JUNE 12, 2022

INTERIM REPORT

DUNDALK
INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA
DHÚN DEALGAN

ANALYSIS AND PREDICTION OF STOCK PRICES

Supervised by: Dr. David O'Keeffe

NISHANT VERMA
D00251774
MSc Data Analytics

Abstract: *This is the era of big data where stock prices or trend prediction has become more famous than ever before by leveraging the data. We have collected the data of 5 years of stock prices from the Stock market index- S&P 500. The plan is to propose deep learning based predictive model that is capable of predicting the stock price trend in near future. We are focusing on short-term predictions in the study. The study includes the preprocessing of the stock market dataset, optimized feature selection, building of a few types of models and then lastly the comparison and evaluation of the models. The aim is that the proposed system achieves the higher accuracy and will be able to generalize the other stock prices as well which will not be present in the training data.*

Keywords: *Stock market, optimization, prediction, deep learning.*

I. Introduction and Background

Stock market has always been a hot topic for investors and traders as their only goal is multiply the amount of capital they put in and drive the profit out of it. The humanity has witnessed a great advancement in technology sector over the last few decades which has changed the ways of doing businesses all around the world (Rouf et al. 2021). There's one word 'Fintech' which has been doing rounds around the world in last few years comes from two different words i.e., Finance and Technology, this shows how the technology has been impactful in financial industry. Over the last few years, trading and investing in stocks has stolen all the limelight in the financial market when it comes to association with technology. Investors and traders seek applications and tools that would make trading or investing easier, would maximize the profit and lower the risk (Upadhyay and Bandyopadhyay 2012). Stock market prediction has always been an intriguing matter for investors and technologists. In this research, our aim is to develop a prediction model that emphasizes on short-term stock price trend prediction.

Stock market prediction is known to be notoriously difficult task due to its nature of being highly volatile, stochastic, non-linear and having high level of noise (Tan et al. 2007). Stock prices depends on the various mini and macro factors ranging from profits to accounting errors or scandals (Gupta et al. 2013). Gathering all the information that affects stock prices is entirely impossible and also is illegal to make use of sensitive and confidential information for trading stocks (Insider Trading). In our research, we have collected the previous prices of stocks and trying to estimate the future trend based on historical data which is publicly available and can be accessed by anyone. In the world of academic, the stock market prediction is a problem of simple time-series analysis & forecasting in which the historical data is examined and based on that, the next values are estimated (Rouf et al. 2021). The efficient market hypothesis describes that stock prices are significantly dependent on the new information and follows completely random path hence, the stock prices cannot be entirely predicted based on the historical data (Yaes 1989) . This theory stood valid until the researchers made use of advanced technology and showed that the stock prices could be predicted to a certain extent from the historical data. The performance of any mathematical model depends on the on the inputs (features) used to develop it and so is the case with stock prediction system, the better the quality of features, the better the performance would be of the system (Inthachot et al. 2016). In 2003, a team of researchers conducted an experiment by building a neural network to predict the stock prices by keeping their main focus on the volume of shares traded (Wang et al. 2003). The significance of volume turned out to be irrelevant in the dataset that they had used to build the model, the data belonged to the stock index S&P 500 and we are using the data of same stock market for this research. (Ince and Trafalis 2008) aimed at short-term forecasting and applied support vector machine (SVM) to predict the stock prices. The entire idea of their research was to compare the results of multi-layer perceptron and support vector machine for stock predictions and SVM emerged as a winner in this comparison although the results differ by different trading strategies. Parallelly, the researchers having financial background were applying traditional statistical methodologies and signal processing techniques to visualize and predict stock prices. In the past, optimization techniques such as principal component analysis (PCA) has also been used in stock price prediction for fetching the most relevant features in the data engineering in order to attain the maximum possible accuracy (Lin et al. 2009) Over the last few years, researchers have not limit themselves in analyzing the stock prices but also took volume (number of shares) into consideration and have witnessed the change in the prediction's accuracy. This change does not only imply the significance of the volume of share but also redefine the scope and potential of the field (Shih et al. 2019). As the artificial intelligence (AI) methodologies and systems grown over the last few years, the researchers tried to make use those evolved system by combining machine learning and deep learning methods and developed a new metrics which later used as a feature in the process of feature engineering (Liu and Wang 2019). Such types of study would help us in our research for the purpose of feature engineering when implementing machine learning algorithms or while implementing any neural network methodology. Liu, Zhang and Ma developed a convolutional neural network (CNN) and long short-term memory (LSTM) neural network to examine the quantifiable approach in stock market (Liu et al. 2017). The CNN is applied for stock filtration approach, which automatically pulls the relevant features based on the quantifiable data and then proceeds to LSTM to keep the time-series affect intact for better results. Recent research has also introduced a similar kind of hybrid neural network, combining convolutional neural network with bidirectional long short-term memory to analyze and predict the stock market trends (Eapen et al. 2019).

