SMART INVESTMENT PREDICTION

1.1 INTRODUCTION :

A stock market, equity market or share market is the aggregation of buyers and sellers (a loose network of economic transactions, not a physical facility or discrete entity) of stocks (also called shares), which represent ownership claims on businesses; these may include securities listed on a public stock exchange, as well as stock that is only traded privately. Examples of the latter include shares of private companies which are sold to investors through equity crowd funding platforms. Stock exchanges list shares of common equity as well as other security types, e.g. corporate bonds and convertible bonds.

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. The efficient-market hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable. Others disagree and those with this viewpoint possess myriad methods and technologies which purportedly allow them to gain future price information.

Predicting how the stock market will perform is one of the most difficult things to do. There are so many factors involved in the prediction – physical factors vs. physiological, rational and irrational behavior, etc. All these aspects combine to make share prices volatile and very difficult to predict with a high degree of accuracy.

Machine learning techniques have the potential to unearth patterns and insights we didn’t see before, and these can be used to make unerringly accurate predictions.

1.2 Objective Of Research:

The main objectives of stock market are:

Raising Money For Business: Stock exchanges around the world enable companies around the world to raise money.

Capital Formation: The primary function of a stock exchange is to help companies raise money.

Security And Transparency: The legitimate sale of stock on any exchange requires reliable and accurate information.

Companies plan to use the newly-raised funds to invest in productive business assets and grow revenues and profits.

1.3 Problem Statement:

We'll dive into the implementation part of this article soon, but first it's important to establish what we're aiming to solve. Broadly, stock market analysis is divided into two parts – Fundamental Analysis and Technical Analysis. Fundamental Analysis involves analyzing the company's future profitability on the basis of its current business environment and financial performance. Technical Analysis, on the other hand, includes reading the charts and using statistical figures to identify the trends in the stock market.

The use of fundamental and technical analysis methods are basis of the predictions of future stock price movement. These tools show a trend on future movement and not the figure of the most likely trade price for any stock in future. It is there- fore desirable to have a tool that does not just point at a direction of price movement. Machine Learning methods that can actually analyse the stock prices over time and gain intelligence, then use this intelligence in prediction, can be used to model such a tool.

1.4 Industry Profile :

The main significant approach, used in this paper for the predicting result is a concept of machine learning and result tested on the Bombay Stock Exchange (BSE) index data set. To seize the best accurate output, the approach decided to be implemented is machine learning along with supervised classifier.

2. REVIEW OF LITERATURE:

In the last two decades forecasting of stock returns has become an important field of research. In most of the cases the researchers had attempted to establish a linear relationship between the input macroeconomic variables and the stock returns. But with the discovery of nonlinear trends in the stock market index returns[4], there has been a great shift in the focus of the researchers towards the nonlinear prediction of the stock returns. Although, there after many literature have come up in nonlinear statistical modelling of the stock returns, most of them required that the nonlinear model be specified before the estimation is done.

But for the reason that the stock market return being noisy, uncertain, chaotic and nonlinear in nature, ANN has evolved out to be better technique in capturing the structural relationship between a stock's performance and its determinant factors more accurately than many other statistical techniques[5] In literature, different sets of input variables are used to predict stock returns.

In fact, different input variables are used to predict the same set of stock return data. Some researchers used input data from a single time series where others considered the inclusion of heterogeneous market information and macro economic variables. Some researchers even pre-processed these input data sets before feeding it to the ANN for forecasting.

1. A recent study (et al.Risul Islam Rasel ,Nasrin Sultana ,NasimulHasan, IEEE 2016) has shown that ANN model can be more advantageous compared to other SVM or LR models and the Advantages are Increase in accuracy with multiple attributes[4]. Works well even if attributes and output do not have a clear relation. Also, some disadvantages also must be considered which are Time required for prediction is more than other methods Can face over fitting problem.

2. Chan., Wong., and Lam., implemented a neural network model using the technical analysis variables for listed companies in Shanghai Stock Market. In this paper performance of two learning algorithm and two weight initialization methods are compared. The results reported that prediction of stock market is quite possible with both the algorithm and initialization methods but the performance of the efficiency of the back propagation can be increased by conjugate gradient learning and with multiple linear regression weight initialization.

