [17], [9]
others	9	[23], [1], [30], [8], [18], [17], [31], [16], [14]
End of Table		

2.3 Some common Time series evaluating techniques

This part of the literature review will give succinct details about some of the most popular machine learning models used for the analysis of financial time series.

- ARIMA: ARIMA model is basically the integration of AR (Auto regressive) and MA (Moving Average) and is capable of working with the non-stationary data. It is also referred to as box Jenkins models as it was popularized by George Box and Gwilym Jenkins [21]. It can be represented mathematically as shown below
  yt = Σpi=1 iyt-I + Σqi=1 iut-I + ut
  where yt stands for the goal variable, with which the values of yt+1, yt+2, and so on can be determined [13]. It is one of the most common statistical methods used for financial time series analysis as from the reviewed literature five of the papers were on ARIMA.

- ANN: ANN or Artificial Neural Network are inspired from the human brain's neurological functions and made in such a manner that it replicates its decisions similar to humans and can be created by programming computer to behave like neurons. Mathematically, it can be represented simply as
  hΘ(x)=1/(1+e-ΘxT)
  where hΘ(x) is the output, x is the input and Θ is the parameter vector [20].

- KNN: KNN or K-Nearest Neighbour can be called as the one of the simplest algorithms which classifies the data point on the basis of its neighbours. It will be suitable for big datasets for analysis and prediction [27].

- LSTM: LSTM or long short-term memory belongs to recurrent neural network architecture and consists of memory cells that store information which can be updated from time to time by input, output and forget gate [6].
- LR: LR or Logistic Regression is statistical analysis method that can be used to predict a binary value on the basis of the data observed. The model works with the binary data, that is 0 meaning the event does not happen, and 1 the event happened [7].
- Hybrid models: Hybrid models are the combination of two or more algorithms, and this is a major approach towards more accurate and reliable methods, because it benefits from the two methods and thus reach higher performance. The examples of the hybrid models are DTFNN, ARIMA-ANN, et cetera [2].

These are the most commonly used algorithms that have been reviewed and used for the analysis of the financial time series.

2.4 Gaps in current literature

Financial time series is a very popular and complex branch of time series analysis. The machine learning techniques have taken the branch to new levels. The aim of the paper was to answer what machine learning techniques that are being used for financial time series analysis, finding the best one and what void still needs to be filled. The common algorithms that are used for the financial time series analysis are ARIMA, ANN, LR, LSTM, SVM, KNN, MLP, decision tree, random forest and the hybridisation of these. Even though the study is not just limited to these machine learning algorithms, but these are the most commonly used ones due to their higher accuracy than the other algorithms or their ease of implementation. Moreover, the accuracy of each of the model also varies, depending on the dataset as well, for example if the dataset is huge, LSTM will work better. All the models have their own pros and cons, but after review of number of papers, hybrid models outperform the basic machine learning models. The gap that needs to be filled for the betterment of the topic is to study and develop more machine learning algorithms, whether it be by creating new ones, or creating new hybrids of existing models for higher accuracy, thus a great opportunity for the researchers in this field. The next part of the dissertation will consist of the development of the artefact, it's evaluation and future work regarding the topic.

3 Research design

The main objective of the project is to model financial time series with machine learning algorithms available. The first part of the project is done with the help of existing machine learning algorithms available and working on them. each of the model tested will be evaluated with RMSE, R2, and mean absolute error. Since, as mentioned in the literature review that each of the model has its own pros and cons, it is a better idea to compare the models in the first phase of the research. For the research I have chosen six models, which will be compared to find out the best. the benefit of doing this would come in handy when hybridising them with the other models. The second phase of the research will be to hybridise them and work on the client portfolios.

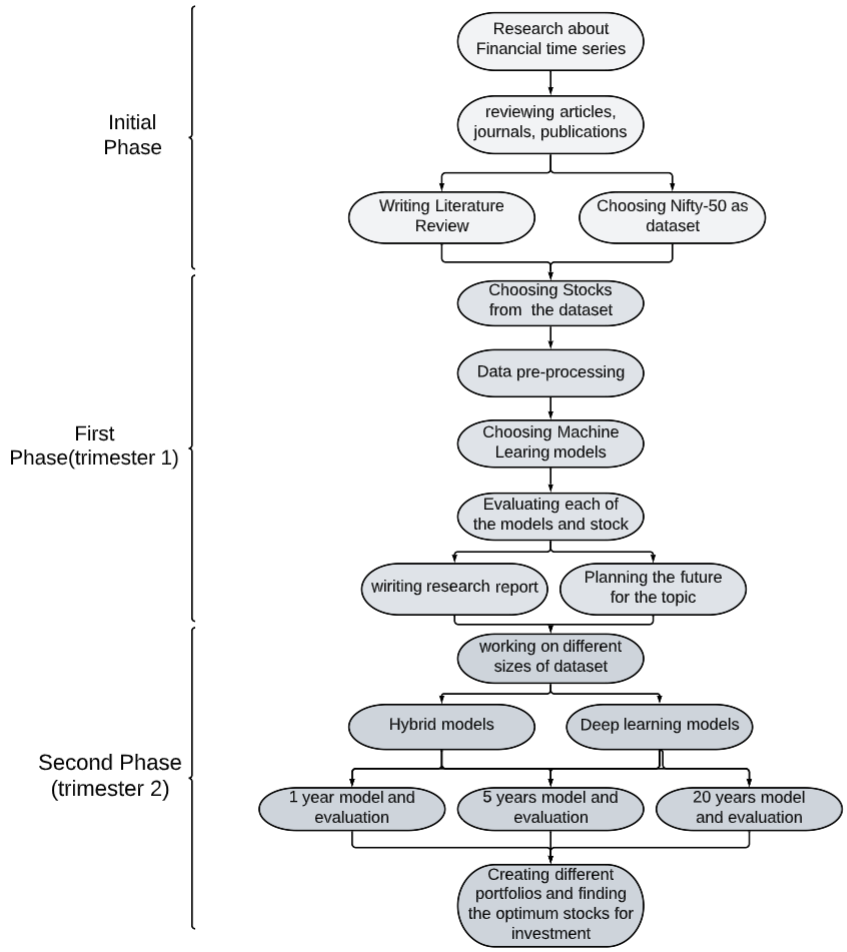

Figure 1: Development Phases for the Project

3.1 Experimentation design

For the analysis of financial time series, the project was divided into number of phases, that has been followed in order to create an optimal research thesis. The steps taken are as follow

- Read and review a number of literatures, articles, book, journals et cetera to find out the current approach and gap in the analysis of financial time series.
- Finding and researching about different machine learning algorithms that can be used to fix the gap in the current learning.
- Create a number of machine learning models and evaluate them and find out if they cover the gaps or not.