Based on the background study, it is most likely that either a machine learning or neural network-based model would provide us the better and accurate results at predicting stocks near future trends than traditional and general time-series model like ARIMA. However, we'd create a basic time-series ARIMA model which would be considered as a base model and would be used later for comparison with other methodology models.

At the very beginning of the project when only the idea of the project was proposed, I had put forward a few research questions which were quite generic – 1) Which index to be considered for the project, 2) Should we target the single company or all the companies of the stock exchange, 3) What would be the source of data, 4) Is the data collected ethically right for the use in the research, 5) What are all the technologies and methodologies to be used for the research and 6) What would be final metric or final way to evaluate the models and results. Through the background research and study over the last few months, we were able to answer the most of them already like – We are conducting the research on stock market index- Standard and Poor's 500 (S&P 500), tracks the performance of large 500 companies in the United States. From the research, one thing became crystal clear that all the company's stocks data cannot fed at once into any model hence each company's stocks data would be treated and dealt with separately. We chose an open-source data repository platform to collect the data from because of mainly two reasons, one is the platform being open-source hence there's no ethical compliances associated with any entity of the platform, every dataset available on the platform can be utilized by any individual in any possible way without the consent or any permission. However, the right way is to seek permission from the owner of the dataset via email, just to be on the safe side. This answers our research question 3 and 4, related to the source and ethical compliance of the dataset. The plan is to make use of the technologies learned over the course like Python programming for the coding, Jupyter notebook as the editor. Several libraries and frameworks like pandas for data analysis, numpy for numerical operations, matplotlib and seaborn for data visualization, stats model for implementing statistical models, scikit learn to implement machine learning models and many more. We have planned to carry out a new framework algorithm i.e., Long short-term memory neural network which comes from the family of recurring neural network (RNN). RNN-LSTM needs to be learned and implemented for the research. We know various metrics like root mean square error (RMSE), mean square error (MSE), mean absolute error (MAE) to evaluate the performance of the state-of-art model. In the process of the research over the last few months, the research questions have now been revised and have become more focused and precise. The overall objective of the research can be encapsulated by the following research questions-

- What would be the best suited parameters to implement ARIMA efficiently?
- Which power transformation method would be appropriate for the data?
- Does the data follow any seasonality and trend?
- How would the data help to develop neural network based prediction model?
- Which model would have the higher accuracy?
- What would be criteria of choosing the best model?

The general behavior of the stocks are the changing prices at a regular interval of time hence this behavior makes it risky to invest the money. The common view of society towards the stock market is that it is highly uncertain for the purpose of investment, and it cannot be trusted at all with one's money and that's majority of people are not even interested to put their money in trading by calling it the gamble. There are many factors that affects stock market, and it is indeed true that it is impossible to have access to all the factors responsible for stock market movement but the factors like season variance and stable movement of any index can help both new traders as well as active traders in understanding the market and make better informed trades. These types of problems could be well countered by Time-series analysis to see the pattern movement of the market or even forecast the future values to some extent. Financial markets seem to be unpredictable and sometime even illogical. Due to such nature of the market, it gets really difficult to find the patterns in the ill-structured financial data and that's where the data analytics comes into play. In order to find the hidden patterns and model the ill structured data, tools of data analysis like machine learning and time-series algorithms are required. The highly efficient methodologies in this area of field are machine learning and deep learning which are capable of unearthing hidden patterns within the data, predict the future movement and also capable of finding complex co-relations between features which further enhances the accuracy of the results.

