GIRIJANANDA CHOWDHURY INSTITUTE OF

MANAGEMENT AND TECHNOLOGY,

GUWAHATI

PROJECT REPORT

ON

"ITC STOCK MARKET PREDICTION"

Submitted in the requirement for the degree of

Bachelor of Technology

in

DEPARTMENT OF

COMPUTER SCIENCE AND ENGINEERING

ASSAM SCIENCE AND TECHNOLOGY UNIVERSITY GUWAHATI

Submitted by:	**Project Guide:**
Himasmita Bharadwaj (170310007022)	Dr. L.P Saikia
Pallabi Saikia(170310007036)	Professor
Debashree Baruah(170310007016)	Department of CSE
Session(2017-2021)	GIMT,Guwahati

DECLARATION

We hereby declare that this project work entitled **“ITC Stock market prediction”** was carried out by three of us under the guidance and supervision of **Dr L.P Saikia**, Professor of department of Computer science and engineering, Girijananda Chowdhury Institute of Management and Technology, Guwahati. This project is submitted to Department of Computer Science and Engineering of Girijananda Chowdhury Institute of Management and Technology, Guwahati during the academic year 2017-21. The work is never produced before any authority except Assam Science and Technology University for evaluation.

Name of Project Members	**Roll No**	**Signature**
1. Himasmita Bharadwaj	170310007022	
2. Pallabi Saikia	170310007036	
3. Debashree Baruah	170310007016	

GIRIJANANJA CHOWDHURY INSTITUTE OF MANAGEMENT AND TECHNOLOGY, GUWAHATI

SESSION 2017-21

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that **Himasmita Bharadwaj, Pallabi Saikia and Debashree Baruah**, students of B.Tech 4th Year, 8th Semester, have completed the project **"ITC Stock Market Prediction"** during this academic session 2020-21 under my guidance and supervision.

I approve this project for submission as required for the completion of the Bachelors of Technology Degree.

…......................................

(Signature of guide)

Dr. L.P Saikia
Professor
CSE, GIMT Guwahati

…......................................

(Signature of HOD)

Dr. Th. Shanta Kumar
Head of the department
CSE, GIMT Guwahati

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to our Principal, **Prof. Thuleshwer Nath,** GIMT-Guwahati, our guide **Dr. L.P Saikia**, Department of Computer Science and Engineering, GIMT-Guwahati and to our HOD, **Dr. Th. Shanta Kumar,** Department of Computer Science and Engineering, GIMT-Guwahati for giving us the freedom to work on this project and also for providing occasional guidance when needed. Our sincere gratitude also goes to all the faculties of Department of Computer Science and Engineering, GIMT-Guwahati for being with us, encouraging us to complete our project with honesty and sincerity. And lastly, we also convey our sincere thanks to all the lab bearers and office bearers for their active and enthusiastic cooperation.

Name of Project Members	**Roll No**	**Signature**
1. Himasmita Bharadwaj	170310007022	
2. Pallabi Saikia	170310007036	
3. Debashree Baruah	170310007016	

ABSTRACT

In ITC stock market prediction, the aim is to predict the future value of the financial stocks of a company. The recent trend in stock market prediction technologies is the use of machine learning which makes predictions based on the values of current stock market indices by training on their previous values. Machine learning itself employs different models to make prediction easier and authentic. The project focuses on the use of LSTM based Machine learning to predict stock values. Factors considered are open, close, low, high and volume.

CONTENTS

CHAPTER 1. INTRODUCTION

1.1.ITC LIMITED

ITC limited is an Indian multinational conglomerate company headquartered in Kolkata, WestBengal. Established in 1910 as the Imperial Tobacco Company of India Limited, the company was renamed as the India Tobacco Company Limited in 1970 and later to I.T.C. Limited in 1974. The company now stands renamed to ITC Limited, where "ITC" today is no longer an acronym or an initialized form. ITC has a diversified presence across industries such as Cigarettes,FMCG,Hotels,Packaging,Paperboards& Specialty Papers and Agribusiness. The company completed 100 years in 2010 and as of 2019–20, had an annual turnover of US$10.74 billion and a market capitalizationof US$35 billion. It employs over 36,500 people at more than 60 locations across India and is part of the Forbes 2000list.

"ITC Limited" was originally named "Imperial Tobacco" and was later renamed "Imperial Tobacco Company of India Limited", succeeding W.D. & H.O. Willson 24 August 1910 as a British-owned company registered in Calcutta.

Since the company was largely based on agricultural resource, it ventured into partnerships in 1911 with farmers from the southern part of India to source leaf tobacco. Under the company's umbrella, the "Indian Leaf Tobacco Development Company Limited" was formed in Guntur district of Andhra Pradesh in 1912. The first cigarette factory of the company was set up in 1913 at Bangalore.

Though the first six decades of the company's business were primarily devoted to the growth and consolidation of the cigarette and leaf-tobacco businesses, ITC's packaging & printing business at Mungerwas set up in 1925 as a strategic backward integration for ITC's cigarettes business. It is today India's most sophisticated packaging house. More factories were set up in the following years for cigarette manufacturing across India.

Fig 1.1: ITC Limited logo

ITC acquired Carreras Tobacco Company's factory at Kidder pore in 1935 to further strengthen its presence. ITC helped to set up indigenous cigarette tissue-paper-making plant in 1946 to significantly reduce the import costs and a factory for printing and packaging was set up at Madras in 1949. The company acquired the manufacturing business of Tobacco Manufacturers (India) Limited and the complementary lithographic printing business of Printers (India) Limited in 1953.

The company was converted into a Public Limited Company on 27 October 1954. The first step towards Indianization was taken in the same year with 6% of the Indian shareholding of the company. ITC also became the first Indian company to foray into consumer research during this time. During the 1960s, technology was given more focus on setting up cigarette machinery and filter-rod manufacturing facilities aimed at achieving self-sufficiency in cigarette-making.