3. DATA COLLECTION:

We need a training data set. It is the actual data set used to train the model for performing various actions. In our experiment, we created our own dataset for investment prediction that contains the following set of columns as:

- Grade of the company - contains the grade value from 1 to 5 which represents from low to high grade.
- No. of offline projects and No. of online projects - refers to the projects that are taken by the related company.
- Net turnover - refers to the overall profits that gained by the company in that specific year.
- Share price - refers to the values holed by the shareholders in the company.

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter.

The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split. An example of a decision tree can be explained using above binary tree.

Let's say you want to predict whether a person is fit given their information like age, eating habit, and physical activity, etc. The decision nodes here are questions like 'What's the age?', 'Does he exercise?', 'Does he eat a lot of pizzas'? And the leaves, which are outcomes like either 'fit', or 'unfit'. In this case this was a binary classification problem (a yes no type problem).

There are two main types of Decision Trees:

1. Classification trees (Yes/No types) What we've seen above is an example of classification tree, where the outcome was a variable like 'fit' or 'unfit'. Here the decision variable is Categorical.
2. . Regression trees (Continuous data types) Here the decision or the outcome variable is Continuous, e.g. a number like 123.

4. METHDOLOGY:

4.1 Exploratory Data Analysis:

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA),[1] which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

The objectives of EDA are to:

• Suggest hypotheses about the causes of observed phenomena • Assess assumptions on which statistical inference will be based

