Using Artificial Intelligence in Stock Market Prediction

Muhammad Asif Khan
S19004470
Wrexham Glyndŵr University
Wales, UK
s19004470@mail.glyndwr.ac.uk

ABSTRACT

In the last few decades, Artificial intelligence has seen a huge growth in various industrial sectors specially in finance where it is being used to meet the ever-increasing demands of customers who want smarter and safer ways to invest, spend and save their money. This project is about exploring the role of Artificial intelligence and its overarching discipline Machine Learning in the field of financial intelligence. As it is a new emerging technology, it has gained a huge interest from both financial experts and academia researchers who are calling it 'the new financial brain'. This research will focus on how to deploy Machine Learning techniques to solve the complex financial problem of stock market share prediction. Due to recent advancement in digital technology and cloud computing, financial data has been generated at an unprecedented rate and it is very noisy, nonstationary, and nonlinear. In the financial world, zillions of calculations need to be done to provide real-time data analysis which is a complex task. By using time series to model financial data, many researchers have tried various techniques from classical Statistics and Computational Mathematics for prediction of financial time, but they are time consuming and inaccurate. There are few studies exploring the idea of using artificial intelligence methodologies in predicting stock market behaviour but what remains to be explored is to apply those ideas in real time to understand their effectiveness. The purpose of this research is to evaluate the prediction of Machine Learning models by using a real-time financial data and to identify the most computational efficient model to generate the trading decisions more effectively [29].

KEYWORDS

Artificial intelligence, machine learning, time series, deep learning, artificial neural network, linear regression, support vector regression, long short-term memory regression

1. INTRODUCTION

1.1 Background

The world financial institutions are generating and accumulating petabytes of data about their customers behaviours. It is commonly known as 'Big Data' and it is very important for banks and financial institutions to understand this data, uncover hidden patterns, and produce accurate models to make important decisions accurately [1]. This study will provide an alternative solution by employing Artificial Intelligence (AI) and its subset Machine Learning (ML) techniques which have revolutionised and transformed the analysis of Big Data [29].

In the last few years, AI based solutions based on various ML algorithms are getting very popular to analyse financial data and make informed decisions instantly. For example, to decide the loan amount for customers based on their live credit scores, detect fraudulent transactions, and provide focussed target marketing to customers [2,29].

1.2 Scope for Research

This study focuses on trading in stock market which is the most popular way of financial investment. In a financial market, investors make profit by selling and buying shares, bonds, precious metals, and foreign exchanges. The focus of this research is to find out how to use AI and ML to predict the share price of a given stock for an investor to gain the maximum profit [29].

The review of past literature reveals that time series is being used to model financial data. According to some researchers, using basic time series and traditional statistical methods to analyse financial data is not the best approach because financial data is very complicated due to non-symmetrical trends and extreme seasonal variations [3].

There is clear research gap in the field of building intelligent computational financial solutions by using ML algorithms to predict stock market index movement because it is relatively a new research domain. However, this research is very important for financial corporations and banks because it enables them to analyse huge amount of data coming out from financial markets, customer transactions, and social media at a record speed. By using intelligent ML financial models, they will be to derive valuable information and offer higher quality of service to their customers at a reduced price [4]. This study also believes that ML models will be helpful for financial policy makers to monitor stock market indices and design effective trading strategies to reduce financial crimes [29].

I am very interested in applied mathematics and computing and this study will help me to understand how to use mathematical models to represent underlying hidden patterns in the given data. I hope that by using ML algorithms and Big Data computational analytics techniques will help me to understand how to model financial data and predict share prices accurately [29].

2. LITERATURE REVIEW

2.1 Introduction to Artificial Intelligence (AI)

In the last two decades, there has been a sharp rise of using computational tools which are based on Artificial Intelligence (AI) and its derivatives such as Machine Learning (ML) and Deep Learning (DL) to provide better predictive performance of stock price [5, 29].

To understand these techniques, a clear understanding of AI and its related fields is very vital to understand subtle differences between each field as shown below in Figure-1.

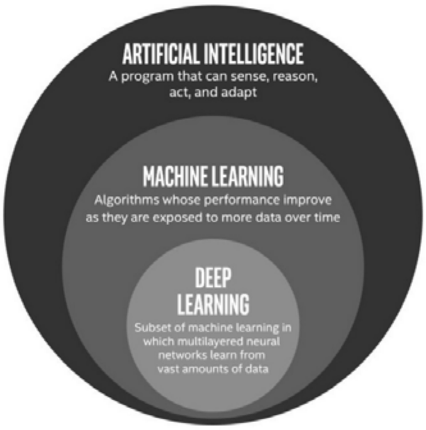

ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

Figure-1

John McCarthy (who is considered as one of the founders of artificial intelligence discipline) defines AI as follows [6],

AI is defined as the science and engineering of the simulation of human intelligence processed by machines especially computer programs.

In the context of this study, AI should be considered as a tool to demonstrate better problem solving and risk management skills than humans.

ML is a branch of AI that builds algorithms by using computer programs to automate learning and make accurate predictions based on the given data without explicitly programmed to do so.

There are three broad categories of ML which are called as supervised, unsupervised, and reinforced learning as shown in Fig-2 below [22, 29].

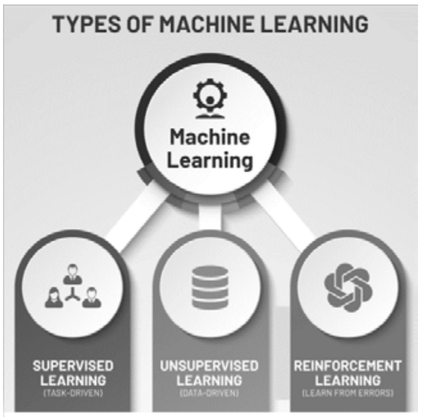


Figure -2

Supervised learning is the type of ML algorithm which is trained on labelled input and output data. In supervised learning, the given data must be structured and should be stored in a machine-readable form such as magnetic disks, optical disks, and barcodes etc. Unsupervised learning deals with working with unlabelled data which is mostly in non-machine-readable form like photographs and hand-written documents. Reinforced learning is based on human learning skills like trial-and-error. It uses a reward system to reinforce favourable outputs and discard non-favourable outcomes [5, 29].