For the project, ten years of the Nifty-50 data from 2011-2020 has been chosen [24]. The whole project works on the rate of return which is defined as
Rate of return= ((closing value-previous close value)/previous close value) x 100
And this will be extremely helpful in finding out if it the valuation of the stock is positive or negative. The statistical models have been chosen for this part of the project and are down as follow

- Auto regression(AR)
- Auto Regressive integrated moving average (ARIMA)
- Auto Regressive integrated moving average with Exogenous variables (ARIMAX)
- Seasonal Auto Regressive integrated moving average (SARIMA)
- Seasonal Auto Regressive integrated moving average with Exogenous variables (SARIMAX)
- Holt-Winter's Exponential Smoothing (HWES)

All of these models have been tested and compared against each other and evaluated. this is the basis on what the whole research design has been created.

3.2 Goals of the research

The goals that this research report needs to answer are:

- Find the current gaps by review the existing literature.
- Design a solution to fix the gaps in the existing literature.
- Test the solution proposed.

4 Artefact Development

The artefact is divided in two parts and the first part is to review literature, find gaps in it, and then create machine learning models, and evaluate them. This will be helpful in finding out what machine learning algorithm works most accurately and what next steps of the project could be. The second part of the model will be more focused in creating hybrid/deep learning models and creating client portfolios. For now, the project is just focusing on the first part of the project.

4.1 Dataset and Data pre-processing

The dataset required for the analysis of financial time series analysis should be related to something finance or money related. This could have included the stock market prices, gold prices, crypto currency evaluation et cetera. For this project I have chosen Nifty-50 stock data from 2011 to 2020 [28] for the testing and prediction for six months. The data consists of top-50 stocks from the national stock exchange, India. All of the data is of format comma separated values, which is ideal format for machine learning and is stored in the local file, but later on during the second iteration of the research project, the data could be shifted to more reliable cloud database like google drive.
Each of the CSV file contains a number of heading but only two of the headings are most important for the project, that is, Previous close value and closing value. These two values are important to find the rate of return, which as defined above is, rate of return is defined as the closing value subtracted from the previous close value divided by previous close value multiplied by 100. The index of the dataset is date starting from 01/01/2011 to 31/12/2020 with the frequency of Business days, since

the stock market is only opened on business days. Furthermore, there are a number of Null values, which have been fixed by assuming the value of close value from last observation.

4.2 Description of models used

- Auto Regression (AR): autoregression works on the past values to predict the future values. It works on the co-relation between different variable to predict and tells what lag is required for the dataset. The stronger the correlation, the more weight will be put on the autoregression variable. The number of lags can be identified from the library stastmodels.tsa.ar model. The autoregression is a part of the package statsmodels.tsa.ar model. When fitting the model, it requires training dataset and lags.

- Auto Regressive integrated moving average (ARIMA): This is a statistical model that integrates autoregressive and moving average terms together to predict the future. The model is based on the terms p, d and q, and there are two different methods to identify them. one of them is drawing ACF and PACF curves and the other method is using the method auto arima, which automatically tells the p, d and q values. When fitting the ARIMA model, training dataset and the order of p, d and q is required.

- Auto Regressive integrated moving average with Exogenous variables (ARIMAX): this model is same as ARIMA model, except for the fact that it takes external factors into consideration while making a prediction. This uses same libraries as the ARIMA model but while fitting the model, we need to declare the exogenous factors along with training dataset and the order of p, d and q.

- Seasonal Auto Regressive integrated moving average (SARIMA): SARIMA model is just one step different from the ARIMA model and that is it takes seasonality into account for the model. The model is SARIMAX imported from the library statsmodels.tsa.statespace.sarimax. when fitting the model, we need to provide training dataset, the order of p, d, and q and also seasonal order.

- Seasonal Auto Regressive integrated moving average with Exogenous variables (SARIMAX): it is just similar model like SARIMA except for the fact that it

takes external variables into consideration while making a prediction. While fitting the model we need to tell the exogenous values along with training dataset, order of p, d and q and seasonal order.

- Holt-Winter's Exponential Smoothing (HWES): this model is used for forecasting time series that has both a trend and a seasonal variation. This model is imported from the library statsmodels.tsa.holtwinters and while fitting the model, it asks for a training dataset, a trend that can either be additive or multiplicative and seasonality which similar to trend can be additive or multiplicative. The trend and seasonality can be determined by looking at seasonality graph.

5 Results

The models have been trained for nine years of data to predict the last year's outcome. There are six algorithms used in the research as mentioned before and all of them have been worked for each of the stock mentioned above. The frequency for the model is business days, as the stock exchange markets only open on the business days, and they are approximately 250 business days, meaning the model is tested for 250 days.

5.1 Comparing resulting graphs for test data and the predicted data for different stocks

There are five stocks in total and all the algorithms' predicted and actual graphs will be compared for the results, and find what model closely follows the trend of the actual graph.

HDFC: HDFC is a financial corporation stock from the Nifty 50, National Stock exchange, India. The six models used showed varied trend to what the actual trend was. There are all six graphs along with the actual trend of the model, which can be used to see what the actual result for HDFC stock was. Auto Regression, ARIMA and Holt Winters Exponential Smoothing, move in a straight line, but when observed closely, they show a little bit of deviation, but cannot be used for prediction. Next model that starts with a higher deviation but later on tends to go towards a straight line is the SARIMA model and like the other three models, this cannot be good for prediction. The most optimal models for the HDFC stock are ARIMAX and

SARIMAX, which closely follows the actual trend, but giving more higher positive as well as negative peaks, which could be a sign that the model works but needs more evaluation in order to find the best one.