• Support the selection of appropriate statistical tools and techniques

• Provide a basis for further data collection through surveys or experiments

4.1.1 Figures And Tables:

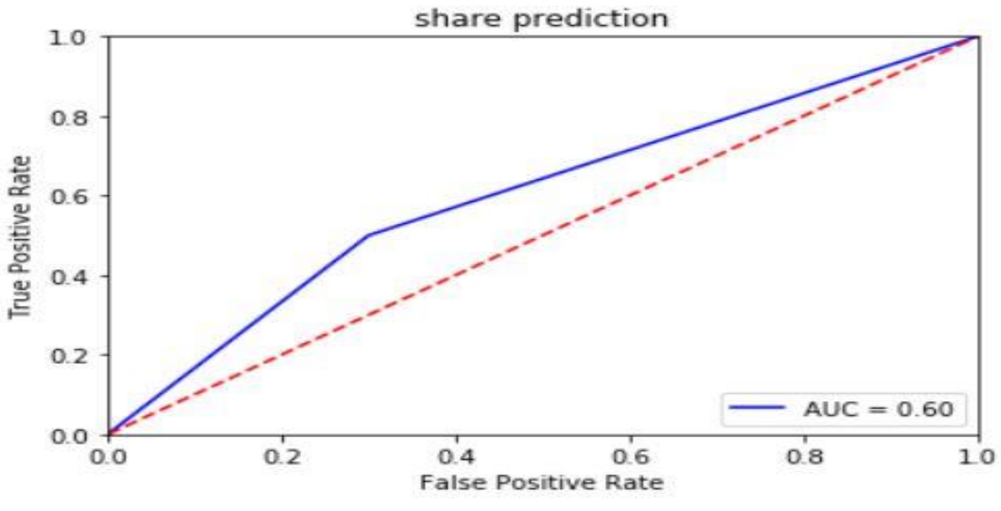
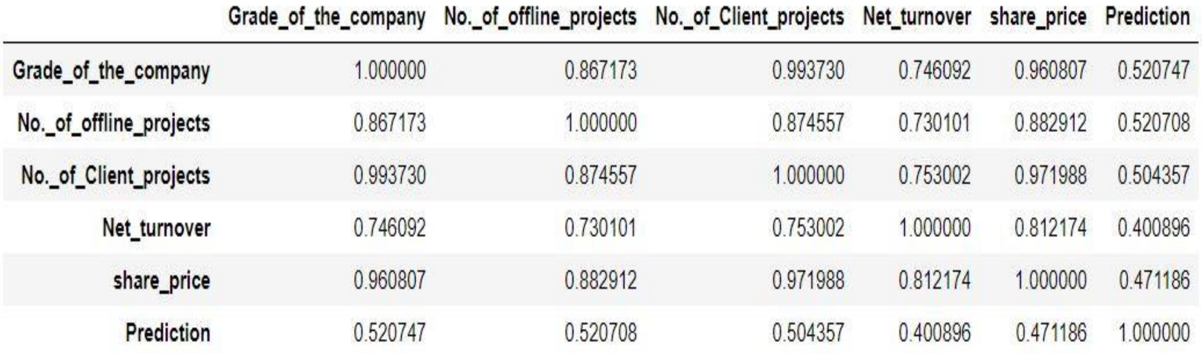
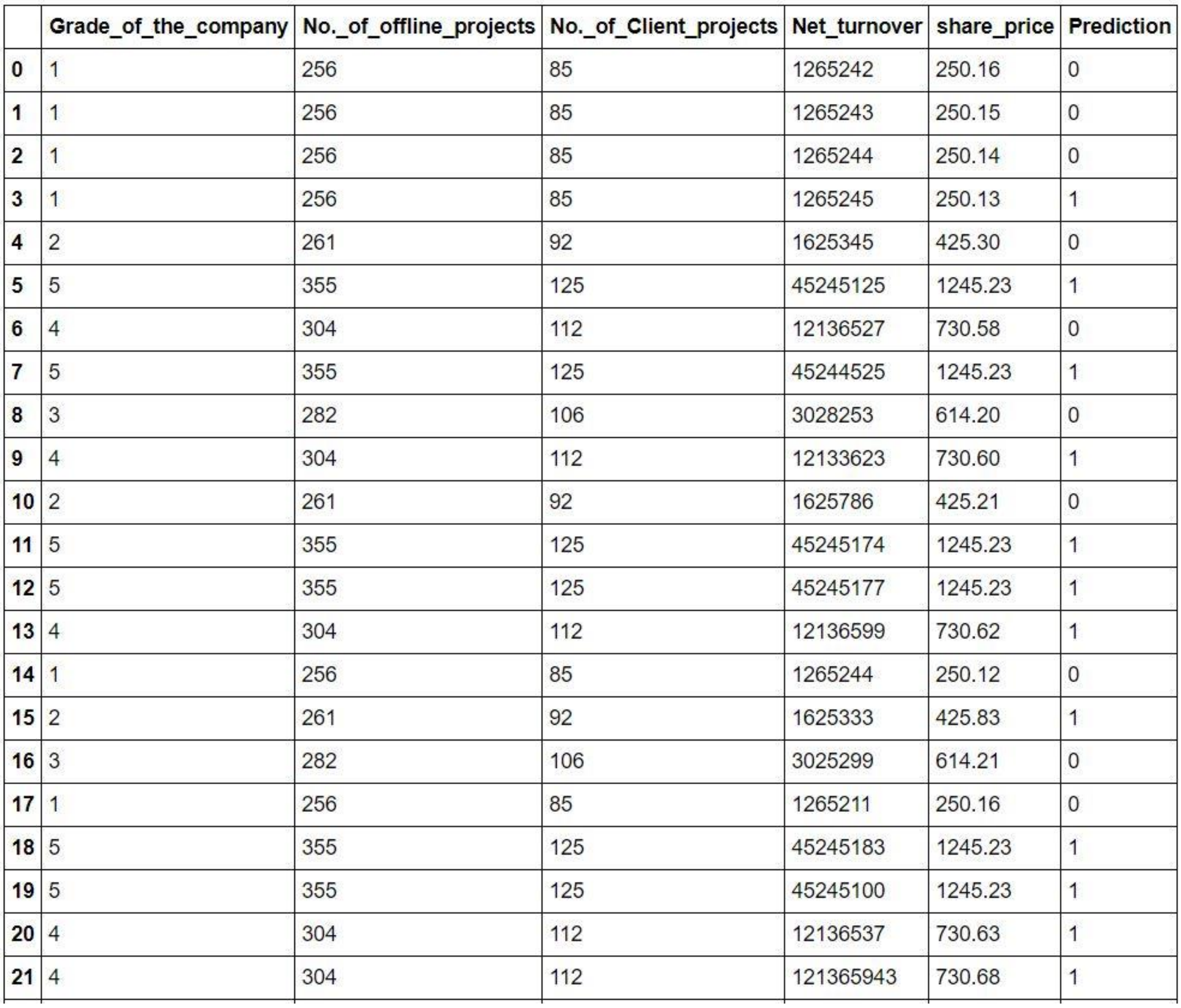


