Stock Market Prices Prediction using Random Forest and Extra Tree Regression

Subba Rao Polamuri, K. Srinivasi, A. Krishna Mohan

Abstract: ***Prediction of Stock price is now a day's an existing and interesting research area in financial and academic sectors to know the scale of economies. There did not exists any significant set of rules to estimate and predict the scale of share in the stock exchange. Many evolutionary technologies are existing such as technical, fundamental, time, statistical and series analysis which help us to attempt the prediction process, but none of the methods are proved as reliable and accurate tool to the society in the estimation of stock exchange or share market scales. Here in this paper we attempted to do innovative work through Machine Learning approach to predict or sense the behaviour tracking of the stock market sensex. Linear regression, Support Vector regression, Decision Tree, Ramdom Forest Regressor and Extra Tree Regressor are the Machine Learning models implemented effectively in predicting the stock prices and define the activity between the exchanges the securities between the buyers and sellers. We predicted the price of the stock based on the closing value and stock price. An algorithm with high accuracy we do the process of comparison for the accuracy of each of the model and finally is considered as better algorithm for predicting stock price. As share market is a vague domain we cannot predict the conditions occur, and also share market can never be predicted, this job can be done easily and technically through this work and the main aim of this paper is to apply algorithms in Machine Learning in predicting the stock prices***.

Keywords: **Decision Tree, Extra Tree repressor, Multi-Variate Linear Regression, Random Forest ,Stock Market.**

I. INTRODUCTION

Stock Market forecasting is an act of analysing a trying to diagnose the market worthiness of the stock existed in the company; it acts like a reliable instrument of financial growth of the company to trade on an exchange. Commerce and trade are the two economical elements that play a vital role in the evolution towards economic of the nation in wide range such as industry market and investors. In this process we can easily predict the value of the stock in case of price raise and praise fall at any period of time. Stock marketing [1,2] is the primary source of any industry whether it is private sector or public sector to raise their funds of business expansions and also further growth of the company. The best actors are investors and industry involved in this stock exchange process of their securities. This is based on the pure concept of economic policy of demand and supply. For example if the demand of stock of the particular company decreases[4,7], there always the fall in the price of the share of that particular company.

Efficient market hypothesis is the technical theoretical and experimental challenges are the motivational for getting efficient results of marketing. The stock prices completely reflect variable information about the constituents and the opportunities foe earning abundant profits. New York Stock[3] Exchange is one of the successful stock exchanges worldwide[8]. It is the world's No: 1 stock exchange that gives the reliable services those who seek. Many industries and companies now a day's involving in this process. This contains huge sets of data which is difficult to extract, analyze and extract information. This is a big task to the users on the manual process. The market pattern and the prediction of time of purchase of stock is revealed by the analysis of the stock market. Significant profits can be achieved if there is a successful prediction process taken place. The historic data of the market represents varying conditions and helps in confirming the time series pattern has statistically significant predictive power and has high possibility of profitable trades and returns in the investment for competitive business.

II. RELATED RESEARCH

Benchmark for comparison with LS-SVM and LS-SVM-PSO models are discussed by Osman Hegazy, Omar S.Soliman and Mustafa Abdul Salam discussed Levenberg-Marquardt [1] (LM). The obtained results showed that the proposed model has better prediction accuracy and the potential of PSO algorithm in optimizing LS-SVM. L. Minh Dang, Abolghasem Sadeghi-Niaraki, Huy D. Huynh, Kyungbok Min And Hyeonjoon Moo [2] Compared by taking set of indexes. Experiments are conducted to check whether summarization help the stock prediction and acquired positive result. Christian Slamka, Bernd Skiera, and Martin Spann[3] comes the deep learning technique which uses the data preprocessing, document labelling, finding indicators and word embedding to predict the price. Later the accuracy is found and comparison with various indexes is also done. This can be applied to short term analysis for better results.

Manuscript published on 30 September 2019
* Correspondence Author

Subba Rao Polamuri*, Research Scholar, Dept Of CSE, University college of Engineering, JNTU Kakinada, East Godavari, AP, INDIA.. Email: psr.subbu546@gmail.com.