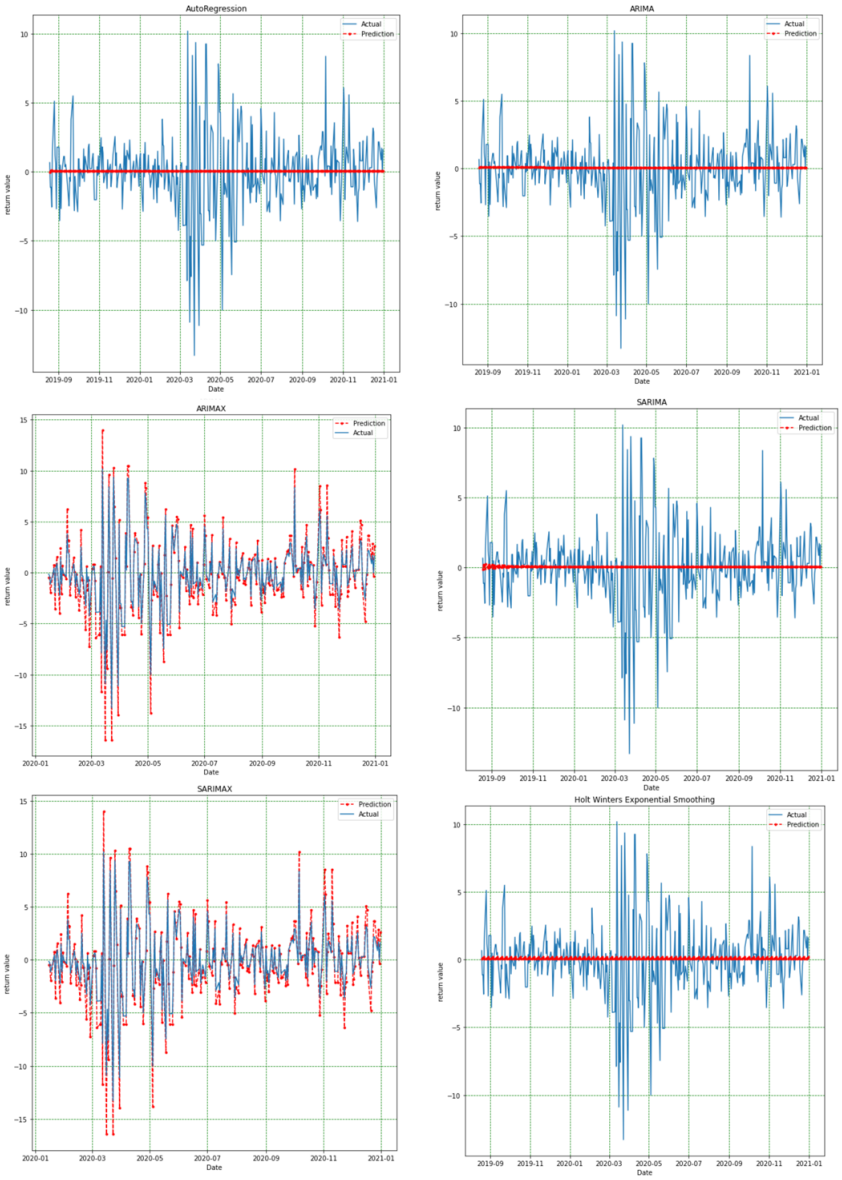

Figure 2: HDFC Stock Prediction with different algorithms

Infosys: Infosys is a stock for an information technology company. The six models used showed varied trend to what the actual trend was. There are all six graphs along with the actual trend of the model, which can be used to see what the actual result for Infosys stock was. when looking at the results for these models, it can be clearly seen that Auto Regression, ARIMA and SARIMA show little to none deviation and work in a straight line and making these models extremely unreliable to work upon. Same is the case with Holt Winter's Exponential Smoothing, but with a higher deviation. The

best working models are ARIMAX and SARIMAX, which show similar kind of trend and needs to be evaluated in order to find the more optimum one. both of these models are not able to reach the actual value's highest peaks.

