**Project Report**

**On**

**Forthcoming Stock Prediction Algorithm**



***Bachelor of Computer Application***

**Final Year (Vth Semester)**

**Of**

**AMITY UNIVERSITY UTTAR PRADESH**

**LUCKNOW CAMPUS**

**SUBMITTED TO: DEVELOPED BY:**

**MR. UMAR KHALID FAROOQUI ANMOL SINGH**

**(A7304816019)**

**DECLARATION**

I **Anmol Singh**, student of BCA hereby declare that the project titled “**Algorithm for predicting future stock market using Machine learning**” which is submitted by me to Amity Institute of Information Technology, Amity University Uttar Pradesh, in partial fulfillment of requirement for the award of the degree of BCA has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

Signature

Lucknow

Date Name and Signature of Student(s)

**CERTIFICATE**

On the basis of declaration submitted by **Anmol Singh**, student of BCA 5 hereby certify that the project titled “**Algorithm for predicting future stock market using Machine learning**” which is submitted to Amity Institute of information Technology, Amity University Uttar Pradesh, Noida, in partial fulfillment of the requirement for the award of the degree of BCA is an original contribution with existing knowledge and faithful record of work carried out by him under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

|  |  |
| --- | --- |
| Date | (Name and Signature of Guide) |

Amity Institute of Information Technology

Amity University Uttar Pradesh

***Acknowledgement***

Any project requires dedication, direction, support of many individuals and hours of brain storming to pull out something intriguing and give a human brain a new possibility towards an empty space where it never went before.

With the above motivation and intention, this project took its shape and I am highly obligated to receive my mentorship from *Mr. Umar Khalid Farooqui*.

Working under his roof of guidance, supervision made me to sail this journey of project towards its successful completion.

My gratitude also extends for all those individuals who helped me by pooling their resources and knowledge. I hope this project worth my effort and for your time.

Thanking You

Anmol Singh

BCA 5th Sem

A7304816019

***Contents***

|  |  |  |
| --- | --- | --- |
| Sr No. | Content Name | Page No. |
| 1. | Title of Project |  |
| 2. | Acknowledgement |  |
| 3. | Abstract |  |
| 4. | Objective |  |
| 5. | MACHINE LEARNING- Introduction |  |
| 6. | Types of ML Algorithm |  |
| 7. | Methodology |  |
| 8. | How these algorithm implemented |  |
| 9. | Feasibility Study |  |
| 10. | Objective |  |
| 11. | Discussion |  |
| 12. | Conclusion & Future Scopr |  |
| 13. | References |  |
| 11 | Code & Output |  |

***Algorithm for predicting future stock market using Machine Learning***

***Abstract***

***The stock market in today’s world is like an oxygen supplier for any country or company’s survival. Like oxygen, if the stock market is depleting for any company then its future survival becomes questionable. Stocks are believed to be very dynamic and if one wants to predict the market the various wary which needs to take care off makes it prone to human error. Even if they succeed, they need to provide a considerable amount of empirical evidence to support their argument, which makes this task very tedious.***

***In the recent development of machine learning or computer learning, it really makes it easier for anyone to predict the weather of the stock market.***

***Machine learning is teaching a handmade machine model by encoding it with an algorithm that uses a large amount of data set to gain its learning for doing a specific for which it is made.***

***In this project, the algorithm uses the complete data set of Microsoft’s stock performance from its origin until current day from renowned data set provider QUANDL.***

***In addition, we train and test our model on that data set with a ratio of 70:30(train and test) respectively and we re-iterate the process of training and testing until and unless we do not get the accuracy of prediction 99%.***

***Machine Learning***

**“Machine learning** is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed**”[1]**

Machine learning can also be defined as “programming computers to optimize a performance criterion using example data or past experience. We need learning in cases where we cannot directly write a computer program to solve a given problem, but need example data or experience”[2].

In amass population, people most frequently are ambiguous between Artificial Intelligence, Deep Learning, and Machine Learning. Artificial Intelligence is like a parent, while machine learning is a child of AI and added deep learning is a subcategory of machine learning.

Today, Machine learning is everywhere and recently its developing so fast that we might get habitual to it without even our realizing, for example,

* Face recognition
* AI voice
* Pattern recognition
* The custom built recommendation system

**Types of Machine learning algorithm**

Machine learning algorithm is broadly distributed in **supervised, semi supervised** and **unsupervised**.

**Supervised algorithms**

“Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance”[3].

Supervised learning can be implemented for problems, which involves classification and regression.