Dr. K. Srinivasi, Professor, Dept Of CSE, V R Siddhartha Engineering College, Vijayawada, INDIA. Email: vrdrks@gmail.com.

Dr. A. Krishna Mohan, Professor, Dept Of CSE, University college of Engineering, JNTU Kakinada, East Godavari, AP., INDIA. Email: krishna.ankala@gmail.com.

Joseph St. Pierre, Mateusz Klimkiewicz, Adonay Resom and Nikolaos Kalampalikis[4] have compared different parameters like usability, implementation effort, general properties and possible security designs. Mustafa Gocken, Mehmet Ozcalıcı, Aslı Boru and Ayse Tugba Dosdogru [5] considered attributes like market history, commodity price which are normalized between [+1,-1].

The NEWS and twitter feed for each day was gathered and processed which are to be declared as Positive or Negative. The results obtained from the Multi-Layer Perceptron algorithm of machine learning have predicted 77% market performance. Tae Kyun Lee, Joon HyungCho, Deuk SinKwon and So YoungSohn[6] have experimentedand stated that Random forest model gave 54.12 accuracy among the here models which are used by author for stock market prediction. Kang Zhang, GuoqiangZhong, JunyuDong, ShengkeWang and YongWang [7] have experimented and explained how Regression involves predicting dependent variables using independent variables. Yi-Fan Wang [8] have experimented how to predict the market performance of Nifty stock index dataset by using some well defined technical analysis, fundamental analysis, time series analysis and statistical analysis, etc. and proposed model gave 96.6% accuracy. Pei-Yuan Zhou, Keith C.C. Chan, and Carol Xiaojuan Ou [9] have experimented with Support Vector Machine is used to train data which includes Support Vector Regression, Loss function and Kernel Function. On completion of training data implementing prediction algorithm is done.

III. THEORETICAL BACKGROUND

Stock value forecast is a work of art and significant issue. With a fruitful model for stock expectation, we can pick up knowledge about market conduct after some time, spotting patterns that would some way or another not have been taken note. With the undeniably computational intensity of the PC, AI will be an effective strategy to take care of this issue.

A. Background of Problem

Securities exchange is profoundly unstable. At the most central level, it is said that free market activity in the market decides stock cost. Be that as it may, it doesn't pursue any fixed example and is additionally influenced by an enormous number of profoundly changing components the financial specialists on the Wall Street are part in two biggest groups of followers; the individuals who accept the market can't be anticipated and the individuals who accept the market can be beat.

B. Related research to solve problem

As of late, a great deal of fascinating work has been done in the territory of applying Machine Learning Algorithms for investigating value examples and foreseeing stock cost. Most stock dealers these days rely upon Intelligent Trading Systems which help them in anticipating costs dependent on different circumstances and conditions. Late looks into utilization input information from different sources and numerous structures. A few frameworks utilize recorded stock information, some utilization money related news stories, some utilization master audits while some utilization a mixture framework which takes different contributions to anticipate the market. Likewise, a wide scope of AI calculations are accessible that can be utilized to plan the framework. These frameworks have various ways to deal with tackle the issue. A few frameworks perform scientific investigation on notable information for expectation while some perform assumption examination on money related news stories and master surveys for forecast.

IV. METHODOLOGIES

A. Linear Regression

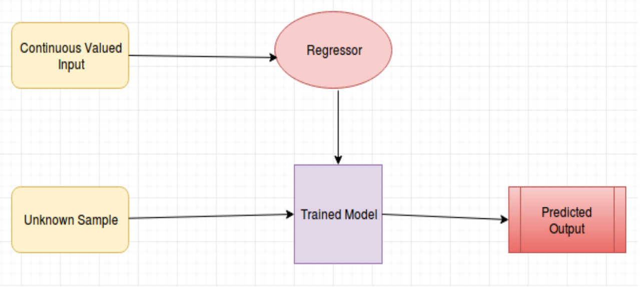

Fig 1. General data Flow diagram of a linear regression model

Linear Regression comes under the simplest type of supervised learning. The goal of this linear regression is to explore the relation between the input feature with that of the target Value and give us a continuous Valued output for the given unknown data. Machine Learning means teaching the machine how to find patterns in data[5]. To extract those and manipulate them. Linear regression finds the pattern in the data and shows us an estimation or prediction.