For this research, I would be following the CRISP-DM methodology. The CRISP-DM stands for cross industry standard process for data mining, is a standard process that operates as a foundation for data science project. CRISP-DM consists of 6 steps –

1. Business Understanding – What is the requirement of the business?
2. Data Understanding – What all the data we already have, what more do we need and from where do we need?
3. Data Preparation – Is the data organized enough for the modelling or it needs some manipulation and wrangling?
4. Modeling – Selecting various algorithms which fits fine with data and serves the purpose of the business goal.
5. Evaluation – How the model created in last step performing with the test data or unforeseen data?
6. Deployment – How the stakeholders can be benefitted from the model results?

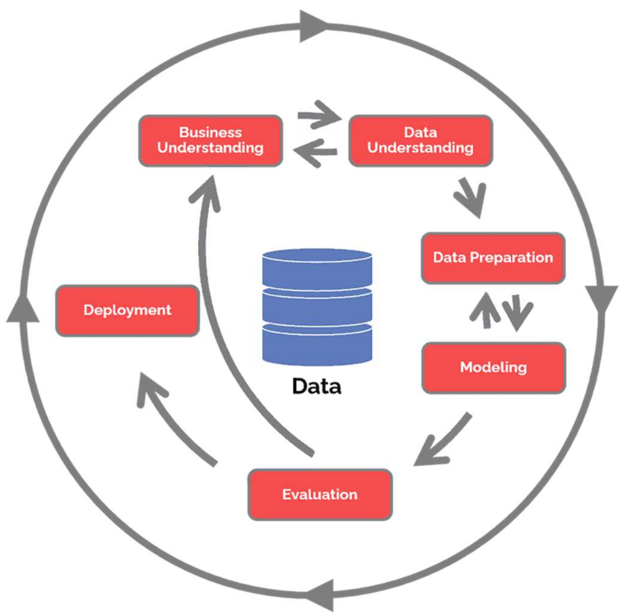


Figure 1:CRISP-DM Diagram. (Source: Wikipedia)

The remaining of this report is structured as follows. “Literature review” that describes relevant work done in the similar field previously. The “Data exploration” that describes the data source and data collection methods to gather data, description of the data, cleaning done in the data and preliminary visualization & analysis. The “proposed future analysis” that describes plan to be followed in the summers in order to reach the end goals that we have set in the form of research questions. The “Conclusion” which is basically an early conclusion based on the research so far and where we stand at the moment, a summary of all the work and findings so far.

II. Literature Review

In this part of the report, we have reviewed the previous related work in the field stock price prediction from both the domains- financial and technology.

In 2000, (Kim and Han 2000) Two researchers proposed an approach of genetic algorithm (GA) in combination with artificial neural networks (ANN) with discretization of features to predict the stock price index. The data which they have used for their study have some technical indicators and also the daily change of direction in Korea stock price index (KOSPI). They have made use of 2928 days of trading data for their research starting from January 1989 to December 1998. They have done feature engineering precisely and have performed feature selection methodology. They have applied a technique for optimization of feature discretization which is similar to the feature reduction technique like PCA. The highlight of their work is the introduction of GA to improve the ANN. There are many limitations to this work due to the evolution in technology sector. For an instance, the learning process of ANN is flawed as the researcher kept the focus only on two factors for optimization. A similar kind of research was conducted in 2016 by Qui and Song in which they have also used ANN to predict the movement of the stock price of the very next day and then they have used GA to enhance the accuracy of the ANN results as an optimization technique (Qiu and Song 2016).