AjitNarainHaskar became the company's first Indian chairman in 1969 and this was crucial in building up the Indian management for the company. As the company's ownership was progressively Indianized, under Haskar's leadership, the name of the company was changed from "Imperial Tobacco Company of India Limited" to "India Tobacco Company Limited" in 1970. ITC also became the first company in India to start from the 1971 Scissor's Cup. Innovative market campaigns and electronic data processing were started in the 1970s.

1.2. STOCK MARKET

A stock market, equity market, or share market is the aggregation of buyers and sellers of stocks(also called shares), which represent ownershipclaims on businesses; these may include securities listed on a public stock exchange, as well as stock that is only traded privately, such as shares of private companies which are sold to investorsthrough equity crowdfundingplatforms. Investment in the stock market is most often done viastockbrokeragesand electronic tradingplatforms. Investment is usually made with an investment strategyin mind.

Stocks can be categorized by the country where the company is domiciled. For example, Nestléand Novartisare domiciled in Switzerlandand traded on the SIX Swiss Exchange, so they may be considered as part of the Swissstock market, although the stocks may also be traded on exchanges in other countries, for example, as American depositary receipts(ADRs) on U.S. stock markets.

Astock exchangeis an exchangewhere stockbrokersand traderscan buy and sell shares,bonds, and other securities. Many large companies have their stocks listed on a stock exchange. This makes the stock more liquid and thus more attractive to many investors. The exchange may also act as a guarantor of settlement. These and other stocks may also be traded "over the counter" (OTC), that is, through a dealer. Some large companies will have their stock listed on more than one exchange in different countries, so as to attract international investors. Stock exchanges may also cover other types of securities, such as fixed-interest securities or derivatives, which are more likely to be traded OTC.

Trade in stock markets means the transfer (in exchange for money) of a stock or security from a seller to a buyer. This requires these two parties to agree on a price. Equities(stocks or shares) confer an ownership interest in a particular company. Participants in the stock market range from small individual stock investorsto larger investors, who can be based anywhere in the world, and may include banks,insurancecompanies,pension fundsand hedge funds. Their buy or sell orders may be executed on their behalf by a stock exchange trader.

The stock market is one of the most important ways for companiesto raise money, along with debt markets which are generally more imposing but do not trade publicly. This allows businesses to be publicly traded, and raise additional financial capital for expansion by selling shares of ownership of the company in a public market. The liquiditythat an exchange affords the investors enables their holders to quickly and easily sell securities. This is an attractive feature of investing in stocks, compared to other less liquid investments such as propertyand other immoveable assets.

History has shown that the price of stocksand other assets is an important part of the dynamics of economic activity, and can influence or be an indicator of social mood. An economy where the stock market is on the rise is considered to be an up-and-coming economy. The stock market is often considered the primary indicator of a country's economic strength and development. Rising share prices, for instance, tend to be associated with increased business investment and vice versa. Share prices also affect the wealth of households and their consumption. Therefore, centralbankstend to keep an eye on the control and behavior of the stock market and, in general, on the smooth operation of financial systemfunctions. Changes in stock prices are mostly caused by external factors such as socioeconomicconditions, inflation, exchange rates. Intellectual capitaldoes not affect a company stock's current earnings. Intellectual capitalcontributes to a stock's return growth. Market participantsinclude individual retail investors, institutional investorse.g., pension funds,insurance companies,mutual funds,index funds,exchange-traded funds,hedge funds, investor groups, banks and various other financial institutions, and also publicly traded corporations trading in their own shares. Robo-advisors, which automate investment for individuals are also major participants.

1.3. TRADE

Trade involves the transfer of goods or servicesfrom one person or entity to another, often in exchange for money. Economists refer to a systemor network that allows trade as a market. An early form of trade, the Gift economy, saw the exchange of goods and services without an explicit agreement for immediate or future rewards. A gift economy involves trading things without the use of money. Modern traders generally negotiate through a medium of exchange, such as money. As a result, buying can be separated from selling, or earning. The invention of moneygreatly simplified and promoted trade. Trade between two traders is called bilateral trade, while trade involving more than two traders is called multilateral trade. In one modern view, trade exists due to specialization and the division of labor, a predominant form of economic activityin which individuals and groups concentrate on a small aspect of production, but use their output in trades for other products and needs. Trade exists between regions because different regions may have a comparative advantage(perceived or real) in the production of some trade-able commodity—including production of natural resources scarce or limited elsewhere. For example: different regions' sizes may encourage mass production. In such circumstances, trade at market pricesbetween locations can benefit both locations.

International trade is the exchange of goods and services across national borders. In most countries, it represents a significant part of GDP. While international trade has been present throughout much of history (see Silk Road, Amber Road), its economic, social, and political importance have increased in recent centuries, mainly because of Industrialization, advanced transportation, globalization,multinational corporations, and outsourcing.

CHAPTER 2. REQUIREMENT ANALYSIS AND PREREQUISITES

2.1Hardware requirements:

- Minimum RAM 8GB, 12GB Recommended.
- 4GB of available disk space minimum, 8GB Recommended.
- 1280*800 minimum screen resolution.
- 64-Bit Operating System.

2.2Software Requirements**:**

- Linux-Ubuntu 18.04 LTS(64-bit).
- Python 3.6.
- Anaconda Navigator
- Jupyter Notebook

2.3TensorFlow

TensorFlow is an open source software library for high performance numerical computation. Its flexible architecture allows easy deployment of computation across a variety of platforms (CPUs, GPUs, TPUs,) and form desktops to clusters of servers to mobile and edge devices.

Originally developed by researchers and engineers form the Google's AI organization, it comes with strong support for machine learning and deep learning and the flexible numerical computation core is used across many other scientific domain.

2.4Keras

Keras is an open-source neural-network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or PlaidML. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible.

2.5Numpy

Numpy is a general-purpose array-processing package. It's provide a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package of scientific computing with Python. Besides its obvious scientific uses, Numpy can also be used as an efficient multidimensional container of generic data.