A relatively new field of AI called Deep Learning which is based on simulation of physical structure of human brain by employing a unique algorithm called Artificial Neuron Networks (ANNs) or simply Neural Networks (NNs). A neural network mimics human decision making by adding hidden layers between its input and output data. It takes the input data and then uses its hidden layers to understand the complex relationship between different features of the given data. It assigns a new weight or factor to each hidden layer and when input data passes through each layer, it combines the weight and the input data together to get the desired output [7, 29].

2.2 Using AI to model stock market data

In the last few decades, it has been a big challenge for researchers to design efficient computational models to predict share prices accurately because financial data is highly fluctuating with irregular patterns. Initially, some researchers used linear models based on regression method but discovered that those models failed to provide better results because algorithms based on linear relationships should not be used to represent non-linear stock market data. It raised another issue; there was no straightforward non-linear algorithm to model time series data [8, 29].

Some research questioned the need of predicting model by pushing forward the ‘Efficient Market’ theory. This theory states that stock price of a given company is based on all the relevant market information in the given time. Each time, new information arises, the stock price would corrects itself, so it is perfectly self-efficient and impossible to model [2, 29].

This theory was rejected by Tsibours and Zeidenberg who showed that stock market prices could be modelled by using neural networks. In 1990, Schwartz and Whitcomb designed a stock market pricing model which was based on Artificial Neural Networks (ANNs) and managed to produce predictions with reasonable accuracy. Abiqueng, Hill, O’Connor and Resmus (1994) proved that a multi-layered feed forward neural network could predict stock prices with a better prediction ratio as compared to previous ANN models. Some studies suggested that the efficiency of multi-layered ANNs stock pricing models could be improved by using some advanced time series features such as Auto Regressive Moving Average (ARMA), Vector Auto Regression (VAR), and Bayesian VAR [9, 29].

By using ARMA based neural network, Yim (2002) predicted Brazil stock market prices successfully with an amazing Root Mean Squared Error (RMSE) value of 0.21. Suddenly, Neural Networks (NNs) started to become dominant because they were producing better predicting models. Researchers started to try various ways to improve the efficiency of NNs models. In 2015, Kumar and Thenmozhi conducted some statistical tests like RMSE and Mean Absolute Error (MAE) on S&P 500 data. They used various linear and non-linear ML models with different computing techniques like RF, SVM, LR and ARMA. After comparing each model RMSE and MAE values, they discovered that by combing two different models to create a composite model would produce accurate predictions and better efficiency than a stand-alone ANN model. Lee (2000) produced a new stock pricing predicting model by combining Support Vector Regression (SVR) with statistical F-scores. He used the model on NASDAQ data with 29 technical indices and got a technical accuracy of 80% and produced better results than any single neural network [29].

The idea of combining different ML techniques to develop a better stock market prediction model started to get very popular. Many researchers started to experiment with different combinations. Kim and Gen used a composite model of NNs and Genetic Algorithm (GA) to predict Singapore stock exchange index rates and achieved 82% accuracy to predict both rising and falling prices of different currencies [10, 29].

3. RESEARCH HYPOTHESIS AND QUESTIONS

3.1 Research Hypothesis:

Based on my research and observational study, I will use the following hypotheses:

H0: There is no linear relationship exist between the adjusted closing price (US$) and the given time (days) of a given stock in a financial market.

H1: There is a linear relationship exist between the closing price (US$) and the given time (days) of a given stock in a financial market.

3.2 Research Inferential Question

1. *How to apply artificial intelligence techniques to model the financial data accurately?*

2. *Which is the best Machine Learning (ML) model to predict share prices accurately?*

3 METHODOLOGY

The intent of this study is to apply AI and data mining ML techniques to understand the financial data to develop better and accurate financial data models. The main task is to evaluate the prediction of Machine Learning (ML) models by using a real-time financial data and to identify the most computational efficient ML model to generate the trading decisions more effectively [29].

3.1 Data Collections

In this study, the data comes from on the Standard and Poor's 500 (S&P 500) data which is the one of the most reliable financial benchmarks to track the performance of the top 500 companies listed in the US stock exchange. I am using a Python package called ‘Yahoo Finance’ or in short ‘YFinance’ to access daily financial trading data from December 2000 to December 2021. It will be impractical to look at all 500 companies, so I will be focussing on two major corporations which are Amazon and FedEx.

Amazon is a multinational internet retail giant company which employs 1.6 million people worldwide with a revenue of 386.1 billion US$ (2021) and it focuses on digital streaming, e-commerce, cloud computing and artificial intelligence. Federal Express or ‘FedEx’ is another multinational business which employs more than 500 million people worldwide with a revenue of 22.6 billion US$ (2021) and it is best known for transportation, e-commerce, and air delivery service.

The purpose of this research is to investigate the relationship between two features which are the share price and the given time in days. By using the cause-and-effect relationship, I believe that time is an independent variable because it causes shares prices to change. Therefore, I will use the time which is being measured in days as an independent variable and the dependent variable is the adjustable closed price index of the given stock [29].

Additionally, I want to clarify that I will use ‘adjusted closing price’ instead of ‘closing price’ of a given share because the adjusted closing price is more accurate than the closing price because it indicates not only the cost of shares at the end of the day but also include other factors like stock splits and dividends. Finally, the control variables will be the time frame and type of financial organisations which should be the same for each ML model.