FIGURE: 4.1.2 AOC Curve



	Grade_of_the_company	No._of_offline_projects	No._of_Client_projects	Net_turnover	share_price	Prediction
Grade_of_the_company	1.000000	0.867173	0.993730	0.746092	0.960807	0.520747
No._of_offline_projects	0.867173	1.000000	0.874557	0.730101	0.882912	0.520708
No._of_Client_projects	0.993730	0.874557	1.000000	0.753002	0.971988	0.504357
Net_turnover	0.746092	0.730101	0.753002	1.000000	0.812174	0.400896
share_price	0.960807	0.882912	0.971988	0.812174	1.000000	0.471186
Prediction	0.520747	0.520708	0.504357	0.400896	0.471186	1.000000

TABLE 4.1.3 CO-RELATION ANALYSIS



	Grade_of_the_company	No._of_offline_projects	No._of_Client_projects	Net_turnover	share_price	Prediction
0	1	256	85	1265242	250.16	0
1	1	256	85	1265243	250.15	0
2	1	256	85	1265244	250.14	0
3	1	256	85	1265245	250.13	1
4	2	261	92	1625345	425.30	0
5	5	355	125	45245125	1245.23	1
6	4	304	112	12136527	730.58	0
7	5	355	125	45244525	1245.23	1
8	3	282	106	3028253	614.20	0
9	4	304	112	12133623	730.60	1
10	2	261	92	1625786	425.21	0
11	5	355	125	45245174	1245.23	1
12	5	355	125	45245177	1245.23	1
13	4	304	112	12136599	730.62	1
14	1	256	85	1265244	250.12	0
15	2	261	92	1625333	425.83	1
16	3	282	106	3025299	614.21	0
17	1	256	85	1265211	250.16	0
18	5	355	125	45245183	1245.23	1
19	5	355	125	45245100	1245.23	1
20	4	304	112	12136537	730.63	1
21	4	304	112	121365943	730.68	1

TABLE 4.1.4: Dataset

4.2 Statistical Techniques And Visualization:

Statistics is a collection of tools that you can use to get answers to important questions about data. You can use descriptive statistical methods to transform raw observations into information that you can understand and share. You can use inferential statistical methods to reason from small samples of data to whole domains. Statistics is a pillar of machine learning. You cannot develop a deep understanding and application of machine learning without it.

- Problem Framing: Requires the use of exploratory data analysis and data mining.
- Data Understanding: Requires the use of summary statistics and data visualization.
- Data Cleaning. Requires the use of outlier detection, imputation and more.
- Data Selection. Requires the use of data sampling and feature selection methods.
- Data Preparation. Requires the use of data transforms, scaling, encoding and much more. Model Evaluation. Requires experimental design and resampling methods.
- Model Configuration. Requires the use of statistical hypothesis tests and estimation statistics. Model Selection. Requires the use of statistical hypothesis tests and estimation statistics. Model Presentation. Requires the use of estimation statistics such as confidence intervals. Model Predictions. Requires the use of estimation statistics such as prediction intervals.

NumPy - is a commonly used Python data analysis package. By using NumPy, you can speed up your workflow, and interface with other packages in the Python ecosystem, like scikit-learn, that use NumPy under the hood. NumPy was originally developed in the mid 2000s, and arose from an even older package called Numeric. This longevity means that almost every data analysis or machine learning package for Python leverages NumPy in some way.

We'll walk through using NumPy to analyze data on wine quality. The data contains information on various attributes of wines, such as pH and fixed acidity, along with a quality score between 0 and 10 for each wine. The quality score is the average of at least 3 human taste testers. As we learn how to work with NumPy, we'll try to figure out more about the perceived quality of wine.