2.6 OpenCV

OpenCV supporta wide variety of programming language such as C++, Python, Java, etc., and is available on different platforms including Windows, Linux, OS X, Android, and Ios. Interfaces for high-speed GPU operations based on CUDA and OpenCL are also under active development. OpenCV-Python is the Python API for the OpenCV, combining the best qualities of the OpenCV C++ API and the Python language.

2.7 Scikit-Learn

Scikit-learn (formerly scikits.learn) is a free software machine learning library for the Python programming language.It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

2.8Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. There is also procedural “pylab” interface based on a state machine(like OpenGL), designed to closely resemble that of MATLAB, through its use is discouraged.

CHAPTER 3. BASIC CONCEPTS USED

3.1Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information.

Machine learning algorithms are often categorized as supervised or unsupervised.

Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

Unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn’t figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

3.2Deep Learning

Deep learning is a class ofmachine learningalgorithmsthat uses multiple layers to progressively extract higher level features from the raw input. For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces. Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks. Learning can be supervised, semi-supervised or unsupervised.

Deep learning architectures such asdeep neural networks,convolutionneural networksand recurrent neural networks have been applied to fields includingcomputer vision, image recognition, speech recognition,natural language processing,audio recognition, social network filtering etc.

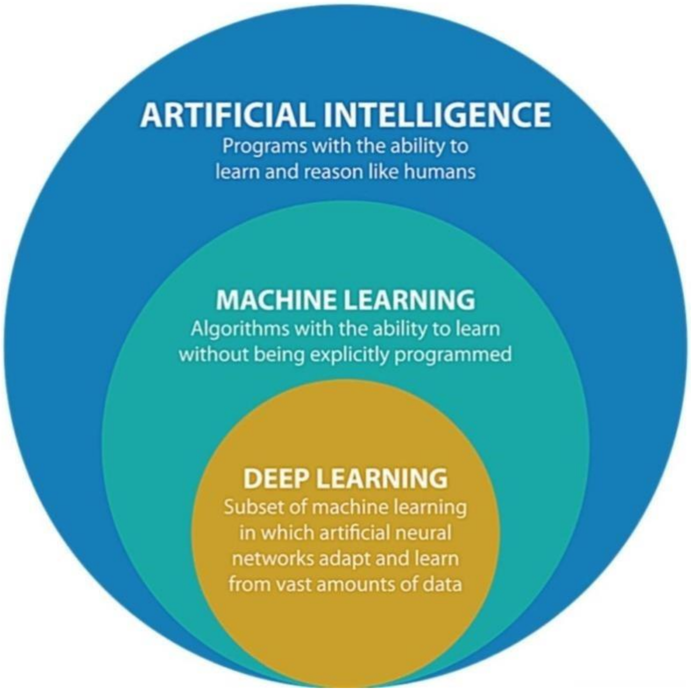

ARTIFICIAL INTELLIGENCE
Programs with the ability to learn and reason like humans
MACHINE LEARNING
Algorithms with the ability to learn without being explicitly programmed
DEEP LEARNING
Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data

Fig 3.1: Relation between AI, ML and DL

3.3Neural Networks

Neural networks are a series of algorithms that mimic the operations of a human brain to recognize relationships between vast amounts of data. A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. A “neuron” in a neural network is a mathematical function that collects and classifies information according to a specific architecture. A neural network contains layers of interconnected nodes. Each node is a perceptron and is similar to amultiple linear regression. The perceptron feeds the signal produced by a multiple linear regression into an activation function that may be nonlinear. Hidden layers fine-tune the input weightings until the neural network’s margin of error is minimal. It is hypothesized that hidden layers extrapolate salient features in the input data that have predictive power regarding the outputs. This describes feature extraction, which accomplishes a utility similar to statistical techniques such as principal component analysis.

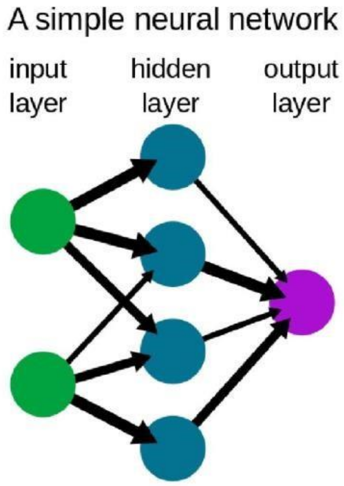


Fig3.2: Simple Neural Network

3.4Python

Python is an interpreted , high-level, general-purpose programming language. Created by Guido van Rossumand first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms including structured, object-oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Python interpreters are available for manyoperating systems. A global community of programmers develops and maintainsCPythonanopen sourcesreference implementation. Anon-profit organization, thePython Software Foundation, manages and directs resources for Python andCPythondevelopment.

From development to deployment and maintenance, Python helps developers be productive and confident about the software they're building.Benefits that make Python the best fitformachine learningand AI-based projects include simplicity and consistency, access to great libraries and frameworks for AI and machine learning (ML), flexibility, platform independence, and a wide community. These add to theoverall popularity of the language.

CHAPTER 4.LITERATURE SURVEY

SL no.	TITLE	AUTHOR	SUMMARY
1	Study on the prediction of stock price based on the associated network model of LSTM	Guangyu Ding & LiangxiQin(School of Computer, Electronics and Information, Guangxi University, Nanning, 530004, Guangxi, China)	In this paper, a multi-value associated network model of LSTM-based deep-recurrent neural network (Associated Net) is proposed to predict multiple prices of a stock simultaneously.