3.2 Method

I have used Python 3.9 in a Jupyter notebook which is web-based interactive development environment of Anaconda with the help of some popular Python libraries for data analysis and ML such as pandas, scikit-learn, statsmodel and TensorFlow [24,25] . A dynamic geometrical software called Autograph is being used to draw and analyse statistical diagrams [26]. For training and testing Deep Learning models, the Google Colaboratory Research program called ‘Colab’ and MATLAB have also been used [27,28].

By using a Python module called ‘YFinance’, the raw dataset for Amazon and FedEx stock prices from Yahoo website was downloaded as comma separated values (csv) format files. Then, this data was loaded in the Juptyer Notebook within Anaconda working environment. By using a Python library called ‘Pandas’, data frame objects were created for testing and training the input data [29].

3.3 Initial Findings

Initially, I noticed that there were few missing values and incorrect data type which are commonly known as NaN (Not a Number) values. I employed mean average to replace NaN values to avoid the problem of raising exceptions while performing statistical calculations. The other pre-processing step was to remove any outliers because they would seriously mislead the training process which might lead to less accurate models and longer training times [7, 29].

My research shows that it is a common practice to convert the raw data into standardised or normalised scales to level up different features for hypothesis testing, but I decided not to use any scaling techniques. The downside of scaling the data is that some ML algorithms like Support Vector Machine (SVM) and Regression could produce wrong prediction as they tend to converge faster on scaled data [12]. Furthermore, normalisation also changes the actual meaning of the required outcome i.e., share price in US$ makes more sense than some random scaled value.

4. DATA ANALYSIS AND RESULTS

During this stage, I am planning to perform different data analysis techniques to understand the data, interpret statistical figures and build accurate ML models. For data analysis, I will use univariate analysis on the adjusted closing share price to look at the spread of the data and identify any hidden patterns. Bivariate analysis is also being used to check if there is a relationship exist between share prices and time in days [19].

4.1 Univariate Data Analysis

For univariate analysis, I am using the key statistical measures like mean, median, mode, range, standard deviation, and Inter-Quarter Range (IQR) to understand the spread and central tendency of the adjusted closing share prices. Some statistical diagrams like frequency table and box plot are being used to understand the distribution of data [7] as shown in Figure 3 and 4 below.

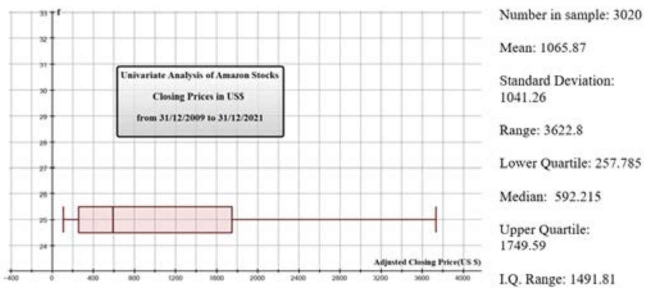


Figure 3: Univariate Analysis of Adjusted Closing Price of Amazon shares from 31/12/2009 to 31/12/2021

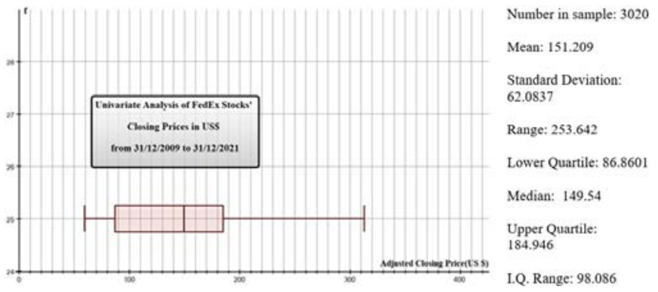


Figure 4: Univariate Analysis of Adjusted Closing Price of FedEx shares from 31/12/2009 to 31/12/2021

Both Figures 3 and 4 shows that median and mean values of adjusted closing price of Amazon shares are higher than FedEx, which means that average price of Amazon shares is more than FedEx. However, Amazon share price is positively skewed which means that there are more values close to higher share prices which indicates that Amazon share prices has increased rapidly in the last decade. On the contrary, FedEx share price shows negative skewed distribution which indicates that higher frequency of its share prices constitutes of lower valued scores which indicates stable share prices [12].

The most significant feature related to my hypothesis is standard deviation or ‘σ’ that is an important metric to calculate risk in investment by looking at how far apart the adjusted closing prices of each share is in the given dataset. It is very clear that Amazon has higher standard deviation (σ =1041.26) which means that its stock prices are volatile and tend to fluctuate very often which causes greater risk to an investor. On the other hand, FedEx has a lower standard deviation (σ =62.0837) which means its share prices are mostly stable and follow a predictable pattern which makes them a safer investment.

4.2 Bivariate Data Analysis

The reason to use bivariate analysis is to understand the hidden relationship between adjustable closed price index of the given stock and the time which is measured in days. Some statistical features like scatter diagrams, correlation coefficient and Spearman's rank coefficient (SRC) are also used to understand the type and strength of correlation. Figures 5 and 6 gives the bivariate analysis of Amazon and FedEx share prices as shown below.