In 2004, Piramuthu carried out detailed research on various features selection methodologies for data extracting applications. For this research, he used diverse dataset that included credit card approval data, traffic on the web data, data of loan defaulters, tam, kiang data and at the end performed a thorough evaluation on how different feature selection methods are enhancing the performance of decision tree. In this study he used several inter-classes and probabilistic distance-based feature selection techniques to study the effectiveness of each technique in processing the input data that would enhance the performance of decision tree(Piramuthu 2004). Author used several inter-class and probabilistic distance-based feature selection methodologies to evaluate their result is the highlight for this research. Although, the algorithm used for evaluation is decision tree which is not capable of handling complex and larger dataset hence we are not sure if the feature selection methodology would perform efficiently when exposed to a larger dataset with a complex model.

In 2005, Hassan and Nath proposed an approach of Hidden Markov Models (HMM) for predicting the stock price of inter-related stock indices. They have narrowed down the features to 4 i.e., opening price of the stock, closing price of the stock, highest price reached in a day and the lowest price in a day of the stock. They performed this study on the stock prices of four different airlines. The study was conducted on only 2 years of data of stock prices for training and testing (Hassan and Nath 2005). The best part of the HMM approach is that it does not require a prior experience in modelling and is a straightforward simple approach. The study fails to generalize the model and is more sensitive to the airlines data hence we cannot be sure if the same approach would be fine with more complex and bigger data. Another limitation of the study would be the size of the data that they have used. The overall dataset isn't so big hence the size of data over which the evaluation has been performed gets smaller. Such a study could easily be exploited in comparison work with relevant studies in the field of stock price predictions.

In 2018, an integrated approach of Rough set (RS) and Wavelet Neural network (WNN) was introduced to enhance the stock price trend prediction capability. In this study, the author first introduced RS as an optimization technique to reduce the features and later the RS was brought in again to decide the structure of neural network i.e., WNN in this study. They made use of quite diverse dataset consisting of several stock indices – Dow Jones Index (USA), Nikkei 225 Index (Japan), All Ordinaries Index (Australia), CSI 300 Index (China) and one the last SSE Composite Index (China). Since the author made use of data of various stock markets around the world, the result was good enough with generalization upon evaluation. It was indeed a great idea of using Rough Set as an optimization technique before processing as it brought down the complexity level in computation (Lei 2018). The authors only emphasized on the parameter manipulation and have not discussed anything about the weakness of the study. We observed that since the model was built on the stock market indices, we cannot be sure if the results would be the same if particular stocks data would be fed in the model.

In 2009, Lee had developed a stock price trend predictive model based on an algorithm Support Vector Machine (SVM). He had used a hybrid feature selection methodology with SVM to predict the stock price trend. The hybrid feature selection technique is named as F-score and Support Sequential Forward Search (F_SSFS), has the capabilities of filter and wrapper techniques to pull out the most subset of most optimum features from the original set of features. In terms evaluation, he compared the performance of the SVM based predictive model in combination with F_SSFS optimization technique with the back propagation neural network (BPNN). He used BPNN for evaluation along with three common feature selection techniques – Information Gain, Correlation based feature selection through paired T-test and Symmetrical Uncertainty. The dataset used for this study is a sample subset of NASDAQ Index from Taiwan Economic Journal Database (TEJD). He performed parameter hyper tuning of kernel function value of SVM by carrying out the grid search methodology which is said to be most expensive function in terms of resources, time complexity and computation power. In the study, the results from the SVM outperform the results from BPNN (Lee 2009). The only limitation found in the study that SVM was only compared with BPNN but not any other machine learning algorithm.

In 2019, Cont and Sirignano proposed a deep learning system built on financial markets' universal feature set. The dataset used for the study contained the records of buying and selling of all the transactions and cancellations of market orders for almost a thousand NASDAQ stock via the database having market order records of the stock exchange. The neural network is comprised of 3 different layers with Long short-term memory (LSTM) units and a feed forward layer with Rectified Linear Units (ReLUs), along with the Stochastic Gradient Descent algorithm (SGD) as an optimization mechanism (Sirignano and Cont 2019). They were able to produce a generalized predictive model which was able to cover the stocks outside the training data as well. They have mentioned a lot of benefits of using the universal model, but it must have been an expensive task to train the model. Due to unclear code of deep learning, it is ambiguous if the irrelevant features were dropped off before feeding the data into the model. Authors have acknowledged a limitation in the paper that it would have been better to carry out a feature selection methodology before training the data as it would have reduced the resource constraints, space complexity and computational power.

