PROJECT: Predicting stock price for large –cap technology companies

Name: Sudeshna Gunjote

Submitted to: Technocolabs

Department of Computer Science

Course : Machine Learning

**Abstract**

In this project, we were asked to implement a stock price prediction

project from all the topics and concepts used in machine learning sessions during the training. All the concepts we learned and implement a mini project on the bases of any data set. And hereby present the report with understanding and considering the learning aspects. Here, is my Report for the project letter.

**Keywords**: Machine Learning, Linear Regression and xboost, Classification, Supervised learning,

Artificial Intelligence.

**Introduction:**

Machine learning is sub domain of working in computer science. Machine Learning and artificial intelligence have boosted the technological aspects of computer science.

Machine learning is Core domain for Artificial intelligence and deep learning. It is the most important field in todays advanced environment and foreseen future in technologies. It is a method to analysis data and understand it. Machine learning also includes predicting and understanding data and forecasting information and relevant problem statement. It Involves computers discovering how a data can perform and extract necessary information with the help of pervious datasets.

Machine Learning includes many Algorithms and learning to implement. It has mainly three major aspects as supervised learning, unsupervised learning and Reinforcement learning. Each has different requirements to be understood according to the dataset we are working on.

Supervised models can be further grouped into regression and classification cases:

* Classification: A classification problem is when the output variable is a category e.g. “disease” / “no disease”.
* Regression: A regression problem is when the output variable is a real continuous value e.g. stock price prediction

Unsupervised models can be further grouped into clustering and association cases.

* Clustering: A clustering problem is where you want to unveil the inherent groupings in the data, such as grouping animals based on some characteristics/features e.g. number of legs.
* Association: An association rule learning is where you want to discover association rules such as people that buy X also tend to buy Y

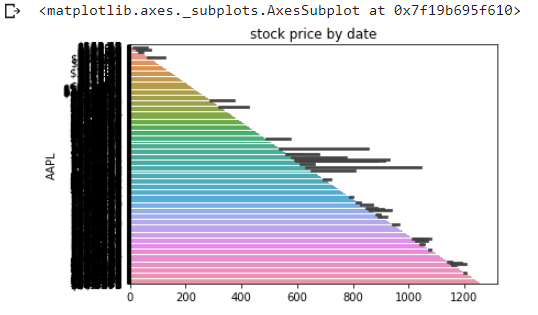
**DATA SET:**

The dataset used is a sample of from nasdaqa and xignite illustrates all the details of prediction stock prices for large-cap technology companies according to the features of all the stock prices for large-cap technology companies.

Each Vector plays an important role in stock prices for large-cap technology companies. The dataset is provided with the column name as:

The dataset has the information about 2000 mobile details and their price range. Each Price range is take under consideration as 10k, 20k, 30k, 40k respect for the numbers 0,1,2,3.

The dataset has various datatypes and the relation between them can be identified by using correlation.



Each Strong correlation is represented y lighter shades. So, I have plot the representations for each among each other and importantly with price range. Stronger the correlation greater is the importance in the dataset and plays important in machine learning.

We can observe theirs a strong correlation between our target variable that price range and ram and also battery power having correlation value near to 1 which is stated a s high correlation among the scale of 0 to 1.

**Problem Statement:**

We are given a data set of prediction stock price too. The data set is show above and we need to predict the stock price of the prediction stock price market by analyzing the dataset given specifications of the prediction stock price .

The stock range is given according to the specifications available in the dataset by date range they are sold. We will be using different algorithms to predict and train our model accordingly.

**Algorithm:**

Linear Regression:

Linear regression is an attractive model because the representation is so simple.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1\*x

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 \* x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

**KNN Algorithm:**

* K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
* It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data

Confusion matrix for Xgboost:

[[ 92 5 0 0]

[ 3 78 8 0]

[ 0 9 86 5]

[ 0 0 5 109]]

Here, The dataset has no null and duplicate values. As we can state it is a clean dataset and we dropped all the unnecessary columns which created complexity of the dataset while implementation.

In the dataset train test split is implemented so we can train dataset and test its predictions using test dataset. Size of test dataset is 20%. KNN classifier is used to train the dataset and understand it, the accuracy for knn is also good with 94%.

Then to implement it with xgboost with xgb classifier the training dataset has 98% accuracy is good score and on the test data set it has 91% accuracy.

Linear regression with default parameter gave 91% accuracy. Overall we can observe the xgboost as xgbclassifier gives good accuracy with training and test dataset and predicted confusion matrix is stated above.