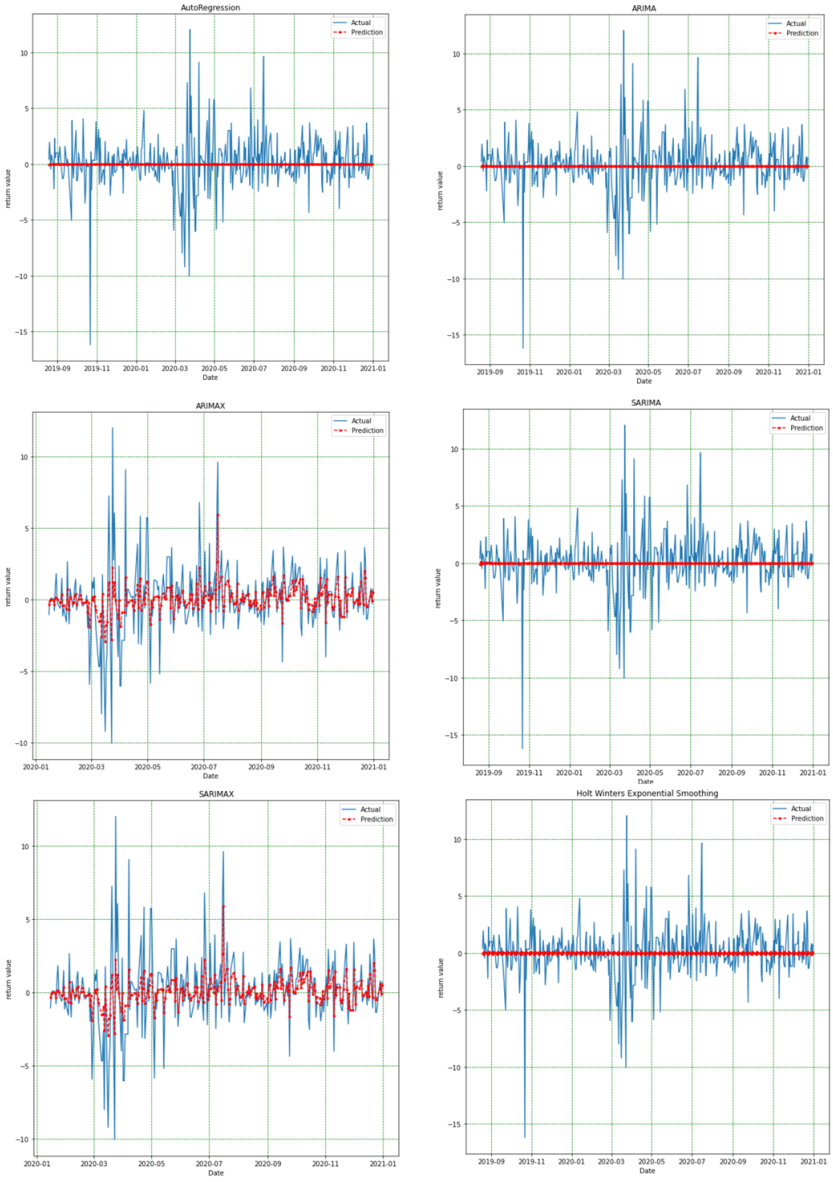

Figure 3: Infosys Stock Prediction with different algorithms

ICICI Bank: ICICI Bank is a stock for a Banking company. The six models used showed varied trend to what the actual trend was. There are all six graphs along with the actual trend of the model, which can be used to see what the actual result for ICICI Bank stock was. Auto regression model shows the least amount of deviation and goes straight meaning can be very unreliable during prediction. Same is the case with ARIMA, SARIMA and Holt Winters Exponential Smoothing, and each of these show a little more deviation than the preceding model. the best fitting model for ICICI

Bank stock are ARIMAX and SARIMAX, which fits the model best, and can be good for predicting positive returns, but can sometimes not be able to predict the negative returns. Since the graphs are similar, it would be a better to evaluate these models before selecting the best model.

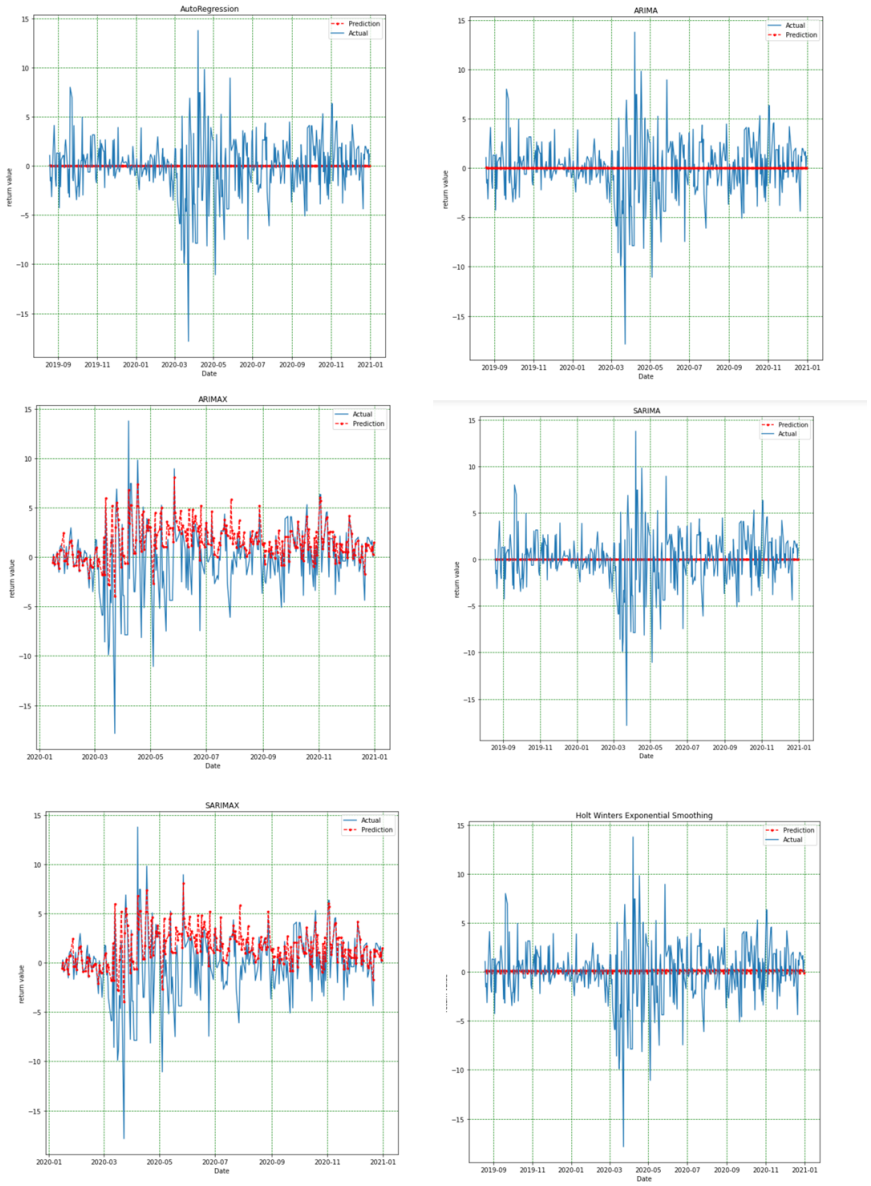

Figure 4: ICICI Bank Stock Prediction with different algorithms

Hindustan Unilever: Hindustan Unilever is a consumer goods company. The six models used showed varied trend to what the actual trend was. There are all six graphs along with the actual trend of the model, which can be used to see what the actual result for Hindustan Unilever stock was. let's start with the amount of deviation that the graphs are showing. The least deviation is shown by Auto Regression, which is almost a straight line and thus cannot be good for prediction. The next two models that show

similar kind of trend are ARIMA and SARIMA, and they show a very interesting pattern if zoomed in and deviates into a little slopier line upwards. The highest deviation is shown by Holt Winters Exponential Smoothing, but the trend continues to be in a straight path. The best fitting models are ARIMAX and SARIMAX. The ARIMAX model fits more than the SARIMAX model. by the looks of the graph, it appears that the ARIMAX model shows a better trend than SARIMAX, except at one occasion when it deviates too negative downwards, but further evaluation will show which one is actually performing better.