Numpy 2-Dimensional Arrays WithNumPy, we work with multidimensional arrays. We'll dive into all of the possible types of multidimensional arrays later on, but for now, we’ll focus on 2-dimensional arrays. A 2-dimensional array is also known as a matrix, and is something you should be familiar with.

Creating A NumPyArray :We can create a NumPy array using the numpy.array function. If we pass in a list of lists, it will automatically create a NumPy array with the same number of rows and columns. Because we want all of the elements in the array to be float elements for easy computation, we’ll leave off the header row, which contains strings.

One of the limitations ofNumPy is that all the elements in an array have to be of the same type, so if we include the header row, all the elements in the array will be read in as strings. Because we want to be able to do computations like find the average quality of the wines, we need the elements to all be floats.In the below code, we:

- Import the numpy package.
- Pass the list of lists wines into the array function, which converts it into a NumPy array.
- Exclude the header row with list slicing.
- Specify the keyword argument dtype to make sure each element is converted to a float.

Pandas - is an open source python library that is built on top of NumPy. It allows you do fast analysis as well as data cleaning and preparation. Pandas is hands down one of the best libraries of python. It supports reading and writing excel spreadsheets, CVS's and a whole lot of manipulation. It is more like a mandatory library you need to know if you’re dealing with datasets from excel files and CSV files. i.e for Machine learning and data science. This is part one of Pandas tutorial. I’m not going to cover everything possible with pandas, however, I want to give you a taste of what it is and how you can get started with it. This tutorial is going to be super short just introducing you to Series object of pandas.

As other libraries, you’d import pandas and reference it as pd.

import pandas as pd

pd.Series()

pd.Series() is a method that creates a series object from data passed. The data must be defined as a parameter. what is a “Series” object in Pandas? It is a data structure defined by Pandas. Basically it looks like a table having rows and columns. Notice that these numbers on the first column were added automatically by pandas. They serve as index. These variables are known as categorical variables and in terms of pandas, these are called ‘object’.

To retrieve information using the categorical variables, we need to convert them into ‘dummy’ variables so that they can be used for modelling. We do that using pandas.get_dummies feature.

First we create a list of the categorical variables. Then we convert these variables into dummy variables. We have created dummy variables for each categorical variables and printing out the head of the new data-frame. You can understand, how the categorical variables are converted to dummy variables which are ready to be used in the modelling of this data-set. But, we have a slight problem here. The actual categorical variables still exist and they need to be removed to make the data-frame ready for machine learning.

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. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged.SciPy makes use of Matplotlib. Matplotlib was originally written by John D. Hunter, has an active development community, and is distributed under a BSD-style license. Michael Droettboom was nominated as matplotlib's lead developer shortly before John Hunter's death in August 2012, and further joined by Thomas Caswell. As of 23 June 2017, matplotlib 2.0.x supports Python versions 2.7 through 3.6. Matplotlib 1.2 is the first version of matplotlib to support Python 3.x. Matplotlib 1.4 is the last version of Matplotlib to support Python 2.6.Matplotlib has pledged to not support Python 2 past 2020 by signing the Python 3 Statement. Pyplot is a Matplotlib module which provides a MATLAB-like interface.

Matplotlib is designed to be as usable as MATLAB, with the ability to use Python, and the advantage of being free and open-source. Several toolkits are available which extend Matplotlib functionality. Some are separate downloads, others ship with the Matplotlib source code but have external dependencies.

- **Basemap:** map plotting with various map projections, coastlines, and political boundaries.

- **Cartopy:** a mapping library featuring object-oriented map projection definitions, and arbitrary point, line, polygon and image transformation capabilities. (Matplotlib v1.2 and above)
- **Excel tools:** utilities for exchanging data with Microsoft Excel • GTK tools: interface to the GTK+ library.

Visualization With Matplotlib- One of Matplotlib's most important features is its ability to play well with many operating systems and graphics backends. Matplotlib supports dozens of backends and output types, which means you can count on it to work regardless of which operating system you are using or which output format you wish. This cross-platform, everything-to-everyone approach has been one of the great strengths of Matplotlib. It has led to a large user base, which in turn has led to an active developer base and Matplotlib's powerful tools and ubiquity within the scientific Python world.