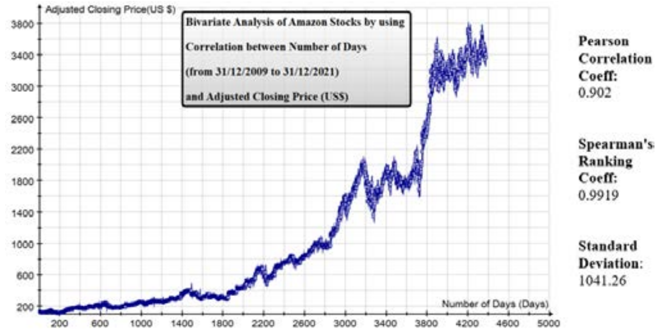


Figure 5: Bivariate Analysis of Adjusted Closing Price of Amazon shares from 31/12/2009 to 31/12/2021

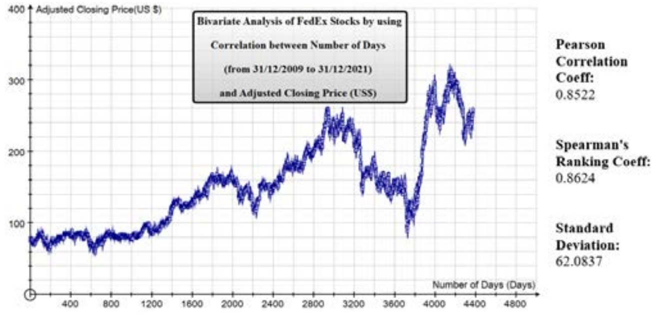


Figure 6: Bivariate Analysis of Adjusted Closing Price of FedEx shares from 31/12/2009 to 31/12/2021

From Figures 5 and 6, it is evident that there is a strong positive correlation between the adjusted closing price (US$) and the given time (days) for both Amazon and FedEx stocks because their Pearson Correlation Coefficients (PCC) have positive values close to 1. However, PCC values will only be effective if there is an assumption that adjusted closing share prices and the time in days are already linearly related.

To overcome this issue, I am assuming that that both the adjusted closing price (US$) and the given time (days) are in monotonic relationship, which means they are not changing at a same constant rate as in linear relationship. As a result, the Spearman Rank Correlation Coefficient (SRCC) is a better choice because it uses ranked values for each variable instead of raw data. SRCC values for both Amazon and FedEx are positive and very close to 1, which means that both share prices and time have a strong positive correlation. Therefore, share price values of both Amazon and FedEx will increase in the future [7].

Based on the above evidence, I am confident that there is a positive correlation between the adjusted closing price (US$) and the given time (days) of both Amazon and FedEx stocks. Now, I need to find out the nature of that relationship, build a mathematical model, and then predict the price of a given stock by using artificial intelligence techniques like Machine Learning (ML) as explained in the next section.

5 USING MACHINE LEARNING MODELS WITH EVALUATIONS

This research uses Machine Learning (ML) to build mathematical models to understand relationships and dependencies between the target output (adjusted closing price in US$) and the input feature (the given time in days) by using the daily financial transaction trading data from December 2000 to December 2021.

As a standard practice, the given data has been split into training, development, and testing sets to build an accurate model. This approach will also help to avoid underfitting or overfitting of the data which may lead to biasing and complexity. The training data set will use 80% of the given data which will be fed into the given ML model to learn about different features. The next data set is the development one which is being used to evaluate the training model by cross checking errors between training data and actual data values. Finally, the last one is the testing set which uses the remaining 20% of the given data and it will calculate the efficiency of a ML model by using various statistical calculations such as R-squared (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAE). The above calculation will be presented in performance matrix to summarise the performance and evaluate the accuracy of share price predictions for each model [9, 29].

For this study, only regression-based ML techniques will be used because it is dealing with predicting share prices accurately which is a regression problem. Additionally, due to time-constraint issues, I have decided to use only Amazon shares prices.

5.1 Using Linear Regression (LR) Model

The Linear Regression or LR is the most popular supervised ML technique which is based on a statistical model that assumes a linear relationship between the input variables and the predicted or output variable. I am using autocorrection and residual analysis to estimate the share price accurately. It is very important to make sure that before training the ML model, all the given data including all inputs(x) and output(y) features are assumed to be linearly correlated; all features are normally distributed; and there should be no outliers in any input features [13]. Figure 7 shows the LR mathematical model of Amazon share prices as shown below.

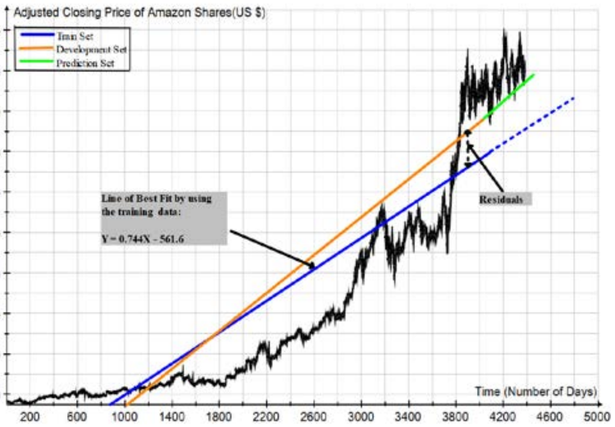


Figure 7: Linear Regression model of Adjusted Closing Price of Amazon shares

During the training phase, a relationship between the input feature ‘X’ which is the given time in days and the output feature ‘Y’ which is the adjusted closing price in US$ is established by drawing a straight line. It is known as regression line and commonly called as the line of best fit (LOBF). It fits closely to the corresponding points of the given training data and represented by a linear equation model in the form of Y= a X + b as shown below:

$$\mathbf{Y = 0.7422X - 561.6}$$

To apply linear regression by using the LOBF to model Amazon share prices is not looking accurate as shown in Figure 7. The LOBF carries big residuals and needs lots of autocorrections to adjust error margins due to higher values of standard deviation. Furthermore, this model lacks accuracy because residuals do not follow a specific pattern. It means that this model is unable to map correct set of values within the given range of training data commonly known as ‘interpolation’ [12]. To fix this problem, I used a 10-fold cross-validation to optimise the linear regression model to produce better predictions as shown by an orange line in Figure 7. Finally, by extending the improved optimised model, some predictions have been made by using the test or prediction set of data which is unknown to our model as shown by a green line in Figure 7.