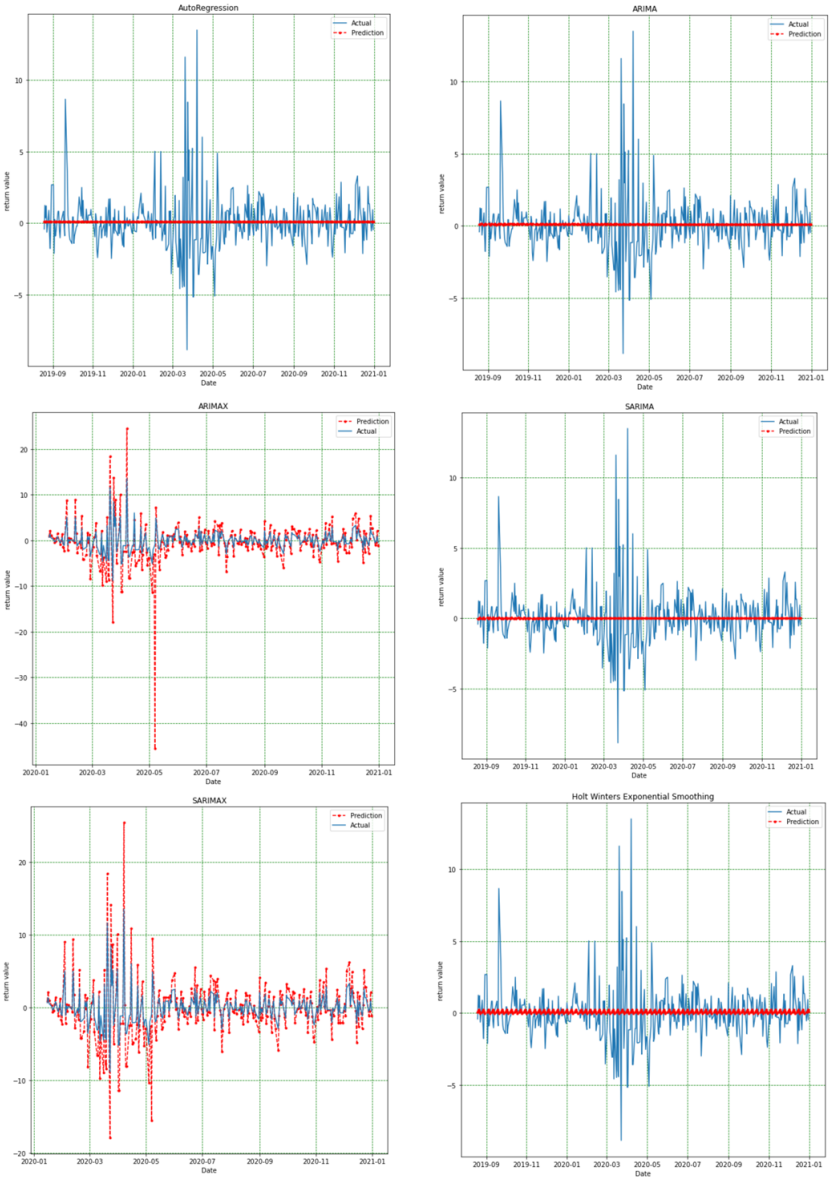

Figure 5: Hindustan Unilever Stock Prediction with different algorithms

Tata Consultancy Services: Tata Consultancy Services, as the name suggests is a consultancy firm. The six models used showed varied trend to what the actual trend

was. There are all six graphs along with the actual trend of the model, which can be used to see what the actual result for TCS stock was. the autoregression model shows only a little bit of deviation in a similar trend, even though when the values were checked, it showed that all the values were different, but only at third or fourth decimal point. Similar kind of trend can be seen in the Holt Winter's Exponential smoothing, but with a lot more deviation than the other two models above. The SARIMA model shows an upward kind of trend which means that the stock prices will increase but there is little to none deviation and would be unreliable in case of market crash. The best models are the ARIMAX and SARIMAX, which follows closely to the actual trend. Further evaluation would show which one works much better than the other.