In recent years, however, the interface and style of Matplotlib have begun to show their age. Newer tools like ggplot and ggvis in the R language, along with web visualization toolkits based on D3js and HTML5 canvas, often make Matplotlib feel clunky and old-fashioned. Still, I'm of the opinion that we cannot ignore Matplotlib's strength as a well-tested, cross-platform graphics engine. Recent Matplotlib versions make it relatively easy to set new global plotting styles (see Customizing Matplotlib: Configurations and Style Sheets), and people have been developing new packages that build on its powerful internals to drive Matplotlib via cleaner, more modern APIs—for example, Seaborn (discussed in Visualization With Seaborn), ggpy, HoloViews, Altair, and even Pandas itself can be used as wrappers around Matplotlib's API. Importing Matplotlib -Just as we use the np shorthand for NumPy and the pd shorthand for Pandas, we will use some standard shorthands for Matplotlib imports.

Plotting from a script- If you are using Matplotlib from within a script, the function plt.show() is your friend. plt.show() starts an event loop, looks for all currently active figure objects, and opens one or more interactive windows that display your figure or figures.

The plt.show() command does a lot under the hood, as it must interact with your system's interactive graphical backend. The details of this operation can vary greatly from system to system and even installation to installation, but matplotlib does its best to hide all these details from you.

4.3 Data Modelling And Visualization:

Node-RED is a flow-based development tool for visual programming developed originally by IBM for wiring together hardware devices, APIs and online services as part of the Internet of Things. Node-RED provides a web browser-based flow editor, which can be used to create JavaScript functions. Elements of applications can be saved or shared for re-use. The runtime is built on Node.js. The flows created in Node-RED are stored using JSON. Since version 0.14 MQTT nodes can make properly configured TLS connections.

In 2016, IBM contributed Node-RED as an open source JS Foundation project.