Table 4.1 Literature Survey

CHAPTER 5. MATERIALS

5.1. Dataset

The dataset on bases of which we have prepared our project is taken from Kaggle. We have 5141 data with each having element :

1. Date -date of the record
2. EQ - It stands for Equity. In this series intraday trading **is** possible in addition to delivery.
3. Prev Close - Previous close is a security's closing price on the preceding day of trading can show substantial changes from a previous close to new open.
4. Open-It is the price at which the financial security opens in the market when trading begins. It may or may not be different from the previous day's closing price.
5. High-The high is the highest price at which a stock traded during a period.
6. Low- The low is the lowest price of the period.
7. Last-last means the final quoted trading price for a particular stock.
8. Close-The close is a reference to the end of a trading session in the financial markets when the markets close for the day.
9. VWAP-The volume weighted average price (VWAP) is a trading benchmark used by traders that gives the average price a security has traded at throughout the day, based on both volume and price. It is important because it provides traders with insight into both the trend and value of a security.
10. Volume-Trading volume is a measure of how much of a given financial asset has traded in a period of time. For stocks, volume **is** measured in the number of shares traded and, for futures and options; it is based on how many contracts have changed hands.
11. Turnover-Share turnover is a measure of stock liquidity, calculated by dividing the total number of shares traded during some period by the average number of shares outstanding for the same period. The higher the share turnover, the more liquid company shares are.
12. Trade- trading refers to the buying and selling of securities.
13. Deliverable Volume-Deliverable quantity or Deliverable Volumeis the quantity of shares which actually move from one set of people (who had those shares in their demat account before today and are selling today) to another set of people (who have purchased those shares and will get those shares by T+2 days in their demat account).
14. % Deliverable-A rise in deliverable volume with falling stock price indicates bearishness on the stock on that particular day. For instance, if total traded volume of share A is 100 on a given day and delivery volume is 50 per cent, it means out of 100 shares, 50 shares actually changed hands, moving from one owner to another.

CHAPTER 6. METHODOLOGIES

6.1. LSTM

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feed forward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unfermented, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems).

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications.

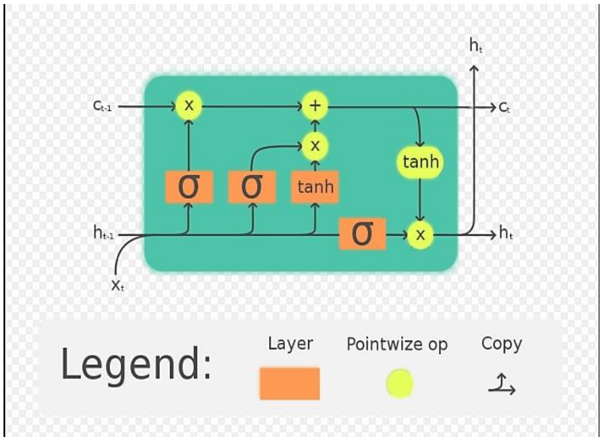


Fig6.1: Long-Short Term Memory

6.1.1Variants

In the equations below, the lowercase variables represent vectors. Matrices W_q and U_q contain, respectively, the weights of the input and recurrent connections, where the subscript q can either be the input gate$i$, output gateo , the forget gate $f$or the memory cell c, depending on the activation being calculated. In this section, we are thus using a "vector notation". So, for example, $c_t = \mathbb{R}^h$ is not just one cell of one LSTM unit, but contains h LSTM unit's cells.

6.1.2LSTM with a forget gate

The compact forms of the equations for the forward pass of an LSTM unit with a forget gates are:

$$f_t = \sigma_g (W_f x_t + U_f h_{t-1} + b_f)$$

$$i_t = \sigma_y(W_i x_t + U_i h_{t-1} + b_i)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)$$

$$\tilde{c_t} = \sigma_c(W_c x_t + U_o h_{t-1} + b_c)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t h_t =$$

$$o_t \odot \sigma_h(c_t)$$

where the initial values are $c_0 = 0$ and$h_0 = 0$ and the operator ⊙ denotes theHadamard product(element-wise product). The subscript t indexes the time step.

6.1.3Variables

- $x_t \in \mathbb{R}^d$: input vector to the LSTM unit
- $f_t \in \mathbb{R}^h$: forget gate's activation vector
- $i_t \in \mathbb{R}^h$: input/update gate's activation vector
- $o_t \in \mathbb{R}^h$: output gate's activation vector
- $h_t \in \mathbb{R}^h$: hidden state vector also known as output vector of the LSTM unit
- $\tilde{c_t} \in \mathbb{R}^h$: cell input activation vector
- $c_t \in \mathbb{R}^h$: cell state vector
- $W \in \mathbb{R}^{h \times d}$,$U \in \mathbb{R}^{h \times h}$ and $b \in \mathbb{R}^h$: weight matrices and bias vector parameters which need to be learned during training

where the superscripts d and t refer to the number of input features and number of hidden units, respectively.

6.1.4Activation functions

σ_g:Sigmoidfunction.
σ_c:Hyperbolictangentfunction.
σ_h: Hyperbolic tangent function or, as the peephole LSTM paper suggests, $\sigma_h(x) = x$

6.1.5Training

A RNN using LSTM units can be trained in a supervised fashion, on a set of training sequences, using an optimization algorithm, likegradient descent, combined withback propagation through timeto compute the gradients needed during the optimization process, in order to change each weight of the LSTM network in proportion to the derivative of the error (at the output layer of the LSTM network) with respect to corresponding weight.

A problem with usinggradient descentfor standard RNNs is that error gradientsvanishexponentially quickly with the size of the time lag between important events. This is due to $\lim_{n\to\infty} W^n = 0$ if thespectral radiusof W is smaller than 1.

However, with LSTM units, when error values are back-propagated from the output layer, the error remains in the LSTM unit's cell. This "error carousel" continuously feeds error back to each of the LSTM unit's gates, until they learn to cut off the value.

6.2. RANDOM FOREST CLASSIFICATION

Random forest is a flexible, easy to usemachine learning algorithmthat produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). In this post we'll learn how the random forest algorithm works, how it differs from other algorithms and how to use it.