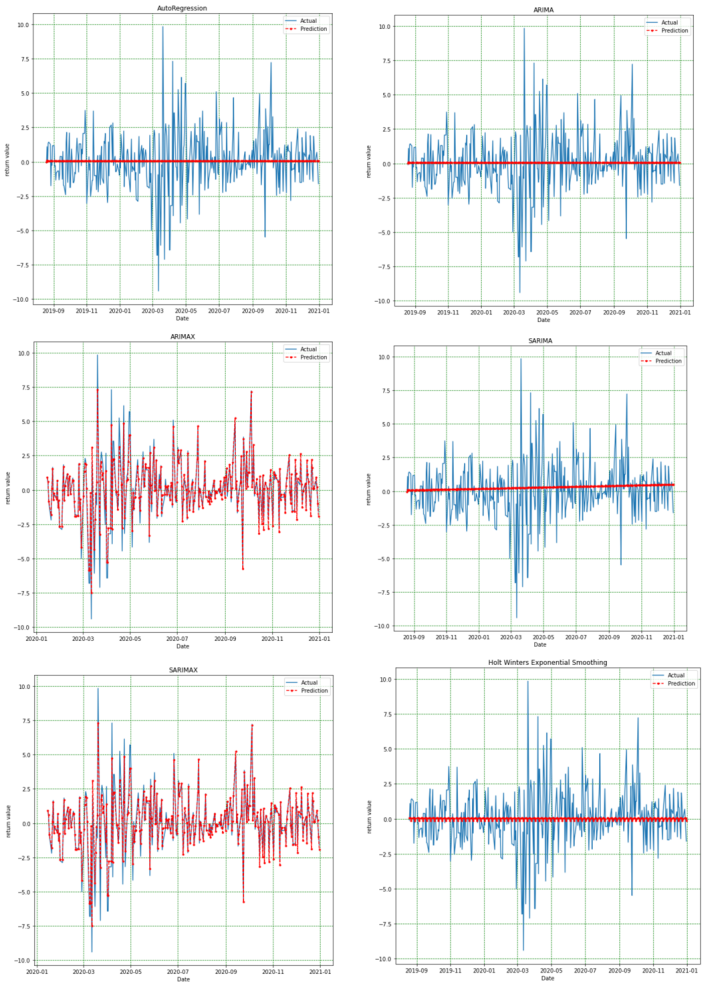

Figure 6: TCS Stock Prediction with different algorithms

6 Evaluation

There are a number of ways to evaluate the performance of machine learning models like accuracy, confusion matrix, F1 score, et cetera. But those methods cannot be used to evaluate the time series models, but can be evaluated by methods like mean absolute error, root mean squared errored, et cetera.

6.1 evaluating techniques used in the paper

For evaluating the financial time series, I have used three most commonly used methods, i.e., mean absolute error, root mean squared error, and R2 score each of which I will be discussing succinctly so as to understand what the results mean.

- Mean absolute error (MAE): the magnitude of difference between prediction and the actual value is mean absolute error. It can be easily understood quantifiable measurement of errors [4]. Even though it can range from 0 to infinity, Lower the value of mean absolute error is, the better the model is working. The mathematical formula for mean absolute error is
  MAE= (Σi=1n|predicted value-actual value|)/n
  Where n is the total number of data points.

- Root mean squared error (RMSE): root mean squared error is, as the name suggests, basically the square root of the mean of the square of all the error. It is extremely helpful to have a single number to judge a model's performance [5]. Same as MAE, RMSE can range from 0 to infinity and lower the value, the better the model performs.
  RMSE= [Σi=1n (predicted value-actual value)2/n]
  Where n is the total number of data points.

- R2 score: R2 is the proportion of variance in the dependent variable that can be explained by the independent variable meaning it is a measure how well the model fits the data. The values range from 0 to 100 and higher the value is, the better the model fits. If the score is 100 percent, it means the dataset is perfectly correlated [11]. And if the score is negative or bigger than 100, then it means that the model does not fit the prediction and should be rejected. The mathematical formula is
  R2=1- (predicted value/initial value)

These criteria will help understand what models are optimal for what stock, which we will see in the next part.

6.2 evaluating stocks and the algorithms used

As mentioned before, this paper consists of five stocks and six machine learning algorithms and each of them will be evaluated using the techniques discussed above.

- HDFC: Here is a table below to evaluate which machine learning algorithm is working best for HDFC stock. For mean absolute error, the value should be as low as possible and by following that rule, ARIMAX is leading the way with the mean absolute error of 0.7387. the next evaluating criteria is root mean squared error and unlike MAE, RMSE is the best for SARIMAX model as the value is the lowest for ARIMAX model. for R2 score, the highest value is 88.28percent, which is of ARIMAX model. this means that the best performing model for the stock HDFC is ARIMAX, since two of the three criteria are in favour of ARIMAX.

Table 3: HDFC stock evaluation

	Mean Absolute Error	Root Mean Squared error	R2 score
Auto Regression	1.8387	2.7466	negative value
ARIMA	1.8383	2.7465	negative value
SARIMA	1.8398	2.7484	negative value
ARIMAX	0.7387	1.0684	88.29 percent
SARIMAX	1.0687	0.7388	88.28 percent
HWES	1.8409	2.7436	0.2 percent
End of Table			

- Infosys: As can be seen from the table below, the Mean Absolute error is the lowest for ARIMAX, that is 1.4171. The RMSE is lowest for SARIMA model, but since it is having a negative R2 score, it has been discarded and chosen the next best value of 1.7900, which is of ARIMAX. Moreover, the highest R2 score which is for ARIMAX, that is 51.67 percent. Overall, the best working model is ARIMAX. here is a table below for comparing different values for evaluation.

Table 4: Infosys Stock Evaluation

	Mean Absolute Error	Root Mean Squared error	R2 score
Auto Regression	1.6024	2.4718	negative value
ARIMA	1.6020	2.4717	negative value
SARIMA	2.4712	1.6016	negative value
ARIMAX	1.1471	1.7900	51.67 percent
SARIMAX	1.1472	1.7902	51.66 percent
HWES	1.6017	2.4719	negative value
End of Table			

- ICICI bank: This is very interesting to see that Mean absolute error, Root Mean Squared Error and R2 score, all three are same for this particular stock. Since SARIMAX model takes extra variables into account such as seasonality, and thus would take more time to work. It would be better to choose ARIMAX, since it would take lesser time.

Table 5: ICICI Bank Stock Evaluation

	Mean Absolute Error	Root Mean Squared error	R2 score
Auto Regression	2.1700	3.1046	negative value
ARIMA	2.1703	3.1047	negative value
SARIMA	2.1726	3.1056	negative value
ARIMAX	1.9182	2.8441	32.53 percent
SARIMAX	1.9182	2.8441	32.53 percent
HWES	2.1619	3.0994	0.25 percent
End of Table			

- Hindustan Unilever: The evaluation and the predicted techniques differ a lot, according to the evaluation, the best working model is Holt Winter's Exponential Smoothing, since MAE, and RMSE are lowest and shows a little bit of correlation with R2 score. But as seen in results, the graph that fits the model best is ARIMAX. This will be further discussed in the discussion.