Y=W1*X + b

1. Y=Predicted value/Target Value
2. X=Input
3. W1=Gradient/slope/Weight
4. b=Bias

The equation is same as that of a straight line (Y=MX+c)

What the question arises what is this W1 and b?

For now let’s say they are the parameters to adjust the straight line to get the best fit. By adjusting the W1 and b we get the algorithm to get the most optimized results.

Algorithm:

Step1: Collect the data.

Step2: Define Hyper parameter i,e. alpha value.

Step3: Calculate W1 and b values using gradient descent function.

Step4: Calculate the Y value using equation Y=W1*X + b.

B. Multi – Variate Linear Regression

Multivariate regression model technique is used. Here we can have more number of variables while prediction, hence the model is called to be multivariate multiple regression. For researchers, who are doing research on regression techniques are more popular and simplest supervised machine learning algorithms. Here a set of explanatory variable are there to predict the response variable, this regression algorithm is used. With the help of matrix operations and PYTHON implementations, “numpy” library this regression technique can be implemented efficiently, which is successfully used and results are achieved by performing the matrix operations which has extensional definitions and related operations for matrix object. The equation of multi-variate linear regression for **n** independent variables are given as

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n$$

Here x is the input and Y is the predicted component of the dependent variable y. Multivariate Regression algorithm for Industry application used heavily in the retail sector

In this retail trade sector customers make a decision based up on number of variables such as label, cost and type of the article. This algorithm well suited in the areas of product selection and find the best collection of factors [6]. This algorithm acts as an decision making agent in selecting the best among the products.

Algorithm:

Step1: Collect the data.

Step2: Define Hyper parameter i,e. alpha value.

Step3: Calculate $\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n$ values using gradient descent function.

Step4: Calculate the Y value using equation $Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_{n.}$

C. Support - Vector - Regression

A To maintain all features Support Vector Machine can also be used as a regression method, which divides depending up on the distinctive qualities the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor dissimilarities. It gives the result in terms of real number, and also infinite number of possibilities in prediction [9], so it is very difficult the information on the tips of fingers. There is also a more complicated reason, this regression has a "epsilon" the margin of tolerance is set in the process of conjecture to the SVM which is already formally requested. Apart from this fact, this is to be taken in consideration that the algorithm is more tangled. Whatever it may be the main theme is always remains to maximizes the margin, that to reduce error as much as possible, individualizing the hyper plane. Error is tolerated by keeping all these things in mind.

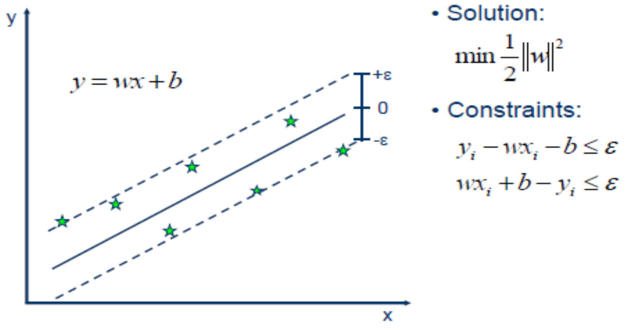
y
$y = wx + b$
+ε
0
-ε
x
• Solution:
$\min \frac{1}{2}\|w\|^2$
• Constraints:
$y_i - wx_i - b \le \varepsilon$
$wx_i + b - y_i \le \varepsilon$

Fig 2. SVM Graph

D. Decision Tree

Decision tree developed regression in the form of tree structure, decision tree is incrementally developed, in this classification; it shatters dataset into as much as possible minute subsets at the same time. The end result is a tree having decision **nodes** and **leaf nodes**. The tested attributes are represented in the decision tree with two or more branches, each representing values for the tested attributes. Numerical target is represented in the Leaf node. The best clairvoyant called **root node** is represented in the highest decision node in a tree which correlates the decision trees can handle both categorical and numerical data.

ID3 by J. R. Quinlan invented the main algorithm for developing decision trees which inculcates a top-down approach, greedy search approach along with the space of possible branches which has no backtracking. Decision tree for regression uses the ID3 algorithm to build the tree by repositioning information gain with Standard Deviation Reduction (SDR).