Random forests or random decision forests are anensemble learningmethod forclassification,regressionand other tasks that operates by constructing a multitude ofdecisiontreesat training time and outputting the class that is themodeof the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit ofover fittingto theirtraining set. Random forests generally outperformdecision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 byTin Kam Housing therandom subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed byLeo BreimanandAdele Cutler, who registered "Random Forests" as atrademarkin 2006 (as of 2019, owned byMinitab, Inc.). The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit andGemanin order to construct a collection of decision trees with controlled variance.

Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

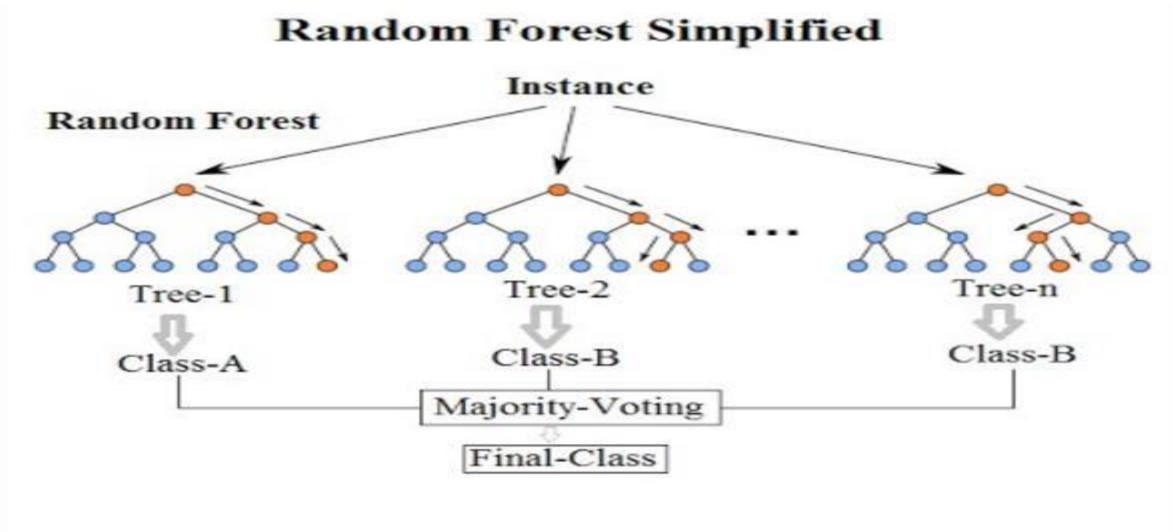


Fig6.2: Random Forest Tree

6.2.1BEGGING

The training algorithm for random forests applies the general technique ofbootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$, bagging repeatedly (B times) selects arandom sample with replacementof the training set and fits trees to these samples:

For $b = 1, ..., B$:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.

2. Train a classification or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$f = \frac{1}{B}\sum_{b=1}^{B} f_b(x')\ ''$$

or by taking the majority vote in the case of classification trees.

This bootstrapping procedure leads to better model performance because it decreases thevarianceof the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees (or even the same tree many times, if the training algorithm is deterministic); bootstrap sampling is a way of de-correlating the trees by showing them different training sets.

Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees on x':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \widehat{f})^2}{B-1}}$$

The number of samples/trees, B, is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. An optimal number of trees B can be found usingcross-validation, or by observing theout-of-bag error: the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample.[15]The training and test error tend to level off after some numbers of trees have been fit.

6.2.2From bagging to random forests

The above procedure describes the original bagging algorithm for trees. Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, arandom subset of the features. This process is sometimes called "feature bagging". The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a fewfeaturesare very strong predictors for the response variable (target output), these features will be selected in many of the B trees, causing them to become correlated. An analysis of how bagging and random subspace projection contributes to accuracy gains under different conditions is given by Ho.

Typically, for a classification problem with p features, √$\boxed{p}$ (rounded down) features are used in each split. For regression problems the inventors recommend $p/3$ (rounded down) with a minimum node size of 5 as the default. In practice the best values for these parameters will depend on the problem, and they should be treated as tuning parameters.

Chapter 7: DISCUSSIONS AND RESULTS

7.1 Matrix evaluation

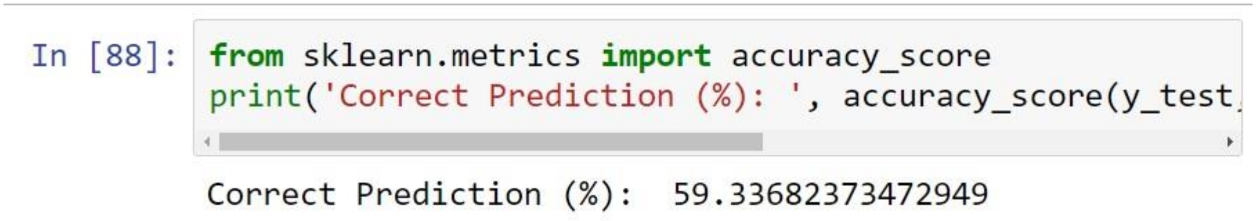
The idea of building machine learning model works on a constructive feedback principle. We build a model, get feedback from metrics, make improvements and continue until we achieve a desirable accuracy. Evaluation metrics explain the performance of a model. An important aspect of evaluation metrics is their capability to discriminate among model results. Many analysts and aspiring data scientists do not even bother to check how robust their model is. Once they are finished building a model, they hurriedly map predicted values on unseen data. This is an incorrect approach. Simply building a predictive model is not your motive. It’s about creating and selecting a model which gives high accuracy out of sample data. Hence, it is crucial to check the accuracy of your model prior to computing predicted values.There are several evaluation metrics like confusion matrix, cross-validation, AUC-ROC curve, etc.Different evaluation metrics are used for different kinds of problems.