In 2011, Ni et al. proposed stock price trend predictive model that was based on Support Vector Machine (SVM). The main idea of the study was how the optimized feature selection can make the difference in the results of predictive model hence they came up with an approach of fractal feature selection as an optimization technique. The dataset that they used had 19 features as technical indicators from Shanghai Stock Exchange Composite Index (SSECI). Prior training the model with data, they made sure that the data processed and ready for training by carrying out feature selection techniques. They performed grid search for finding the best suited parameters, which is K-Fold cross validation. Moreover, the evaluation of feature selection methodologies is also quite broad. In this study, they kept their main focus on the technical indicators came in the dataset, they did not pay attention to any factors related to financial domain as they also have mentioned in their conclusion (Ni et al. 2011).

In 2018, McNally et al. proposed a recurring neural network (RNN) – Long short-term memory (LSTM) solution with a goal of predicting the direction of bitcoin price in USD. The data used in this study was taken from the Bitcoin Price Index. The dataset included the prices of Bitcoin of each day ranging from 19th August to 2013 to 19th July 2016. They used Boruta Algorithm for a specific purpose of feature engineering as an optimization technique. Boruta algorithm works in a similar way to the random forest classifier. In this study, they performed hyper-tuning for parameters of LSTM by making use of Bayesian algorithm. They also used various different types of optimization methods to enhance the performance of neural networks (McNally et al. 2018). The downside of their work was overfitting. The problem of bitcoin price prediction is very similar to the problem of stock price prediction. The difference is that bitcoin market is more dynamin and uncertain in nature which makes it more complex when it comes to predicting. The authors couldn't eliminate the noise and levels present in the data. However, the optimization techniques used in the study and feature engineering part of the study are the strengths of this research.

In 2018, Weng et al proposed a short-term stock price prediction system by making use of ensemble methods. They used multiple sets of data for the study, sourced 5 datasets. They pulled the required datasets from 3 open-source APIs and a package from R named TTR. For the study, they used four machine learning approaches – Support Vector Regression Ensemble (SVRE), AdaBoost with unpruned regression trees as a base model (BRT), Random Forest with unpruned regression trees as a base model (RFR) and Neural Network Regression Ensemble (NNRE). A broad study of ensemble methods has been put forward in this research with a prime focus on short-term stock price prediction. Author did a comprehensive review of the previous related work and precisely chose 8 technical indicators to perform evaluation on their 5 datasets. The strength of their work is that they developed a system for the investor using R programming in which user need not to feed their own data but rather APIs in place will do the job by pulling the data from the online source (Weng et al. 2018). Nevertheless, there are a few limitations to this work. The system can only be able to predict from 1 to 10 days, it won't work for investors looking for return in 2 weeks or intra-day traders which buy and sell the shares the same day. Also, this system was built only on 20 specific US stocks hence it might fail in generalizing the other market's stocks.

In this literature review, we found many methodologies relevant to our dataset, optimization techniques that can be utilized for feature engineering and various evaluation methodologies. From such a vast survey, it is evident that researchers do not focus much on preprocessing of the data and related mechanisms. Neither the researcher attempted to build such a mechanism that facilitates the preprocessing of the data. From technical domain, researchers are highly likely to concentrate on building predictive model with higher accuracy. When it comes to feature selection, I have a common trend where the researchers first list all the features used in the previous works, then apply the optimization algorithm and then based on the results from algorithm, they choose the best voted features. This concludes the summary of feature engineering and predictive modelling from the survey.