Table 6: Hindustan Unilever Stock Evaluation

	Mean Absolute Error	Root Mean Squared error	R2 score
Auto Regression	1.2627	1.9784	negative value

Continuation of Table 6			
	Mean Absolute Error	Root Mean Squared error	R2 score
ARIMA	1.2640	1.9786	negative value
SARIMA	1.2554	1.9786	negative value
ARIMAX	1.5563	3.4731	negative value
SARIMAX	1.4158	2.2328	negative value
HWES	1.2542	1.9776	0.01 percent
End of Table			

- Tata Consultancy Services: As can be seen from the table below, the Mean Absolute error is the lowest for SARIMAX, that is 0.3424. the RMSE is lowest for ARIMAX, that can be seen from the table is 05925. the value for R2 score had to be checked up to for three decimal points, since they were same till second decimal point. The R2 score is lowest for SARIMAX model. Overall, the best working model is SARIMAX.

Table 7: TCS Stock Evaluation

	Mean Absolute Error	Root Mean Squared error	R2 score
Auto Regression	1.4590	2.1002	too big value
ARIMA	1.4587	2.1001	too big value

Continuation of Table 7			
	Mean Absolute Error	Root Mean Squared error	R2 score
SARIMA	1.4586	2.1009	negative value
ARIMAX	0.3425	0.5925	93.4738 percent
SARIMAX	0.3424	0.5936	93.4749 percent
HWES	1.4554	2.1009	negative value
End of Table			

7 Discussion

After looking at the resulting graphs of different models and evaluating them gives us a clear understanding of how each of the model works and what kind of results can be expected. Let's discuss each of the stock's performance.

- HDFC: looking at the graph clearly tells that ARIMAX and SARIMAX are the best models as they are fitting the graph properly. The other models just produce a straight line which means they cannot be a good indicator of the prediction and shows that there is a minimal risk of investment and it will remain constant throughout. While evaluating the stock with MAE, RMSE and R2 score, it was found that the MAE was lowest for ARIMAX, RMSE was lowest for SARIMAX and the R2 score was highest for ARIMAX, making ARIMAX best model as it matches two out of three criteria. the interesting thing is that even though HWES model produces a straight line, it still correlates as it gives a R2 score of 0.2percent, while other models gave negative correlation. Auto Regression was found to be the worst model that performs according to graph, but according to evaluation, the worst model is SARIMA.

- Infosys: according to the graphs the best fitting models are ARIMAX and SARIMAX. These two have almost similar trend, which means that there might not be any particular seasonality order in the SARIMAX model. this was contradicted by the evaluation as a MAE, RMSE and R2 score, as the ARIMAX was performing better than SARIMAX, but just by a little bit. The worst performing model was SARIMA in this case as it had the highest MAE, and RMSE and negative R2 score. Not only this, SARIMA also performed worst graphically as well, giving a straight line. This model might be helpful in creating a hybrid model with a deep learning model, but this is a study for the future.

- ICICI Bank: this stock is basically the worst stock of the lot as it is having the highest MAE, RMSE and R2 values, as well as the prediction of any model fits perfectly on the actual test dataset. still the best models are ARIMAX and SARIMAX, since their curve is similar and their RMSE, MAE and R2 score are same. When a situation like this arises, it is always better to choose a model with lesser external factors, and in this case, it is ARIMAX, since it is not having any Seasonality. The interesting factor for this stock is that even though HWES's graph shows a straight-line trend with little deviations, there is a little bit of correlation and gives the R2 score of 0.25 percent. the worst performing model in this case is also SARIMA as it has highest MAE, and RMSE and negative value for R2 score.

- Hindustan Unilever: for this particular stock, the worst performing stock is ARIMAX, as according to evaluation, have the highest MAE, RMSE and negative value for R2 score. But when looking at the graph, SARIMAX model should give the best results and ARIMAX giving close results to this. Instead, during evaluation it was found that HWES model is the best one for this stock as it has the lowest, but these results cannot be seen in the graph. This is a very controversial model that does not work as it is supposed to. There might be some other model that might fit this stock better, which can be further studied in the second iteration of this unit.

- Tata Consultancy Services: for this stock, the graphs are a little bit different than the other stock. Even though the four of the models, i.e., AR, ARIMA, HWES and SARIMA follow a trend that is in straight line, but the SARIMA model gives an upper kind of trend meaning the stock will increase. Moreover, the SARIMA is not the worst performing model in this case. The worst performing model is Auto Regression as it is having highest MAE and R2 score. The RMSE is still the highest for SARIMA. ARIMAX and SARIMAX are the

optimal models for this stock according to the graphs, but when evaluated, it turns out that SARIMAX fits the data much more perfectly than ARIMAX. This is the only stock where the SARIMAX works better than ARIMAX.

8 Conclusion and Future Work

This paper summarises financial time series in a manner that even a layman can understand. There is still some work that is to be done in the future for the financial time series analysis, but for now let's see what the conclusion is.

8.1 Conclusion

From the research we can see that the model that works the best is ARIMAX, which is basically the ARIMA model with exogenous variables. Even though the SARIMAX model should work the best, since it is having seasonality along with external factors. The worst performing models is Auto Regression and for all of the stocks it just predicts a straight line with a little bit of deviation. This model cannot be used for the prediction of stocks. The next model that shows a little bit of deviation and some results are ARIMA and SARIMA. The maximum deviation can be seen by Holt Winter's Exponential Smoothing but still it can predict just a straight-line outcome. It is interesting to see that for the stock Hindustan Unilever, the other R2 score are negative except for HWES, meaning it is the only model that shows a little bit of correlation. if taking RMSE and MAE, into consideration, the SARIMAX model for the TCS stock gives the best fitting. the RMSE and MAE is lowest and the R2 score is 93.4749 percent. But overall, ARIMAX performs better. It is interesting to see that according to graphs the works performing model should be Auto Regression, but when evaluating the SARIMA model gives the highest values of MAE, RMSE and gives negative or big values for R2 score. There is still some work needed to be done for the topic in the future and some of the trends of the model needs to be studied as well.