7.1.1Classification accuracy

Classification accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples. It works well only if there are equal number of samples belonging to each class. The formula for Classification accuracy is as follows:

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

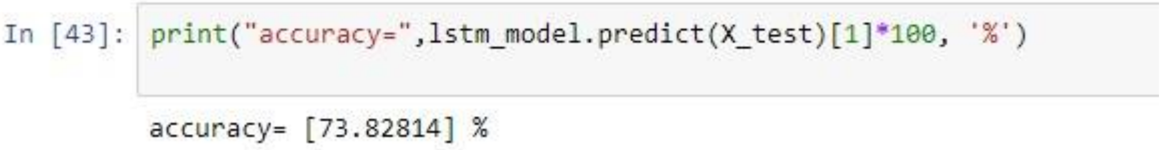
Classification accuracy can also be calculated by summing up the diagonal elements dived by the total amount of test-data.



```
In [88]: from sklearn.metrics import accuracy_score
         print('Correct Prediction (%): ', accuracy_score(y_test_

         Correct Prediction (%):  59.33682373472949
```

Fig7.1: Accuracy Score of Random Forest Classification

Accuracy score for this dataset came out to be 59% using Random forest classification.



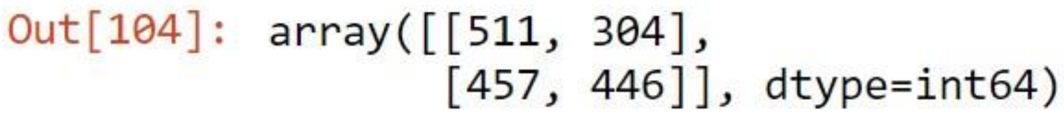
```
In [43]: print("accuracy=",lstm_model.predict(X_test)[1]*100, '%')

         accuracy= [73.82814] %
```

Fig7.2: Accuracy Score of LSTM

Accuracy Score for our dataset after using LSTM is 79 %   , which is higher than random forest. Hence lstm tend to give more accurate prediction in our dataset.

7.1.2Confusion Matrix

A confusion matrix is an N X N matrix, where N is the number of classes being predicted. The confusion matrix can only be used when the true values are known meaning if we have an evaluation procedure i.e. train, test, split then we have a test-data for which the true values are known & when we make predictions on the test-data then we can use a confusion matrix to evaluate the performance.



```
Out[104]:  array([[511, 304],
                  [457, 446]], dtype=int64)
```

Fig7.3 : Confusion matrix for the project.

Here are a few definitions; we need to remember for a confusion matrix:

- **True Positives**: The cases in which we predicted YES and the actual output was also YES.
- **True Negatives**: The cases in which we predicted NO and the actual output was NO.
- **False Positives**: The cases in which we predicted YES and the actual output was NO.
- **False Negatives**: The cases in which we predicted NO and the actual output was YES.

7.1.3Precision

Also called positive predictive value. The ratio of correct positive predictions to the total predicted positives.

Formula for Precision is as follows:

$$P = \frac{TP}{TP + FP}$$

7.1.4Recall

Also called Sensitivity, Probability of Detection, True Positive Rate. Ratio of correct positive predictions to the total positive. The formula for Recall is as follows:

$$R = \frac{TP}{TP + FN}$$

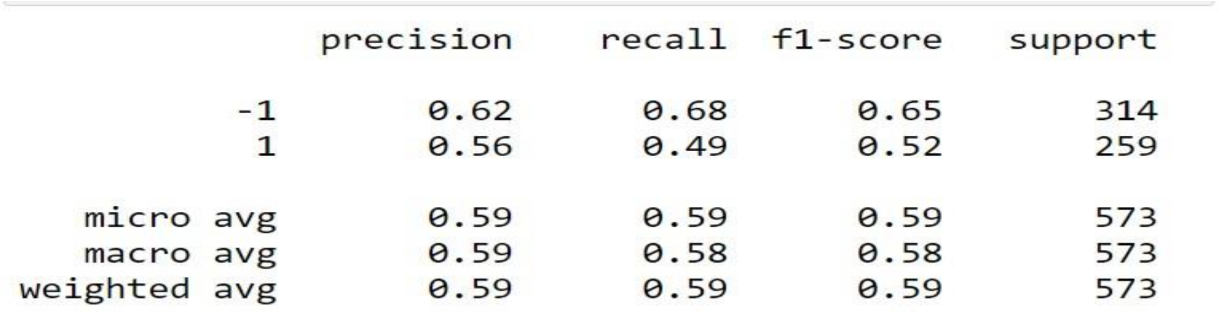
7.1.5 F1-Score

F1-Score is the harmonic mean of precision and recall values for a classification problem. The formula for F1-Score is as follows:

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

7.1.6Support

Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing.



```
              precision    recall  f1-score   support

          -1       0.62      0.68      0.65       314
           1       0.56      0.49      0.52       259

   micro avg       0.59      0.59      0.59       573
   macro avg       0.59      0.58      0.58       573
weighted avg       0.59      0.59      0.59       573
```

Fig7.4: precision, recall, f1-score, support for this project

7.2 GRAPH

7.2.1 Graphical representation for sample dataset

7.2.1.1 Close value v/s predicted value graph for Random Forest Classification

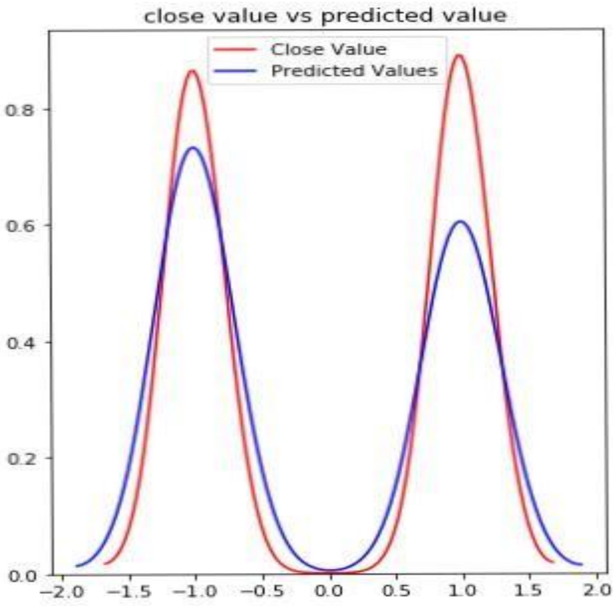


Fig7.5: close value v/s predicted value graph for sample dataset

In this Graph, we can see how the close value curve is different from the predicted value curve. Predicted value curve shows that it was predicted to close the valve at low rate where as the actual closing value went higher than predicted.