5.2 Evaluation of Linear Regression (LR) Model

As discussed earlier, the performance of the LR model will be examined by using key statistical calculations as shown in Table I below.

TABLE I. Performance Matrix of Linear Regression Model

Statistical Calculation	R-Squared	Mean Squared Error	Root Mean Squared Error	Mean Absolute Percentage Error (%)
Symbol	**R^2**	**MSE**	**RMSE**	**MAPE**
Training Set	0.29	149189.06	386.25	73.45
Validation Set	0.32	129931.41	360.46	70.29
Testing/Prediction Set	0.23	201502.23	448.89	79.15

This linear regression model is looking very poor. The R Squared values are very low which means that this model does not fit very well in the given data. The higher values of both MSE and RMSE indicates that it is struggling to follow lots of noise and variations in the given financial data. The higher values of MAPE show that the model is overfitting because it is failing to capture many important trends and patterns of both the training and testing data sets.

5.3 Using Support Vector Machines (SVMs) Model

The SVMs are very powerful ML supervised learning methods which can be applied to solve both classification and regression problems. They are also very useful to model non-linear models by using generalised algorithm. Many researchers have used SVM models to analyse financial time series because they are very effective when dealing with data with lots of fluctuations and non-stationary features. I am using the SVM regression model commonly known as SVR because the future price prediction is a regression problem. The SVR uses a linear classification called ‘optimal line’ to separate sets of two class objects. If linear classification is not possible, it maps the given objects into a high-dimensional space by using a phi(φ) function. Then, it uses different decision boundaries commonly known as ‘hyperplane’ to separate objects of two classes and predict the output. A set of complex mathematical functions called ‘Kernel’ is used to transform the given data into the higher dimensional space. I will also use non-linear kernels like Polynomial and Radial Basis Function (RBF) because I am aware that I am handling data set which is full of high fluctuations [13, 29].

The Figure 8 shows the Support Vector Regression or SVR implementation on the given data.

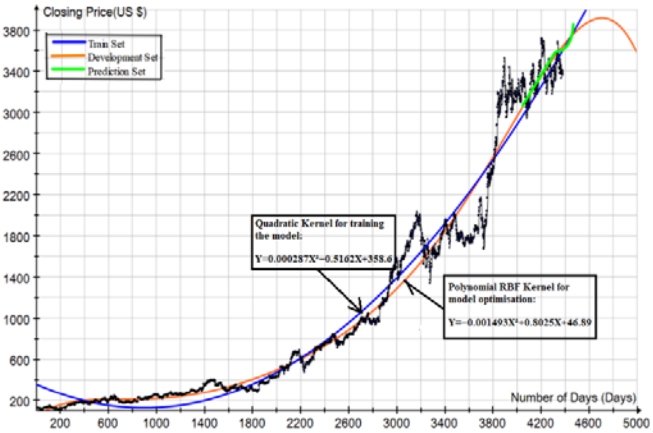


Figure 8: Support Vector Machine (SVM) model of Adjusted Closing Price of Amazon shares

It is clear that that Support Vector Regression Machine model offers better fitting and accuracy as compared to the previous Linear Regression model as shown above. To train the SVM model, a quadratic (non-linear) kernel is being used to work out the decision boundaries of hyperplane to separate each data point. It is shown by a blue curve in Figure 8 with the following equation:

$$\mathbf{Y = 0.000287X^2 - 0.5162X + 358.6}$$

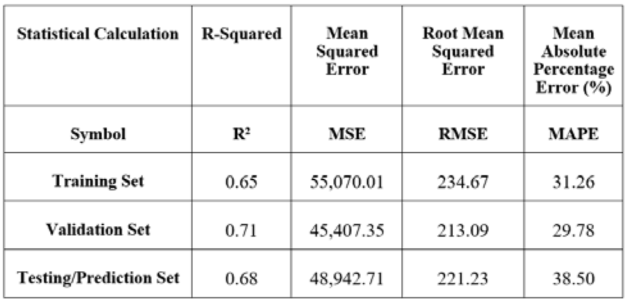
Then, after running few optimisation cycles by using a more versatile non-linear RBF kernel, a new polynomial hyperplane function is produced which can be shown as an orange curve in Figure 8 with the following mathematical model:

$$\mathbf{Y = -\,0.001493X^2 + 0.8025X + 46.89}$$

5.4 Evaluation of SVMs Model

The evaluation of the SVM regression model by using key statistical calculations is shown below in the Table II.

TABLE II. Performance Matrix of SVM Model



Statistical Calculation	R-Squared	Mean Squared Error	Root Mean Squared Error	Mean Absolute Percentage Error (%)
Symbol	R^2	MSE	RMSE	MAPE
Training Set	0.65	55,070.01	234.67	31.26
Validation Set	0.71	45,407.35	213.09	29.78
Testing/Prediction Set	0.68	48,942.71	221.23	38.50

The regression model by using SVM has produced better results than the linear regression. The R Squared values of all data sets are above 0.50 which indicates that the predicted values are closer to the regression line as shown by the green curve in Figure 8. Unfortunately, both MSE and RMSE values are still higher which is reducing the performance of this model. Again, it is very clear from Figure 8 that SVM model is struggling to manage sudden fluctuations in the given data. MAPE values are lower for both training and validation stage which shows that this model performs well on known data. However, during the testing stage, MAPE value increases to 38.50% which shows that it is producing poor results for the unseen data [17].

6 INTRODUCTION TO DEEP LEARNING

Deep learning (DL) is relatively a new technology which is a subset of machine learning. It consists of three or more multiple layers commonly known as neural networks or 'NNs'. The idea of using model layers to learn knowledge is very similar to the working of human brain. This novel approach allows NNs to use hidden multiple layers to understand complex features in a given data and make accurate predictions. DL has been very successful to solve various challenging AI problems such as cybersecurity, online frauds detection, digital assistants, and self-driving vehicles [14, 29].