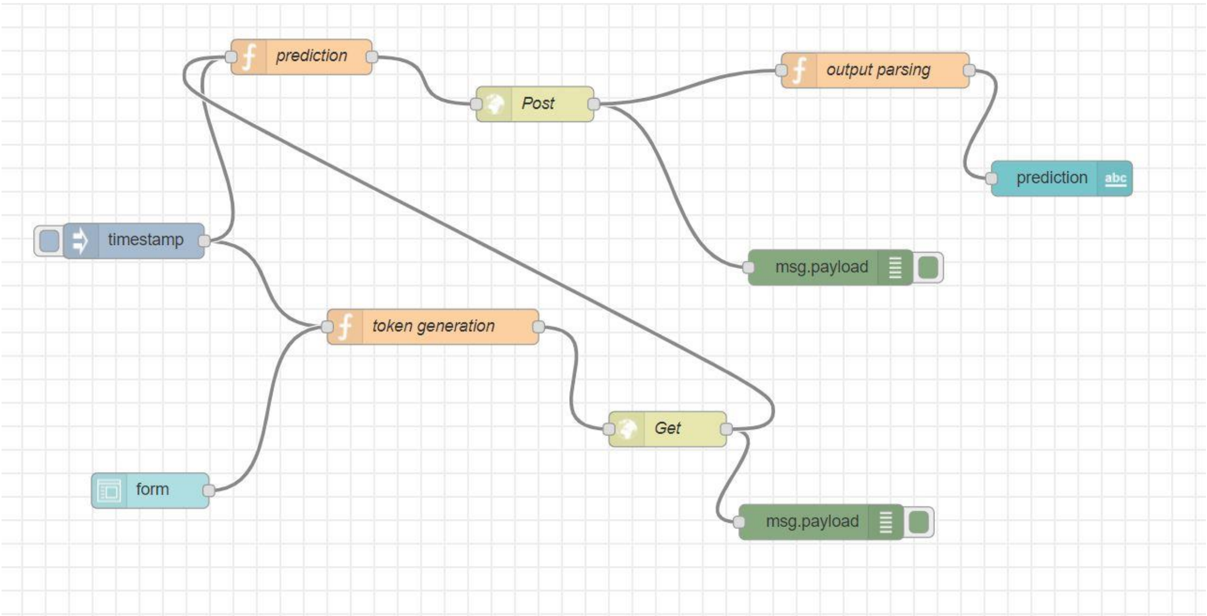
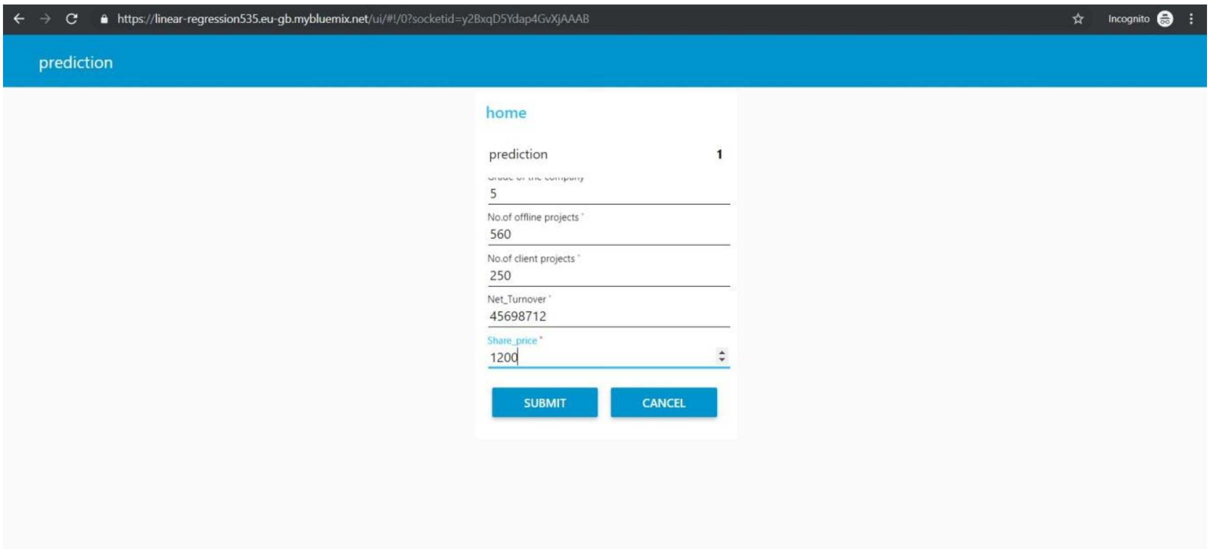


FIG-4.3.1: Flow Diagram


https://linear-regression535.eu-gb.mybluemix.net/ui/#!/0?socketid=y2BxqD5Ydap4GvXjAAAB
Incognito
prediction
home
SUBMIT
CANCEL

prediction 1

Grade of the company
5

No.of offline projects *
560

No.of client projects *
250

Net_Turnover *
45698712

Share_price *
1200

FIG-4.3.2: User Interface

5. FINDINGS AND SUGGESTIONS:

The Securities market emerges out to mobilize primary savings from the public to serve as a resource of funds by issusing shares and providing liquidity to these instruments through regular quotations in the financial markets and thus traded and pave the path for wealth creation. Milloins of investors are backbone of Indian Securities market. The Bombay Stock Exchange (BSE) and National Stock Exchange (NSE) are the leading stock exchanges in India.BSE has the distinction of being largest in the country.

Securities market is known for tax-free returns, an effortless, easy entry into the stock market, higher returns, any time liquidity and to deliver higher real returns than any other investments. Based on the findings, following suggestions are given to investors, brokers and SEBI to overcome the problems faced by them.

6. CONCLUSION:

Stock market prediction is important factor in finance. It is considered to be dynamic in nature. The paper presented how to predict stock values based on the data using Machine Learning algorithms: Decision Tree. We also concluded this algorithm is better than the other algorithms because, within a certain range, the difference between actual price and predicted price is quite small as compared to those in Logistic Regression and Random Forest. Also, Random Forest is better than Logistic Regression, but inferior to MLP and Decision Tree, in predicting stock values