III. Data Exploration

The data is sourced from an open-source platform- Kaggle and requires no permission to use or manipulate the data for any purpose. We have taken the data of all the stocks that belongs to a US stock market index – Standard & Poor 500 (S&P 500) starting from 8th February 2013 to 7th February 2018. The data consists of a few features as explained below in detail-

Table 1:Data Description

S.No.	Feature	Data Type	Description
1	Date	Date	The date of the day
2	Open	Float	The opening price of the stock for the day
3	High	Float	The highest price of the stock for the day
4	Low	Float	The lowest price of the stock for the day
5	Close	Float	The closing price of the stock for the day
6	Volume	Float	The number of shares traded in the day
7	Name	Object	Ticker name of the stock

The overall data contains 619040 records including the stock prices of all the companies listed in the stock market index.

At the very first glance of the data, it contains some of the missing values in open, high and low hence those values need to be imputed. There is no way possible that we can have those records removed as we are dealing with a time series problem here and records of each date of the entire date range is mandatory in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 619040 entries, 0 to 619039
Data columns (total 7 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   date    619040 non-null  object
 1   open    619029 non-null  float64
 2   high    619032 non-null  float64
 3   low     619032 non-null  float64
 4   close   619040 non-null  float64
 5   volume  619040 non-null  int64
 6   Name    619040 non-null  object
dtypes: float64(4), int64(1), object(2)
memory usage: 33.1+ MB
```

Figure 2:Data Information

We were also able to see if the regular are available in the data or not. In time-series analysis and forecasting, one must have the records at a regular frequency. In our problem, we need to have stock prices of every day within our date range.

	date	open	high	low	close	volume	Name
0	2013-02-08	15.07	15.12	14.63	14.75	8407500	AAL
1	2013-02-11	14.89	15.01	14.26	14.46	8882000	AAL
2	2013-02-12	14.45	14.51	14.10	14.27	8126000	AAL
3	2013-02-13	14.30	14.94	14.25	14.66	10259500	AAL
4	2013-02-14	14.94	14.96	13.16	13.99	31879900	AAL

Figure 3:First 5 records of Dataset

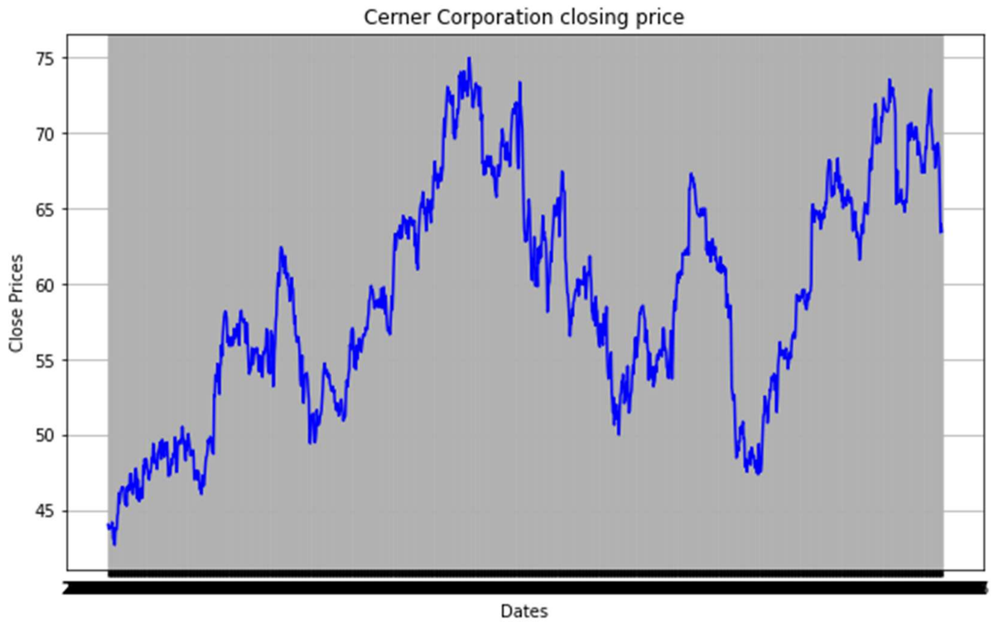
In figure 3, the very first date is 8th February 2013, and the second date is 11th February 2013. In both successive dates, 2 dates are missing. There could be many such instances in the entire dataset. To deal with such a problem, we'd be using Up Sampling and Interpolation methodology to fill these gaps in the data.

Let us visualize the stock price (Close) fluctuations of a few stocks available in our dataset below-

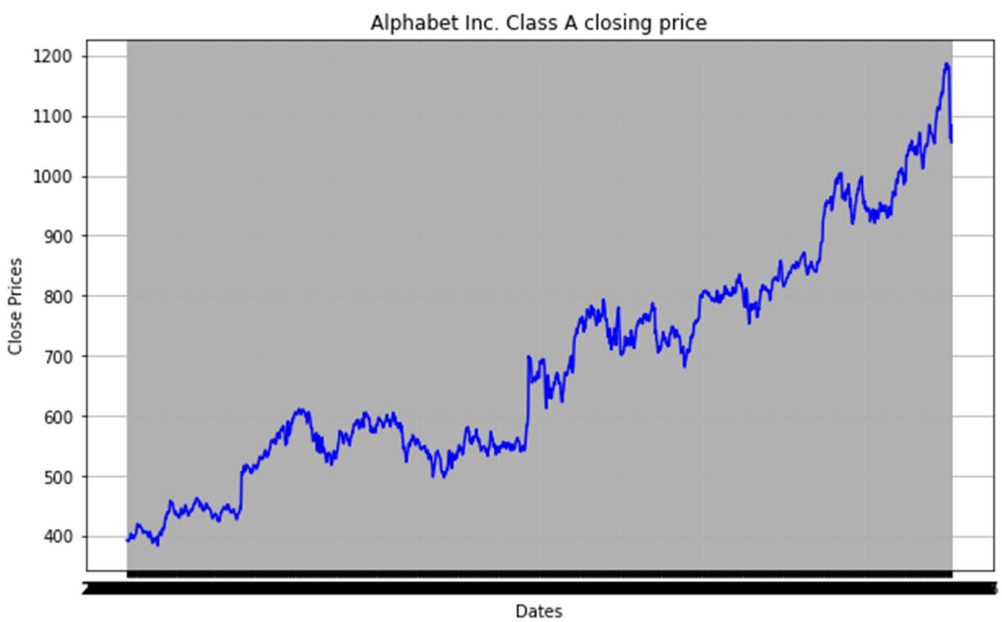
1. AAL (American Airline Group)

2. AAPL (Apple Inc)

3. CERN (Cerner Corp.)



4. GOOGL (Alphabet Inc. Class A)