* **Classification**: A classification problem is one in which the output variable result to be a part of the specific category, such as “animals”, ”humans”, “object” etc.
* **Regression**: A regression problem is one applicable when the output variable is resulting in a real number like, “height”, ”INR”.

**Unsupervised Machine Learning**

“Unsupervised learning is where you only have input data (X) and no corresponding output variables.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

These are called unsupervised learning because unlike supervised learning above, there are no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.”[3]

Unsupervised algorithms are often helpful in dealing with problems which fall in these:

* **Clustering**: A problem is dealt with the help of clustering when we want to find out the hidden group in the given dataset. For example, finding a group of customers with the same buying pattern behavior.
* **Association**: This algorithm is used when you want to learn and find out what are the rules that govern for the humongous amount of data, for instance, customers who buy shoes and most probably will buy socks too.

**Semi-Supervised Machine Learning**

“Semi-supervised learning uses both labeled and unlabeled data to perform an otherwise supervised learning or unsupervised learning task. It is also defined with words like Learning from labeled and unlabeled data, transductive learning”[4].

These problems in which semi supervised is used, sit in between both supervised and unsupervised learning.

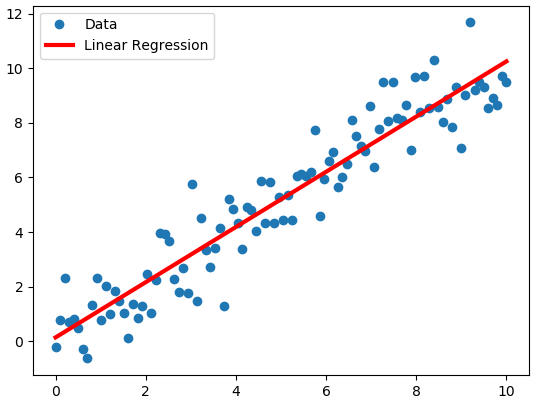
Many real world machine-learning problems fall into this area. This is because it can be expensive or time-consuming to label data as it may require access to domain experts. Whereas unlabeled data is cheap and easy to collect and store.[3]

You can use unsupervised learning techniques to discover and learn the structure in the input variables.[3]

**Methodology**

**Linear Regression Algorithm**

Linear regression attempts to model the relationship between two variables by setting a linear equation to the observed data one variable is considered an explanatory variable(X) while the other is considered to be a dependent variable(Y).



The dependent variable is the variable whose values we want to explain or forecast. The independent or explanatory variable is a variable that explains the other variables and the values are independent. So the dependent variable can be denoted as Y.

So if Linear Regression is to be defined in a line it will be to establish a relation between two variables or see if there is a statistically significant relationship between the two variables. For instance, if we want to see how raise in taxes has an effect on how many cigarette packs consumed, how many sleep hours versus discourse or Pokémon versus of urban density.

The second application is to forecast new observation; we can use what we already know to forecast unobserved value, for example, the value of the stock over the time, or value of fidget spinner over the time.

So the linear regression can think off as **“line of best fit”.** And this line of best fit can be represented by the linear equation

Y=A+BX

Or

Y=B\_0+B\_1\*X

Or

Y=MX+B

Here, B is the intercept

M is the slope or the gradient

**B is the intercept if we change this variable the intercept goes down or up along the Y axis.**

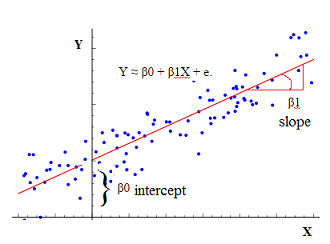
**M is the slope or the gradient if we change this, then the line rotate along the intercept.**

So data is a series of X and Y as shown in a scatter plot they don’t follow a straight line, however, they do follow a linear pattern hence the term **Linear Regression.**

Let us assume that we have the best-fit line already, we can calculate the error term **Ɛ –epsilon** known as residual, and this would be the term, that we would like to minimize along all the points in our data series.

Let us we have our linear equation represented in statistical notation.

**Y = B\_0 + B\_1 \* X + Ɛ**

****

**Support Vector Machine regression algorithm**

Support vector machine (SVM) is best suited for extreme cases. It looks at the extremes of the dataset and draws a decision boundary, also known as the **HYPERPLANE.** So SVM is a frontier which best segregates the two classes (Hyperplane/Line).

**How does it work**

So let us say we have some sample data or features to classify whether the given data belongs to which class X or Y. So we create a decision boundary between X and Y, which will decide whether the observed data belongs to X or Y. So decision boundary is the arbitrary separation line and uses intuition to draw it in between the most extreme final value of each corresponding to each other and find the best or optimal decision line which differentiates both the class X and Y.