7.2.1.2 Close value v/s predicted value graph using LSTM

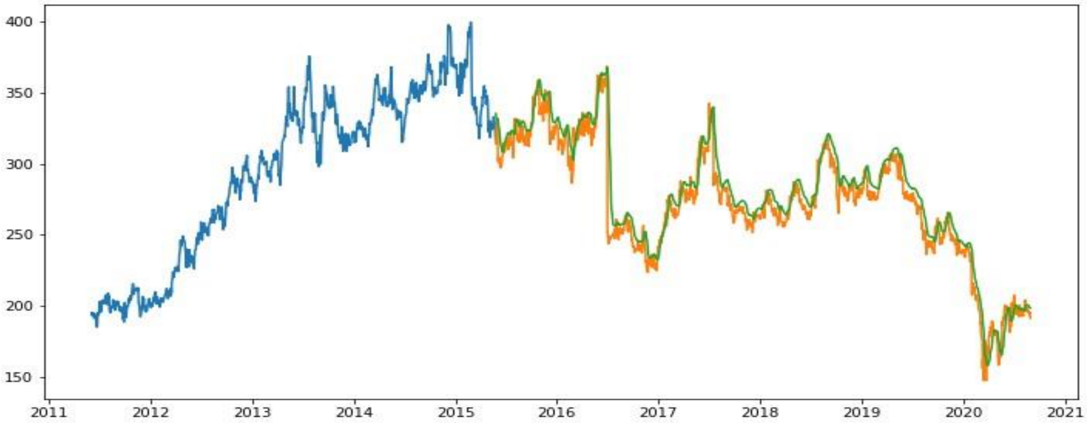


Fig7.6: close v/s predicted value graph using Lstm for sample data set

Here we can see that after training the dataset with LSTM we get more accuracy on the close value and predicted value .The orange curve on the graph depicts the close value where as the green depicts the predicted value by LSTM. Hence, LSTM are supposed to be more beneficial while working with stocks, trading or markets.

7.2.2 Graphical representation for actual dataset

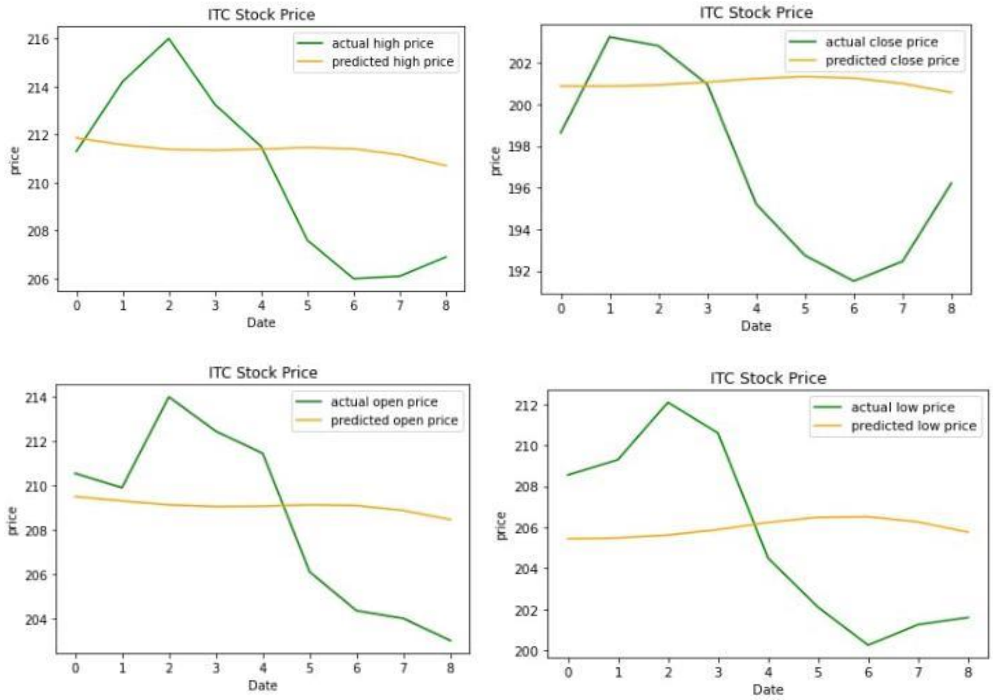


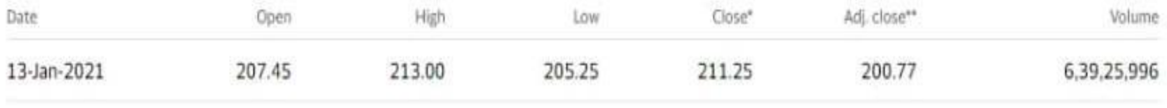
Fig 7.7: Actual v/s Predicted Price for High, Adj Close, Open and Low price

Here we can see that after training the dataset with LSTM we get more accuracy on the close, open, high, low actual value and predicted value .The orange curve on the graph depicts the predicted price where as the green depicts the actual price by LSTM.

Here we have used LSTM in the actual ITC dataset because we have got more accuracy using LSTM rather than random forest classification.

7.3 RESULTS

We have obtained the actual dataset of ITC through yahoo**.** Here we will compare our prediction and the actual price of the same data for Date: 13 Jan 2021 side by side to compare how much the difference of results both the price has.