Algorithm:

Step 1: target is mathematically determined using standard deviation

Step 2: attributes are splitted into distinctive attributes. Then the standard deviation for every branch of nodes is calculated. Before the split, the resultant standard deviation is the minus value from the standard deviation. Thus the result we get is the standard deviation reduction.

Step 3: The attribute which is having relatively greatest standard deviation reduction is selected for the decision node.

Step 4a: depending up on the resultant values of the selected attribute, the dataset is divided. This process continues repeatedly on the non-leaf branches, until all data is computed.

The algorithm works for instance, when coefficient of deviation (**CV**) for a branch becomes smaller than a certain threshold like 10% or when too few instances (**n**) remain in the branch, we need some termination criteria.

Step 4b: If because CV is less than the threshold,. Subset of dataset does not need any further splitting. Then the related leaf node has the average of the subset dataset.

Step 4c: However, further splitting is need when the branch has an CV more than the threshold.

Step 5: we stop further branching process if the number of data points for all branches is equal or less than the number of branches and to the related leaf node is assigned the average of each branch.

V. PROPOSED METHODOLOGIES

In the numerical data, first we pre-process it to eliminate missing values, if any. Later the data set is split into training and Test data. We apply machine learning models like linear regression, multivariate regression, SVM and decision tree regression algorithms and final output is predicted by using these models. To get large financial payoff, strong predictive models in stock market data is interesting to analyze and taken as a further incentive. The financial data on the web is huge and endless. Companies can be hard to come by A large and well structured dataset on a wide array. Last 5 years historical stock prices are taken as dataset members for all companies currently found on the S&P 500 index. From these the price of the stock for the sixth day will be predicted. Given explain the Architecture, What is our problem, How to solve that best training parameters (training plan) and implementation of model.

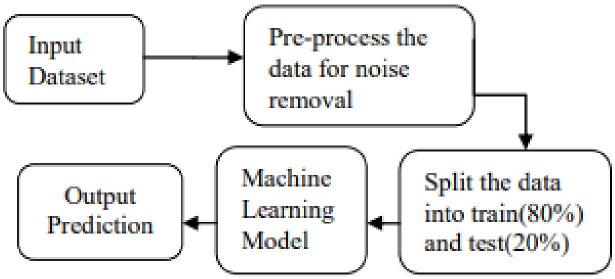
Input Dataset
Pre-process the data for noise removal
Split the data into train(80%) and test(20%)
Machine Learning Model
Output Prediction

Fig 3. Numerical dataset Algorithm Architecture

A. Random Forest Regressor

The ramdom forest means data about data estimator. It fits a number decision tress on various sub samples of the given data. It control over-fitting. It improve the predictive accuracy.

Algorithm:

Step 1: From the dataset pick N random records.

Step 2: Based on N records, build a decision tree.

Step 3a: From your algorithm, choose the number of trees and repeat steps 1 and 2.

Step 3b: In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output).

B. Extra Tree Regressor

Extremely **ra**ndomized **Tree**s means Extra-Tree method. In the context of input features(numerical), choice of optimal cut-point is responsible for a big portion of the variance of induced tree. This is the main objective in randomizing tree building.

In point of Statistical view, dropping the bootstrapping idea gives an advantage in terms of bias. The cut-point randomization usually gives excellent variance reduction effect. Many high-dimensional complex problems give best results using this method. Extra-Tree method produces piece-wise multilinear approximations in the view of functional point, rather than the piece-wise constant ones of random forests.

C. *Our Problem Process*

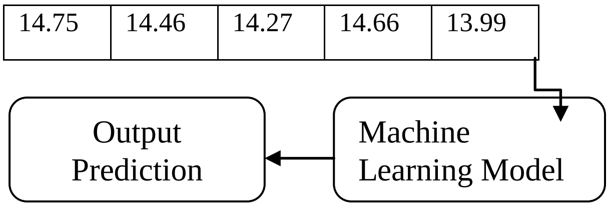

Fig 4. Theme for Numerical Data

TABLE- I: TRAINING PLAN

Model Name	Linear Regression	Multivariate Regression	Radom Forest	Extra Tree Regresser
Input data shape	(619040, 7)	(619040, 7)	(619040 , 7)	(619040, 7)
Train size	80%	80%	80%	80%
Test size	20%	20%	20%	20%