Through visualisation of all these stocks, it is evident that these stock prices are carrying trend and levels. We need to perform deeper analyse to confirm the presence of seasonality and white noise. Although, through the visualisation so far, we can say that the data is not stationary and is not fit for time-series modelling at the moment. We further need to perform data cleaning to make the data stationary in order to build a time-series model.

Through the literature review, we got know how to deal with our data. We have the data of all the stocks in our dataset hence we need to segregate the data stocks wise and must deal with each stock at a time. There is no way possible that all the data can fed into either a machine learning or deep learning or a time-series model hence each stock must be treated, trained and tested separately.

IV. Proposed Future Analysis

We were able to understand problem so far through a thorough study of previous and related work and have reached to a point where we can carve the path for the upcoming summer to meet desired goal of this research. The following steps could be a potential direction to reach our end goals-

1. Cleaning of data to make it stationary, suitable for time-series modelling.
2. Prepare an ARIMA model as a base model.
3. Perform a grid search or Auto-ARIMA to get the best parameters of ARIMA model.
4. Perform various data transformations to witness any change in the ARIMA model.
5. Learn and implement deep learning model – Recurring neural network and long short-term memory algorithm.
6. Compare the results of advance model with my base model.
7. Evaluation of the model via metric system.
8. Finding the best metric system that fits with the problem statement.

V. Conclusion

This is the preliminary conclusion that states the clarity over the study and progress of the research. Through the research so far, I was able to understand that the tools for time-series analysis and relevant knowledge of the subject would be the right direction to follow in the research. Later I got to know how to deal with such a huge data – Dividing the data according to the stocks and deal with each stock separately. Built a common time-series model ARIMA and set it a base model and later develop an advance model. Compare the results of the advance model with the base model and decide which one is more accurate in terms of bringing the returns.

VI. References

Eapen, J., Bein, D. and Verma, A. (2019). Novel deep learning model with CNN and bi-directional LSTM for improved stock market index prediction. In: *2019 IEEE 9th Annual Computing and Communication Workshop and Conference, CCWC 2019*.

Gupta, R., Garg, N. and Singh, S. (2013). Stock market prediction accuracy analysis using kappa measure. In: *Proceedings - 2013 International Conference on Communication Systems and Network Technologies, CSNT 2013*.

Hassan, M.R. and Nath, B. (2005). Stock market forecasting using Hidden Markov Model: A new approach. In: *Proceedings - 5th International Conference on Intelligent Systems Design and Applications 2005, ISDA '05*.

Ince, H. and Trafalis, T.B. (2008). Short term forecasting with support vector machines and application to stock price prediction. *International Journal of General Systems*, 37(6).