Artificial Neural Networks (ANNs) uses deep learning principle and use logic node structures like neurons which are the building block of human brains. The only visible layers in ANNs are input and output layers. The rest the network is consist of multiple hidden layers of interconnected nodes with unique combinations of data inputs, weights, and bias. At each layer, input is analyzed against the given function, a bit of refining is done by adjusting some weights to optimise the final prediction. This process is called as 'forward propagation' which allows a deep network to learn itself without the help of any labelled data or explicit programming [16]. In deep learning, the data is fed into the input, then hidden layers does all the necessary processing and finally the result is produced by the output layer as shown in Figure 9 below.

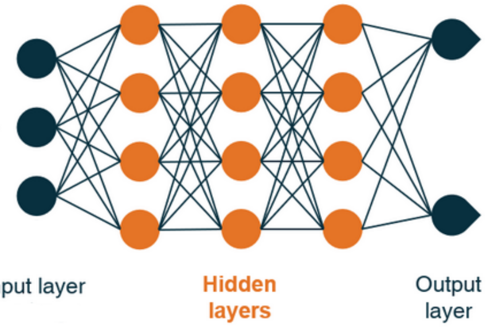


Figure 9: Artificial Neural Network (ANN) by using Deep Learning

There are different types of Deep Learning Networks. The Modular Neural Networks (MNNs) are used to solve complex tasks by breaking it into a system of separate and mostly independent subtasks that work together. They are similar to the specialized functionally regions in the brain that process colour and contrast separately. They use fully connected neural network where every neuron in one layer is connected to every other neuron in the next layer. Therefore, MNNs are called as a feed-forward neural network (FENNs) because inputs are processed only in the forward direction [20, 29].

Recurrent neural networks (RNNs) use MNNs to predict a class to each feature vector in a sequence and are often used in text processing because sentences and texts are naturally sequencing of either words, punctuation marks or sequences of characters.

A Convolution neural network (CNN) are a special type of feed-forward neural networks that significantly reduce the number of parameters in a deep network without diminishing the quality of the model so it uses a lot in image and text processing. A typical CNN will have a three-dimensional arrangement of neurons, instead of the standard two-dimensional array and contains the following layers: an input layer, the convolutional layers, the rectified linear unit layers, the pooling layers, the fully connected layer, and the output layer.

Based on my research, I have decided to use Recurrent Neural Network (RNN) model of Deep learning with an additional Long short-term memory (LSTM) unit which will allow the model to remember data features over a long amount of time [16, 29].

6.1 Using Long Short-Term Memory (LSTM) Deep Learning Model

The LSTM network has a unique ability to learn complex sequences like time series by using its long-term memory. In order to create a LSTM network, you need to develop Recurrent Neural Network (RNN) first. RNN networks uses short term memory to remember previous state by using weights of their cells in each individual layer just like a neuron in human brain. However, they struggle to learn complex sequences because they do not have a long-term memory. To solve this problem, the middle block of RNN hidden layers is replaced by a LSTM block which is designed to learn long-term states of a given data by constantly updating input weight values and add more recurrent subnetworks called memory blocks dynamically to remember the previous state.

During this research, I have learned that the main reason of failure of ML models based on Linear Regression and Support Vector Machines is their inability to learn about sudden fluctuations and important patterns in the given data [23]. By using Deep Learning models, I am hoping to develop a ML model which could predict share prices accurately.

The Figure 8 shows the implementation of LSTM model on Amazon share prices as shown below.

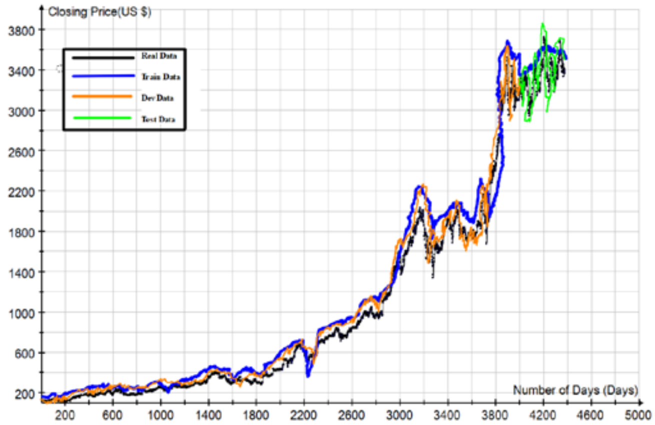


Figure 10: LSTM Deep Learning model of Adjusted Closing Price of Amazon shares

There is no doubt that Long Short-Term Memory (LSTM) offers the better fitting of the given data as compared to Linear Regression and Support Vector Machines.

In the first step, a LSTM model is created within RNN with one input layer to process raw data and one output layer to predict the corresponding share price. The hidden layer which is in this case is a LSTM which contains 50 blocks or neurons whose function is to learn various patterns in the given financial data by using a Rectifier Linear Unit or ReLU activation function.

To improve the model accuracy, some other traditional activation functions like sigmoid(σ) and hyperbolic tangent (*tanh*) are being used. However, ReLU has produced the the best result because it is the most appropriate activation function to learn about non-linear relationships to model dependencies between various features [20].

The LSTM RNN model is being trained by running 100 epochs or cycles to be able to understand the given complex data. During each epoch, a random set of test points are picked on the given time series input data and being processed by a LSTM block to learn about the data. Sets of unknown data points are picked, and consequent prediction are made. Then, the LSTM network helps the model to learn dynamically by using Mean Squared Error (MSE) to compare its accuracy and updates as shown by a blue curve in Figure 10. During the development stage, the output of LSTM model is feed into an additional Recurrent Neural Network (RNN) to improve the accuracy which is indicated by an orange curve. Finally, after compiling and optimising the LSTM model, the test data will be used to make final predictions as indicated by the green line in Figure 10.