VI. EXPERIMENTAL RESULTS

A. Linear Regression

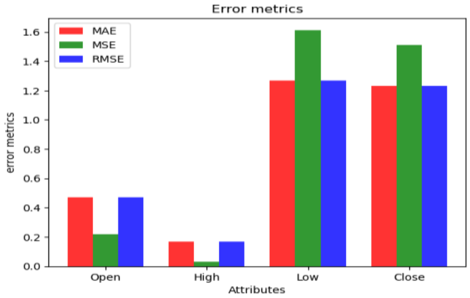

Fig 5. Error Metrics for Linear Regression

B. Multi-Variant Linear Regression

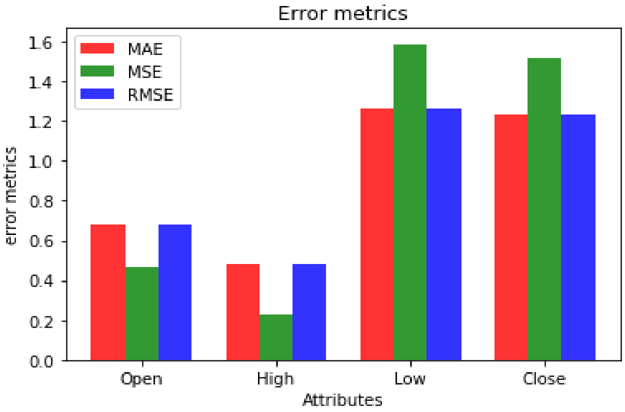

Fig 6. Error Metrics for Multivariate

C. Random Forest Regressor

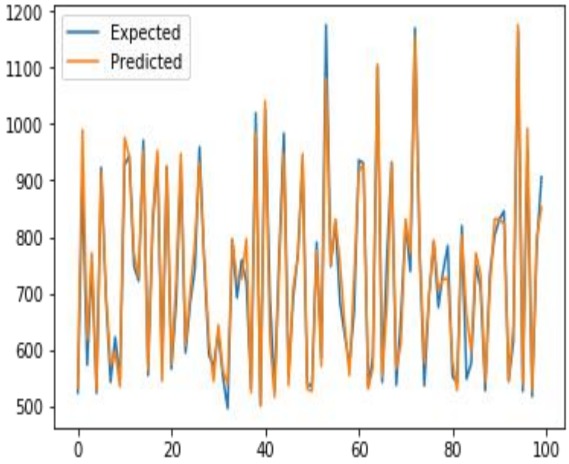

Fig 7. Prediction Graph for Decision Tree

Retrieval Number C4314098319/19©BEIESP
DOI: 10.35940/ijrte.C4314.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

D. Extra Tree Regressor

MAE: 453.082191781
MSE: 477185.787671
RMSE: 690.786354578

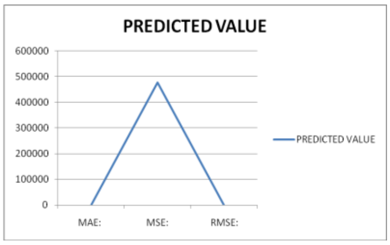

Fig 8. Prediction graph for Regressor

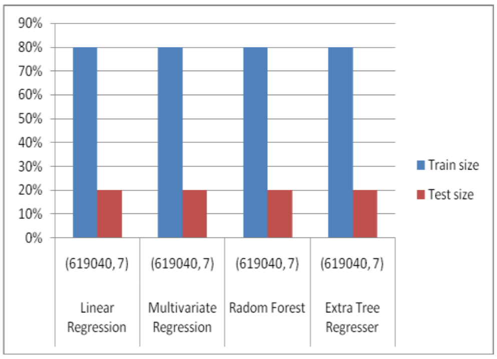

Fig 9. Tested Results

VII. CONCLUSION

After all the inclusions of results, applied Machine Learning algorithms. We conclude that all these algorithms are good each to give good prediction of stock price but Decision Tree and Random Forest Regressor are best regression algorithm among them after comparing the results. The whole process was done in the context of machine learning. The algorithms and system which are traditional systems may not efficiently solves problems associated with this huge amount of data and may leads to the systems run very slowly and cannot yield the best and accurate result of prediction. But with the help of Python environment, we can handle large data very efficiently without alternating the methods in the existing procedures. Experimentation and results are achieved with the help of Numerical data using Python. Future work can be proceeded using deep learning algorithms like Long Term Memory (LTM), Short Term Memory (STM), CNN as further implementation of this work.