Inthachot, M., Boonjing, V. and Intakosum, S. (2016). Artificial Neural Network and Genetic Algorithm Hybrid Intelligence for Predicting Thai Stock Price Index Trend. *Computational Intelligence and Neuroscience*, 2016.

Kim, K.J. and Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications*, 19(2).

Lee, M.C. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, 36(8).

Lei, L. (2018). Wavelet Neural Network Prediction Method of Stock Price Trend Based on Rough Set Attribute Reduction. *Applied Soft Computing Journal*, 62.

Lin, X., Yang, Z. and Song, Y. (2009). Short-term stock price prediction based on echo state networks. *Expert Systems with Applications* [online], 36(3), pp.7313–7317.

Liu, G. and Wang, X. (2019). A new metric for individual stock trend prediction. *Engineering Applications of Artificial Intelligence*, 82.

Liu, S., Zhang, C. and Ma, J. (2017). CNN-LSTM Neural Network Model for Quantitative Strategy Analysis in Stock Markets. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

McNally, S., Roche, J. and Caton, S. (2018). Predicting the Price of Bitcoin Using Machine Learning. In: *Proceedings - 26th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, PDP 2018*.

Ni, L.P., Ni, Z.W. and Gao, Y.Z. (2011). Stock trend prediction based on fractal feature selection and support vector machine. *Expert Systems with Applications*, 38(5).

Piramuthu, S. (2004). Evaluating feature selection methods for learning in data mining applications. *European Journal of Operational Research*, 156(2).

Qiu, M. and Song, Y. (2016). Predicting the direction of stock market index movement using an optimized artificial neural network model. *PLoS ONE*, 11(5).

Rouf, N., Bashir Malik, M., Arif, T., Sharma, S., Singh, S., Aich, S. and Kim, H.-C. (2021). electronics Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. , 2021. Available from: https://doi.org/10.3390/electronics10212717.

Shih, D.H., Hsu, H.L. and Shih, P.Y. (2019). A study of early warning system in volume burst risk assessment of stock with big data platform. In: *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2019*.

Sirignano, J. and Cont, R. (2019). Universal features of price formation in financial markets: perspectives from deep learning. *Quantitative Finance*, 19(9).

Tan, T.Z., Quek, C. and Ng, G.S. (2007). Biological brain-inspired genetic complementary learning for stock market and bank failure prediction. *Computational Intelligence*, 23(2).

Upadhyay, A. and Bandyopadhyay, G. (2012). Forecasting Stock Performance in Indian Market using Multinomial Logistic Regression. *Journal of Business Studies Quarterly*, 3(3).

Wang, X., Phua, P.K.H. and Lin, W. (2003). Stock Market Prediction Using Neural Networks: Does Trading Volume Help in Short-term Prediction?. In: *Proceedings of the International Joint Conference on Neural Networks*.

Weng, B., Lu, L., Wang, X., Megahed, F.M. and Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. *Expert Systems with Applications*, 112.

Yaes, R.J. (1989). The Efficient Market Hypothesis. *Science*, 244(4911).