The most extreme value corresponding to each other are chosen and called as data points or support vectors or points that are close to the opposing class.

So the algorithm implies that only support vectors are important whereas other training examples are ignorable.

Y1

Y19 vvvY1 Y19

X11 X5 X7

X4

X2 X3

* Where X5,x7,x4,x2,x3,x11 are data sets belonging to X class
* And y1 and Y19 are data set belonging to Y class.
* Y19 and X11 are appointed as the data points or support vectors or points that are close to the opposing class.
* And the middle line in between Y19 and X11 is called HYPERPLANE

The middle line is called Hyperplane because SVM can be used in the multi-dimensional dataset and the data points are referred to as vectors as they coordinate within the space of data.

So what we discuss so far is called LINEAR SUPPORT VECTOR MACHINE (LSVM) because classes are linearly separable.

**How these algorithms are used in this Project**

Linear Regression

Linear Regression is used to form a linear equation between the given data coordinates by obtaining the explanatory variable(X1) and the dependent variable(Y).

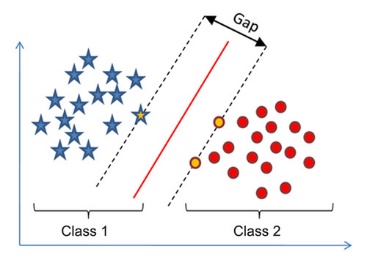
Likewise, in the algorithm, the same process is performed where we split the data obtained from X1 and Y into training the model and making it fit, for its score in testing.

After getting the result from linear regression, we check the score of our algorithm or accuracy of our algorithm.

**Linear Support Vector Machine(LSVM)**

LSVM believed to be the most accurate of all algorithm developed. We use support vector algorithm to double check our accuracy given by our Linear Regression.

In addition, change our approach to make our algorithm work with more accuracy by making amendment in our choosing attributes used in LR model and removing if there is any NAN value or we are choosing wrong attributes for our algorithm to work.



Feasibility Study

The existence of this algorithm would still be a dream without the technical support that leads to the accomplishment of this milestone of scoring 99% accuracy in the test of the algorithm.

These honorable technical feasible supports, which made this project feasible, are-

* Quandl- for providing a dataset of Microsoft’s stock from its origin until today.
* Anaconda3- to provide jupyter notebook as the workspace.
* python, numpy, matplotlib, interpolation.
* Algorithms- Linear Regression, Linear Support Vector Machine.

Objective

The primary objective of writing this algorithm is first off gaining the knowledge about machine learning. Machine learning’s basic terminologies like supervised, unsupervised and semi supervised. Learning about Algorithms, which make machine learning field a broiling topic to learn. Also in developing a new technology like Artificial Intelligence.

Machine Learning’s application is much broader than one could imagine for instance, “Reinforcement learning, which is learning to learn”, “in the prediction of having Cardiac Arrest 2min before happening”, “recognizing breast cancer in its early stages”.

These all motivated me in taking this project in a new direction, which is “future prediction”. In addition, I applied this potential future implication in stocks, because every day amasses population on this earth invest in stocks, and worried as too what will happen to their invested asset. Therefore, my algorithm can be a great arsenal to have in hands and invest into the stocks without much of a worry.

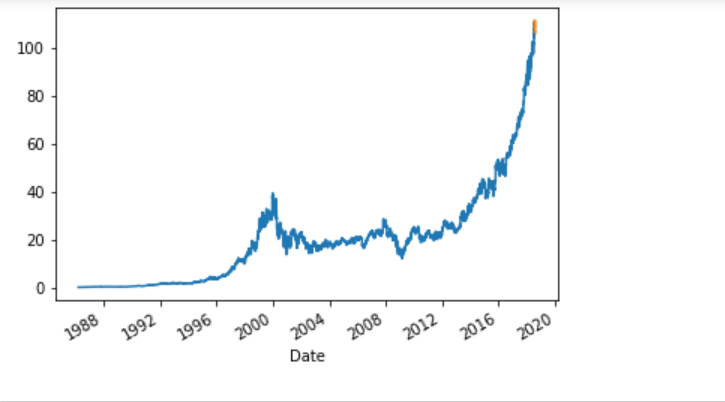
Discussion

*End Result Comparison*

This algorithm is supposed to give an accurate result until some catastrophe happens within the company, which leads to the sudden bankruptcy of its entire asset.

Comparison of prediction initiate from the start of the company until today, and with the promising close result in the prediction of stocks.