REFERENCES

1. Osman Hegazy et al, "A Machine Learning Model for Stock Market Prediction", Volume 4, Issue 12,December 2013.
2. L. Minh Dang et al, "Deep Learning Approach for Short-Term Stock Trends Prediction based on Two-stream Gated Recurrent Unit Network", Volume: 6, Issue: 2, September 2018.
3. Christian Slamka et al, "Prediction Market Performance And Market Liquidity: A Comparison of Automated Market Makers"Volume:60, Issue:1,Date: April 2012.
4. Joseph St. Pierre et al, "Trading the stock market using Google search volumes: a long short-term memory approach", Volume: 3, Issue: 1, Date: 2019.
5. Mustafa Goçken et al, "Stock price prediction using hybrid soft computing models incorporating parameter tuning and input variable selection", Volume:31, Issue: 2, Date 2019.
6. Tae Kyun Lee et al, "Global Stock Market Investment Strategies Based On Financial Network Indicators Using Machine Learning Techniques", Volume: 117, Issue: 1, Date: 2019.
7. Kang Zhang et al, "Stock Market Prediction Based on Generative Adversarial Network", Volume:147, Issue:2, Date:2019.
8. Yi-Fan Wang," On-Demand Forecasting of Stock Prices Using a Real-Time Predictor", Volume: 15, Issue: 4, Date: 2003.
9. Pei-Yuan Zhou, et al, "Corporate Communication Network and Stock Price Movements: Insights from Data Mining", Volume:5, Issue: 2, Date: 2018.
10. Pei-Chann Chang et al," Integrating a Piecewise Linear Representation Method and a Neural Network Model for Stock Trading Points Prediction", Volume: 39, Issue:1, Date:2009.
11. Min-Wen et al," Stock Market Trend Prediction Using High-Order Information of Time Series", Volume: 7, Issue: 4,Date:2019.
12. Liming Zhang et al," A Novel Instantaneous Frequency Algorithm and Its Application in Stock Index Movement Prediction", Volume: 6, Issue:4, Date:2012.
13. Mehak Usmani et al," Stock Market Prediction Using Machine Learning Techniques"Pankajkumar and Dr.AnjuBala," Intelligent Stock Data Prediction using Predictive Data Mining Techniques".
14. Han Lock Siew and Md Jan Nordin," Regression Techniques for the Prediction of Stock ", 2012 ICSSBE, IEEE.
15. Polamuri Subba Rao, Dr. K.Srinivas, Dr.A. Krishna Mohan, A Survey on Stock Market Prediction using Machine Learning Techniques, ICDSMLA, June 2019.

AUTHORS PROFILE

Mr. Subba Rao Polamuri, Research scholar, currently perusing Ph.D in computer science and Engineering at Jawaharlal Nehru Technological University, Kakinada, East Godavari, Andhra Pradesh.. My areas of research include Bigdata, Data Mining, Artificial Intelligence, Machine learning, and Deep Learning

Dr.Kudipudi Srinivasi, done his Ph.D in CSE,M.Tech (CSE), Presently Working as Professor in department of CSE at VR Siddhartha Engg College, Vijayawada, Andhra Pradesh, India. He is eminent faculty member Ratified by JNTUK in CSE, and also recognized Research Supervisor of JNTUK and his areas of interest are Image Processing, Bio Informatics, Data Mining, Bigdata, Artificial intelligence, deep learning and Machine learning.

Dr.A.Krishna Mohan, Professor, Department of Computer Science & Engineering Jawaharlal Nehru Technological University, Kakinada. He is an eminent personality in guiding more no of students of B.Tech, M.Tech, and Ph.D. scholars in and affiliated colleges of JNTU Kakinada and his areas of interest are Data Mining, Bigdata, Artificial Intelligence, deep learning and Machine learning.