8.2 Future for Financial Time series

This was the first part of the research project which was to compare and evaluate different machine learning algorithms. The next part of the project that can be done

in SIT724 and beyond that could be

- Working on different sizes of dataset: the first and foremost task could be working on different sizes of the dataset. For this part of the project, I worked on ten years of data, which is basically used in the actual market for the forecast of three years and five years of data. Next step could be working with twenty years, which is extremely big data, five years, which is medium sized dataset and one year, which is the small set of the dataset.
- Working on hybrid and deep learning models: the next part that could be done is working on the deep learning machine learning models such as LSTM or working on hybrid models. Further these models could be used for comparison with models studied in this iteration of the unit
- Creating portfolios: creating portfolios and choosing the best stocks can be a real headache and it would be good idea to create portfolios on the requirements of the client's investment's period of duration.

These all work could be done in the SIT724 or maybe even beyond that.

References

[1] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, An empirical comparison of machine learning models for time series forecasting, Econometric Reviews, 29 (2010), pp. 594–621.

[2] S. Ardabili, A. Mosavi, and A. R. Várkonyi-Kóczy, Advances in machine learning modeling reviewing hybrid and ensemble methods, Lecture Notes in Networks and Systems, (2020), pp. 215–227.

[3] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, Stock price prediction using the arima model, 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, (2014).

[4] C3.AI, Mean absolute error, 2022.

[5] ——, Root mean squared error, 2022.

[6] J. Cao, Z. Li, and J. Li, Financial time series forecasting model based on ceemdan and lstm, Physica A: Statistical Mechanics and its Applications, 519 (2019), pp. 127–139.

[7] J. Cramer, The origins of logistic regression, SSRN Electronic Journal, (2003).

[8] G. W. Crawford and M. C. Fratantoni, Assessing the forecasting performance of regime-switching, arima and garch models of house prices, Real Estate Economics, 31 (2003), pp. 223–243.

[9] D. S. de O. Santos Júnior, J. F. de Oliveira, and P. S. de Mattos Neto, An intelligent hybridization of arima with machine learning models for time series forecasting, Knowledge-Based Systems, 175 (2019), pp. 72–86.

[10] D. Ersan, C. Nishioka, and A. Scherp, Comparison of machine learning methods for financial time series forecasting at the examples of over 10 years of daily and hourly data of dax 30 and samp;p 500, Journal of Computational Social Science, 3 (2019), pp. 103–133.

[11] J. Frost, How to interpret r-squared in regression analysis, 2022.

[12] J. D. Hamilton, Time Series Analysis, Princeton University Press, 2020.

[13] R. J. Hyndman and G. Athanasopoulos, Forecasting, O Texts, 2 ed., 2018.

[14] A. Khazaee Poul, M. Shourian, and H. Ebrahimi, A comparative study of mlr, knn, ann and anfis models with wavelet transform in monthly stream flow prediction, Water Resources Management, 33 (2019), pp. 2907–2923.

[15] T. Kim, S. Sharda, X. Zhou, and R. M. Pendyala, A stepwise interpretable machine learning framework using linear regression (lr) and long short-term memory (lstm): City-wide demand-side prediction of yellow taxi and for-hire vehicle (fhv) service, Transportation Research Part C: Emerging Technologies, 120 (2020), p. 102786.

[16] S. Kouadri, C. B. Pande, B. Panneerselvam, K. N. Moharir, and A. Elbeltagi, Prediction of irrigation groundwater quality parameters using ann, lstm, and mlr models, Environmental Science and Pollution Research, 29 (2021), pp. 21067–21091.

[17] M. Kumar and M. Thenmozhi, Forecasting stock index returns using arima-svm, arima-ann, and arima-random forest hybrid models, International Journal of Banking, Accounting and Finance, 5 (2014), p. 284.

[18] H. Liu, H.-q. Tian, and Y.-f. Li, Comparison of two new arima-ann and arima-kalman hybrid methods for wind speed prediction, Applied Energy, 98 (2012), pp. 415–424.

[19] Q. Ma, Comparison of arima, ann and lstm for stock price prediction, E3S Web of Conferences, 218 (2020), p. 01026.

[20] D. A. Otchere, T. O. Arbi Ganat, R. Gholami, and S. Ridha, Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ann and svm models, Journal of Petroleum Science and Engineering, 200 (2021), p. 108182.

[21] D. I. Pardoe, Regression methods, 2022.

[22] S. S. Poorna, P. M. V. D. S. Baba, G. L. Ramya, P. Poreddy, L. S. Aashritha, G. J. Nair, and S. Renjith, Classification of eeg based control using ann and knn — a comparison, 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), (2016).

[23] X.-Y. Qian, Financial series prediction: Comparison between precision of time series models and machine learning methods, 2017.

[24] R. Rao, nifty-50 stock market data (2000 - 2021), 2022.

[25] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, Financial time series forecasting with deep learning : A systematic literature review: 2005–2019, Applied Soft Computing, 90 (2020).

[26] S. Siami-Namini and A. S. Namin, Forecasting economics and financial time series: Arima vs. lstm, 2018.

[27] P. Soucy and G. Mineau, A simple knn algorithm for text categorization, Proceedings 2001 IEEE International Conference on Data Mining.

[28] R. S. Tsay, Analysis of Financial Time Series, John Wiley and Sons inc, 2 ed., 2006.

[29] V. H. Wentz, J. N. Maciel, J. J. Gimenez Ledesma, and O. H. Ando Junior, Solar irradiance forecasting to short-term pv power: Accuracy comparison of ann and lstm models, Energies, 15 (2022), p. 2457.

[30] C.-C. Young, W.-C. Liu, and W.-L. Hsieh, Predicting the water level fluctuation in an alpine lake using physically based, artificial neural network, and time series forecasting models, Mathematical Problems in Engineering, 2015 (2015), pp. 1–11.

[31] V. Ülke, A. Sahin, and A. Subasi, A comparison of time series and machine learning models for inflation forecasting: empirical evidence from the usa, Neural Computing and Applications, 30 (2016), pp. 1519–1527.