Microsoft Legitimate Stocks:

Algorithm’s prediction



Conclusion and future scope

My project used machine learning in predicting the future stock market. This knowledge of machine learning proved to a boon in developing this algorithm.

And my algorithm is showing the profitable result which in agreement when it is compared with the real time stocks of the company, and this algorithm can be used in future prediction.

Future scope of this algorithm is enormous in application to a person’s daily life of work. We can make this algorithm available for the normal user by developing an Android/IOS application for this algorithm.

Moreover, making this android app show the user prediction for upcoming days and weeks or months instead showing it for upcoming years.

**References**

[1]Samuel, Arthur (1959).

"Some Studies in Machine Learning Using the Game of Checkers".

IBM Journal of Research and Development.

3 (3): 210–229. doi:10.1147/rd.33.0210

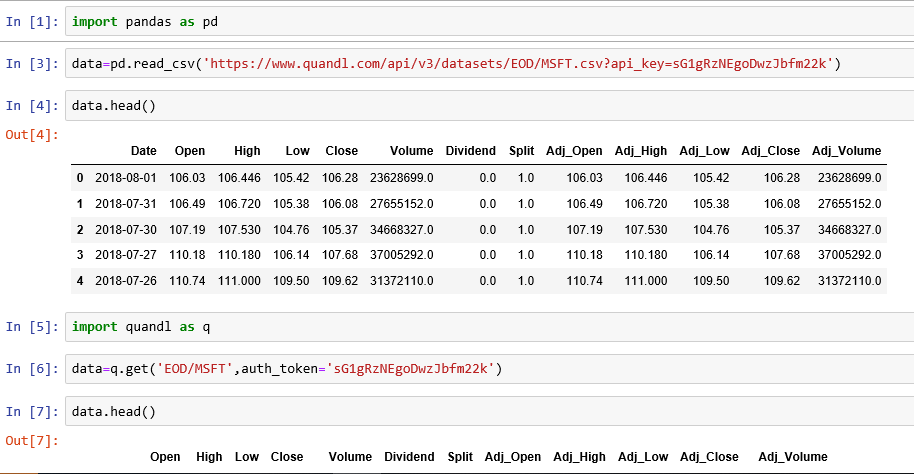
[2]Introduction to Machine Learning By Ethem Alpaydin

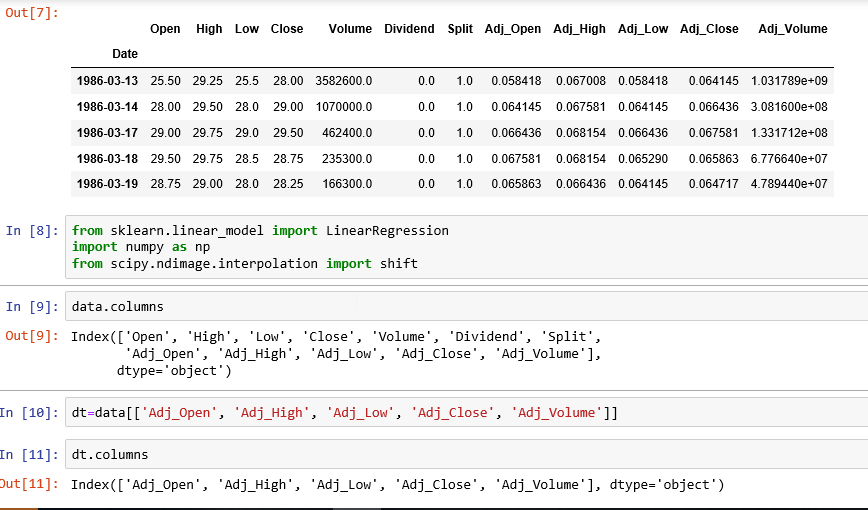
[3]https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

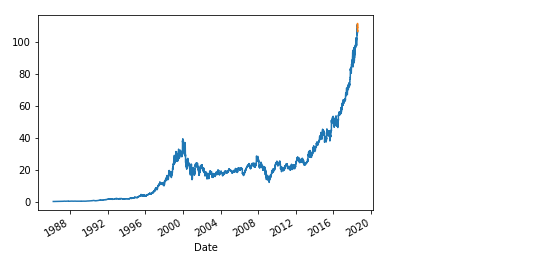
[4]Semi-Supervised Learning

Xiaojin Zhu, University of Wisconsin-Madison

**Code & Output**

****

****

****

ntcc

*by* Anmol Singh

**Submission date:** 20-Aug-2018 12:04PM (UTC+0530)

**Submission ID:** 991439943

**File name:** anmol.docx (115.17K)

**Word count:** 2112

**Character count:** 11168