Date	Open	High	Low	Close*	Adj. close**	Volume
13-Jan-2021	207.45	213.00	205.25	211.25	200.77	6,39,25,996

Fig 7.11 Actual historical price for ITC Stock on 13/01/2021 as found on website

Date	**Open**	**High**	**Low**	**Close***	**Adj. Close****	**Volume**
13 Jan 2021	207.86	210.12	205.096	------	200.09	------------

Table 7.1: Predicted price for ITC Stock by our model on 13/01/2021

As we can see in the above two figures, the data found in yahoo finance and the predicted data done by us are comparatively similar. On date: 13 Jan 2021, the opening prices for ITC have only the difference of 0.41, similarly the high prices of both the figures have the difference of 3.00, similarly the low prices have the difference of 0.154, and the adj. closing price have the difference 0.68. For close price and volume we have not done any prediction as our main aim was to see if our model can predict an accurate value or not, which we got to be very precise.

Hence, we can conclude that using LSTM in our project have given us more accurate results as we can see there is only a minimal difference between our predicted results and the results found on yahoo .

CHAPTER 8. APPLICATIONS

1. **Field of Trading:** This project aims to determine the future movement of the stock value of a financial exchange. Stock traders need to predict trends in stock market behaviour for accurate result. The accurate prediction of share price movement will lead to more profit investors can make.

2. **Field of Education:** This project can be proven beneficial for students to use as a reference model for creating other similar kind of classification tools and therefore add up new techniques and improve the efficiency of the already existing models and aiding in a more rapid and efficient classification process.

3. **Field of Research and Development**: This project can be used as the base model for further research or development of other classification or detection models that are related to other company's stock market.

CHAPTER 9. ADVANTAGES

1. **More accurate result**: In determining the prediction a large number of data is collected and this data is to be compiled, analyzed and interpreted. This requires use of certain statistical methods like standard deviation, standard error, application of tests of statistical significance like Z Test, unpaired and paired t test and chi-square test. Statistical methods are time consuming. With the help of Machine learning, large number of statistical calculations can be performed in a very short time.

2. **Readability of large datasets:** In our project Machine Learning has been used whichout performs other techniques as the data size is large**.** This approach is effective for new applications, or for applications which will have a relatively big number of output categories. This is a relatively less popular approach because, with the rate of learning and large volumes of data, the networks typically take significantly more time to train but provides faster test results.

3. **Continuous improvement:** As machine learning gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. As the amount of data we kept are growing, the algorithm learn to make more accurate predictions faster.

4. **Easily identifies trends and patterns:** Machine learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance for an e-commerce website, it serves to understand the browsing behaviors and purchase histories of its user to help cater to the right product, deals and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

5. **No human intervention needed (automation):** With machine learning we don't need to babysit our project every step of the way. Since it gives the machine the ability to learn, it lets them make predictions and also improve the algorithm on their own. It is also good at recognizing spam.

6. **Ability to deliver high-quality results:** Humansget hungry or tired and sometimes make careless mistakes. When it comes to neural networks, this isn’t the case. Once trained properly, a machine learning model becomes able to perform thousands of routine, repetitive tasks within a relatively shorter period of time compared to what it would take for a human being. In addition, the quality of the work never degrades, unless the training data contains raw data which doesn’t represent the problem you’re trying to solve.

CHAPTER 10. LIMITATIONS

The following are the limitations of the project:

1. The main aim of this system is to provide a general idea of where the ITC stock market is headed. It is only limited to a very basic prediction model. Thus, it cannot be used as a critical decision making tool. Since, there are many indeterminate parameters that directly affect stock market, each and every one of them cannot be taken into account. So, our model only depends on the relationship of our selected parameters with the share price.

2. The project has the use of machine learning which has some disadvantages too:-

   a) It requires very large amount of data in order to perform better than other techniques.

   b) While training the model takes lot amount of time.

   c) It is extremely expensive to train due to complex data models.

CHAPTER 11. CONCLUSION

ITC stock market prediction lays the foundation for democratizing machine learning technologies for retail investors, connecting predictions made by machine learning models to retail investors. It helps investors navigate through the stock markets with additional analysis and help them make more informed decisions.

The findings demonstrated that the application provides significance in trend prediction. When compared to the baseline, the prediction shows useful trend tendency with the real stock trend. Through this project, the user can easily compare the predictions and model scores from different machine learning models, then choosing the one that fits their preference. The models used in the application will continue to improve it by searching for a better model topology, structure and hyper parameters through evolution algorithm. The findings concluded the usefulness of evolution algorithm in lowering the mean squared error when predicting stock prices, which is helpful for improving the trend prediction for retail investors.

Therefore, with the application and research findings, to large extent the project team achieved the aim of creating an user-friendly system for retail investors whom does not have previous technical knowledge to navigate the machine model predictions result with useful benchmarks.

CHAPTER 12. FUTURE SCOPE

ITC stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit. Potential improvement can be made in our data collection and analysis method. Future research can be done with possible improvement such as more refined data and more accurate algorithm. It will help a lot in the trading sector and will help the economy to grow further.

BIBLIOGRAPHY

1. Guangyu Ding &Liangxi Qin's Study on the prediction of stock price based on the associated network model of LSTM https://link.springer.com/article/10.1007/s13042-019-01041-1#citeas

2.https://upload.wikimedia.org/wikipedia/commons/thumb/f/ff/ITC_Limited_Logo.svg/250pxITC_Limited_Logo.svg.png

3. https://mlpython.in/wp-content/uploads/2020/08/f1-1-e1596985579733.png

4.https://upload.wikimedia.org/wikipedia/commons/thumb/9/99/Neural_network_example.svg/220 px-Neural_network_example.svg.png

5.Dataset

https://www.kaggle.com/rohanrao/nifty50-stock-market-data?select=ITC.csv

https://in.finance.yahoo.com/quote/ITC.NS/history/