6.2 Evaluation of Long Short-Term Memory (LSTM)

Again, LSTM model is evaluated by comparing its predictions with actual values in the test data. The following performance matrix in Table III summarises key statistical calculations.

TABLE III. Performance Matrix of LSTM Model

Statistical Calculation	R-Squared	Mean Squared Error	Root Mean Squared Error	Mean Absolute Percentage Error (%)
Symbol	**R^2**	**MSE**	**RMSE**	**MAPE**
Training Set	0.87	4,291.56	65.51	13.91
Validation Set	0.90	3,728.32	61.06	11.26
Testing/Prediction Set	0.94	3,1174.20	56.34	9.25

The Table III clearly shows that LSTM model is the more accurate than SVM and LR models because it has managed to understand random spikes in the given time series data by using its long-term memory feature. The R Squared values of all data sets are very close to 1 which indicates that this model fits very well to each data set. Specially, during the testing stage, a 0.94 R-Squared value shows that predicted values are very closer to the actual values. Similarly, both MSE and RMSE values are also reduced which means that this model has handled sudden fluctuations very well. Finally, MAPE values are very close to 10% for both validation and testing stages which shows that our LSTM model does not has underfitting problem because it performs very well on both seen and unseen data [29].

7. CONCLUSION

This study focuses on how to use historical stock market data to build and train financial models by using various Machine Learning (ML) algorithms. It uses the daily trade data of two major corporations (Amazon and FedEx) from December 2000 to December 2021 which is being collected from the Standard and Poor's 500 (S&P 500).

By using univariate analysis, it is concluded that the adjusted closing prices of Amazon are higher than FedEx due to higher median and mean values. However, Amazon share prices have a higher standard deviation as compared to FedEx, which indicates that investment in Amazon is risky because its shares prices are volatile and tend to fluctuate regularly.

The bivariate analysis shows that there is a strong positive correlation between the adjusted closing price (US$) and the given time (days) for both Amazon and FedEx stocks because their Pearson Correlation Coefficients (PCC) and Spearman Rank Correlation Coefficient (SRCC) values are very close to 1. It shows that share prices of both companies will increase in the future.

By using three ML models namely Linear Regression, Support Vector Machine and Long Short-Term Memory has been very helpful to find out if these models can predict the future movement of stock price values precisely by considering all fluctuations and seasonal variations. To evaluate the performance of each model, traditional statistical measures such as R-squared (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) have been used. The Linear Regression (LR) produced an inaccurate model due to big residuals between its predicted values and actual values. It also used a linear mathematical model to explain dependencies between share prices and time but failed to follow frequent variations in the given data. The second technique called Support Vector Machine (SVM) offered better fitting by using a non-linear kernel than LR. However, it also failed to capture sudden fluctuation in the time series data resulting in poor prediction. Finally, deep learning is being used to develop a Long Short-Term Memory (LSTM) unit along with a Recurrent Neural Network (RNN) to remember complex features of the given data. This model has produced very accurate results. Because the model uses its long-term memory feature and non-linear activation function, it manages to learn regular fluctuations and complex patterns very well.

After analysing the finding of this research and related performance review matrices; I believe that there is enough evidence to accept the null hypothesis (Ho) which states that there is no linear relationship exist between the adjusted closing price in US$ and the given time in days of a given stock. This study has proved that machine learning techniques which uses a non-linear relationship has improved the accuracy of predicting stock prices. It is shown by using quadratic kernel in SVM and polynomial kernel in LSTM which has produced more accurate better fitting models as compared to linear models. This study has also shown that deep learning technique can be used to understand complex trends and irregular variations in the financial trade data by using the power of its hidden multiple layers and non-linear optimisation kernels [29].

8. RECOMMENDATIONS

At the end of this study, I would like to make the following recommendations:

- This study has established that any financial data including stock price indicators change rapidly with high fluctuations as it is very dynamic, noisy, and nonlinear. Unfortunately, it is also very sensitive to certain qualitative (non-numerical) factors such as natural disasters, political instabilities, and terrorism. In this research, I did not consider those unpredictable external factors. It is highly recommended to include these factors while designing a stock market prediction system [29].
- This research was focussing solely on two major companies namely Amazon and FedEx daily trading data which is taken from the Standard and Poor's 500 (S&P 500) trading platform. However, it should be acknowledged that lots of important financial information like market fluctuation and price trends could be gained by analysing social media. Future researchers are recommended to learn how to analyse and extract meaning information from social media by using web scrapping tools.
- If you are planning to use Python code in Anaconda environment to test ML models, it is highly recommended to avoid working in the most recent configuration as most supporting deep learning libraries are not compatible with the newer version of Python. I will also recommend using other tools like Google Colab or MATLAB to avoid any compatibility issue.

9. FUTURE OF DEEP LEARNING IN FINANCE

This study has proved that LSTM unit coupled with RNN deep learning model offers the best predicting accuracy of the given data.

However, this is not the only combination which works very well to model complex financial data. Recently, there are some studies which have highlighted the importance of integration of different ML models to design efficient prediction models.

Some researchers have already achieved lots of success by using this approach. For example, Bae, Yue, and Rao (2017) designed a new stock market prediction based on deep learning features like Convolutional Neural Networks (CNNs), wavelet transformations, and auto encoders. Their framework was further improved by Gao (2016) who opted for using a Recurrent Neural System (RNN) and combined it with Long Short-Term Memory (LSTM) mechanism to remember information for long periods of time to improve qualitative processing skills of his model. Hiransha, Gopalakrishnan, Menon, and Soman (2018) improved deep learning-based stock price prediction model by adding more layers commonly known as Multi-Layer Preceptors (MLPs). They concluded that combination of deep learning models produced a better performance overall as compared to a stand-alone neural network [21, 29].

10. REFERENCES

[1] Dash R, Dash PK, A Hybrid Stock Trading Framework Integrating Technical Analysis with Machine Learning Techniques, The Journal of Finance and Data Science, 2016.

[Online]. Available:

DOI: 10.1016/j.jfds.2016.03.002

[2] J. D. Spiegeleer, D. B. Madan, S. Reyners, W. Schoutens, Machine learning for quantitative finance: fast derivative pricing, hedging, and fitting, Quantitative Finance, 18:10, 1635-1643, 2018.

[Online]. Available:

DOI: 10.1080/14697688.2018.1495335

[3] M. Nikou, G. Mansourfar, J. Bagherzadeh, Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms, First published: 03 December 2019.

[Online]. Available:

https://doi.org/10.1002/isaf.1459

[4] A. Bahrammirzaee, A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system, and hybrid intelligent systems, 2010.

[Online]. Available:

https://doi.org/10.1007/s00521-010-0362-z (Links to an external site.)

[5] S. Jansen, Machine Learning for Algorithmic Trading: Predictive models to extract signals from market and alternative data for systematic trading strategies with Python,

Packt Publishing; 2nd edition, 31 July 2020.

[6] J. McCarthy, What is artificial intelligence? Computer Science Department, Stanford University, Stanford, CA 94305, Nov 2004.

[Online]. Available:

http://www-formal.stanford.edu/jmc/

[7] Y. Hilpisch, Python for Finance: Mastering Data-Driven Finance, O ′ Reilly; 2nd edition, January 2019.

[8] Jiang ZY, Xu DX, Liang JJ, A deep reinforcement learning framework for the financial portfolio management problem, 2017.

[Online]. Available:

https://arxiv.org/abs/1706.10059

[9] A. D. Aydin, S. C. Cavdar, Comparison of prediction performances of artificial neural network (ANN) and vector autoregressive (VAR) models by using the macroeconomic variables of gold prices, Borsa Istanbul (BIST) 100 index and US dollar–Turkish lira (USD/TRY) exchange rates. Procedia Economics and Finance, 30, 3–14, 2015.

[10] W. Bao, , J. Yue, , Y. Rao, A deep learning framework for financial time series using stacked autoencoders and long - short term memory. PLoS ONE, 12(7), e0180944., 2017.

[Online]. Available:

https://doi.org/10.1371/journal.pone.018094

[11] Z. Guo, H. Wang, Q. Liu, , J. Yang, A feature fusion-based forecasting model for financial time series. PLoS ONE, 9(6), e101113, 2014.

[Online]. Available:

https://doi.org/10.1371/journal.pone.0101113

[12] R. Kissell, Algorithmic Trading Methods: Applications using Advanced Statistics, Optimization, and Machine Learning Techniques,

Academic Press, Second Edition, September 2020.

[13] I. Portugal, P. Alencar, D. Cowan, The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems with Applications, 97, 205–227, 2018.

[14] T. Hill, L. Marquez, M. O'Connor, W. Remus, Artificial neural network models for forecasting and decision making. International Journal of Forecasting, 10(1), 5–15, 1994.

[15] D. K. Bebarta, B. Biswal, P. K. Dash, Comparative study of stock market forecasting using different functional link artificial neural networks. International Journal of Data Analysis Techniques and Strategies, 4(4), 398-427, 2012.

[16] Y. Kara, , M. A. Boyacioglu, , Ö. K. Baykan, Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert systems with Applications, 38(5), 5311-5319, 2011.

[17] J. Patel, , S. Shah, , P. Thakkar, K. Kotecha, Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42(1), 259-268, 2015.

[18] A. Geron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems,

O'Reilly, Oct. 2019.

[19] S. Guido, A. C. Mueller, Introduction to Machine Learning with Python: A Guide for Data Scientists,

O'Reilly Media, 1st edition, 25 May 2016.

[20] A. Khashman, Neural networks for credit risk evaluation: An investigation of different neural models and learning schemes. Expert Syst Appl, 37(9):6233-6239, 2010.

[Online]. Available:

https://doi.org/10.1016/j.eswa.2010.02.101

[21] M. Raza, P. Cinquegrana, Deep Learning: The Latest Trend in AI and ML, 2021.

[Online]. Available:

https://www.qubole.com/blog/deep-learning-the-latest-trend-in-ai-and-ml/

Accessed on: 4th January 2022

[22] What Is Machine Learning: Definition, Types, Applications and Examples, Potential Analytics

[Online]. Available:

https://www.potentiaco.com/what-is-machine-learning-definition-types-applications-and-examples/

Accessed on: 4th January 2022

[23] Overfitting and Underfitting with Machine Learning Algorithms by Jason Brownlee

[Online]. Available:

https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Accessed on: 4th January 2022

[24] Python 3.9.0 Documentation

[Online]. Available:

https://www.python.org/downloads/release/python-390/

Accessed on: 12th February 2022

[25] Anaconda Individual Edition

[Online]. Available:

https://www.anaconda.com/products/individual

Accessed on: 12th February 2022

[26] Autograph 5 Documentation

[Online]. Available:

https://completemaths.com/autograph/5

Accessed on: 10th March 2022

[27] MATLAB Download UK

[Online]. Available:

https://uk.mathworks.com/products/matlab.html

Accessed on: 15th March 2022

[28] Introduction to Colab by Google Research

[Online]. Available:

https://colab.research.google.com/?utm_source=scs-index

Accessed on: 15th March 2022

[29] M. A. Khan, Using Artificial Intelligence in Stock Market Prediction: [Unpublished Master’s assignment for the Applied Research Methods Module], MSc Computer Science with Big Data Analytics, Wrexham Glyndwr University, 2022.