**Combined Cycle Power Plant Data Set**

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of dataset** | **Dataset Characteristics** | **Number of attributes** | **Number of instances** |
| Regression | Multivariate | 4 | 9568 |
|  |  |  |  |

**Introduction:**

The dataset used here is Combined Cycle Power Plant dataset which is been taken from UCI Repository <https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant>.

**Data Set Description:**

The dataset contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (EP) of the plant.

A combined cycle power plant (CCPP) is composed of gas turbines (GT), steam turbines (ST) and heat recovery steam generators. In a CCPP, the electricity is generated by gas and steam turbines, which are combined in one cycle, and is transferred from one turbine to another. While the Vacuum is collected from and has effect on the Steam Turbine, the other three of the ambient

variables effect the GT performance.

Predicting full load electrical power output of a base load power plant is important in order to maximize the profit from the available megawatt hours. The base load operation of a power plant is influenced by four main parameters, which are used as input variables in the dataset, such as ambient temperature, atmospheric pressure, relative humidity, and exhaust steam pressure. These parameters affect electrical power output, which is considered as the target variable.

**Attribute Information:**

Features consist of hourly average ambient variables

The following inputs:(X – independent variable)  
 - Temperature (T) in the range 1.81°C and 37.11°C,  
 - Ambient Pressure (AP) in the range 992.89-1033.30 milibar,  
 - Relative Humidity (RH) in the range 25.56% to 100.16%  
 - Exhaust Vacuum (V) in the range 25.36-81.56 cm Hg

The only output is: (y – Target variable)  
 - Net hourly electrical energy output (EP) 420.26-495.76 MW

The averages are taken from various sensors located around the plant that record the ambient variables every second. The variables are given without normalization.

**Machine Learning Procedure:**

**Supervised Machine Learning:**

Supervised Learning, also known as Supervised Machine Learning is defined by its use of labelled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately.

This occurs as part of the cross validation process to ensure that the model avoids overfitting or underfitting.

Supervised Learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in Supervised Learning include Neural Networks, Naive-Bayes, Linear Regression, Logistics Regression, Random Forest, Support Vector Machine(SVM) and more.

**Regression Algorithm:**

In Machine Learning, we use various kinds of algorithms to allow machines to learn the relationships within the data provided and make predictions based on patterns or rules identified from the dataset.

Regression is a supervised machine learning technique which is used to predict continuous values. Predicting prices of a house given the features of house like size, price etc is one of the common examples of Regression.

The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data. The three main metrics that are used for evaluating the trained regression model are variance, bias and error .

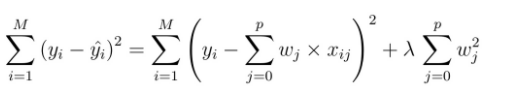
Types of algorithms used for this dataset are Linear, Ridge, Lasso, ElasticNet, SVM, Random Forest and Gradient Booster Algorithm.

**Linear Regression Algorithm:**

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables.

**Ridge Algorithm:**

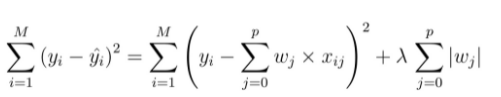
In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. This causes a challenge in notation since we must somehow indicate whether the variables in a particular formula are standardized or not. As far as standardization is concerned, all ridge regression calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back into their original scale. However, the ridge trace is on a standardized scale.



**Lasso Algorithm:**

The only difference in Lasso over Ridge Algorithm is instead of taking the square of the coefficients, magnitudes are taken into account.

In [statistics](https://en.wikipedia.org/wiki/Statistics) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning), lasso (**least absolute shrinkage and selection operator**; also Lasso or LASSO) is a [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) method that performs both [variable selection](https://en.wikipedia.org/wiki/Variable_selection) and [regularization](https://en.wikipedia.org/wiki/Regularization_(mathematics)) in order to enhance the prediction accuracy and interpretability of the resulting [statistical model](https://en.wikipedia.org/wiki/Statistical_model).Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).



**ElasticNet Algorithm:**

Elastic net is a popular type of regularized linear regressionthat combines two popular penalties, specifically the L1 and L2 penalty functions. Elastic Net is an extension of linear regression that adds regularization penalties to the loss function during training.

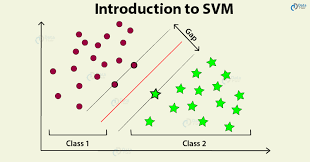
In [statistics](https://en.wikipedia.org/wiki/Statistics) and, in particular, in the fitting of [linear](https://en.wikipedia.org/wiki/Linear_regression) or [logistic regression](https://en.wikipedia.org/wiki/Logistic_regression) models, the elastic net is a [regularized](https://en.wikipedia.org/wiki/Regularization_(mathematics)) regression method that [linearly combines](https://en.wikipedia.org/wiki/Linear_combination) the [L1](https://en.wikipedia.org/wiki/Taxicab_geometry) and [L2](https://en.wikipedia.org/wiki/Norm_(mathematics)" \l "Euclidean_norm) penalties of the lasso and [ridge](https://en.wikipedia.org/wiki/Tikhonov_regularization) methods. The elastic net method overcomes the limitations of the [LASSO](https://en.wikipedia.org/wiki/Lasso_(statistics)) (least absolute shrinkage and selection operator) method.

**Support Vector Machines (SVMs):**

SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), support-vector machines (SVMs, also support-vector networks) are [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data for classification and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis). SVMs are one of the most robust prediction methods, being based on statistical learning frameworks. SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

The algorithm creates a line or a hyperplane which separates the data into classes.

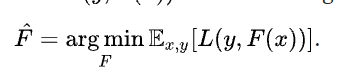


**Gradient Booster Algorithm:**

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. Gradient boosting algorithm can be used for predictingnot only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier).

A Gradient Boosting Machineor GBM combines the predictions from multiple decision trees to generate the final predictions. So, every successive decision tree is built on the errors of the previous trees. This is how the trees in a gradient boosting machine algorithm are built sequentially.

Gradient boosting is a machine learning technique for regression, classification and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms random forest.



**Random Forest Regression:**

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

1. Pick at random k data points from the training set.

2. Build a decision tree associated to these k data points.

3. Choose the number N of trees you want to build and repeat steps 1 and 2.

4. For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values.

**Data Preprocessing:**

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project and while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset. In this Computer Hardware Dataset there is no null values available so, we proceed to next step.

**Outliers and Missing Treatment:**

One of the biggest downfalls for any model performance is the outliers present in the data. Outliers ideally are the extreme values for the specific column which affects the generalization of the data and model. Outliers mostly affect the regression models as it changes the equation drastically. It is an observation that is numerically distant from the rest of the data.

When reviewing a boxplot, an outlier is defined as the data points that is located outside the whiskers of the boxplot. IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

IQR is the range between the first and the third quartiles namely Q1 and Q3:**IQR = Q3 – Q1.** The data points which fall **below Q1 – 1.5 IQR** or **above Q3 + 1.5 IQR** are outliers.

The intention of Outlier detection and treatment is to ensure us to get the best model out of the data considering the fact that your data is qualified to work with the algorithm.

**Algorithm for for IQR:**

1. Sort the data in increasing order.

2. Calculate the first quartile(q1) and third quantile(q3).

3. Find InterQuantile Range(q3-q1).

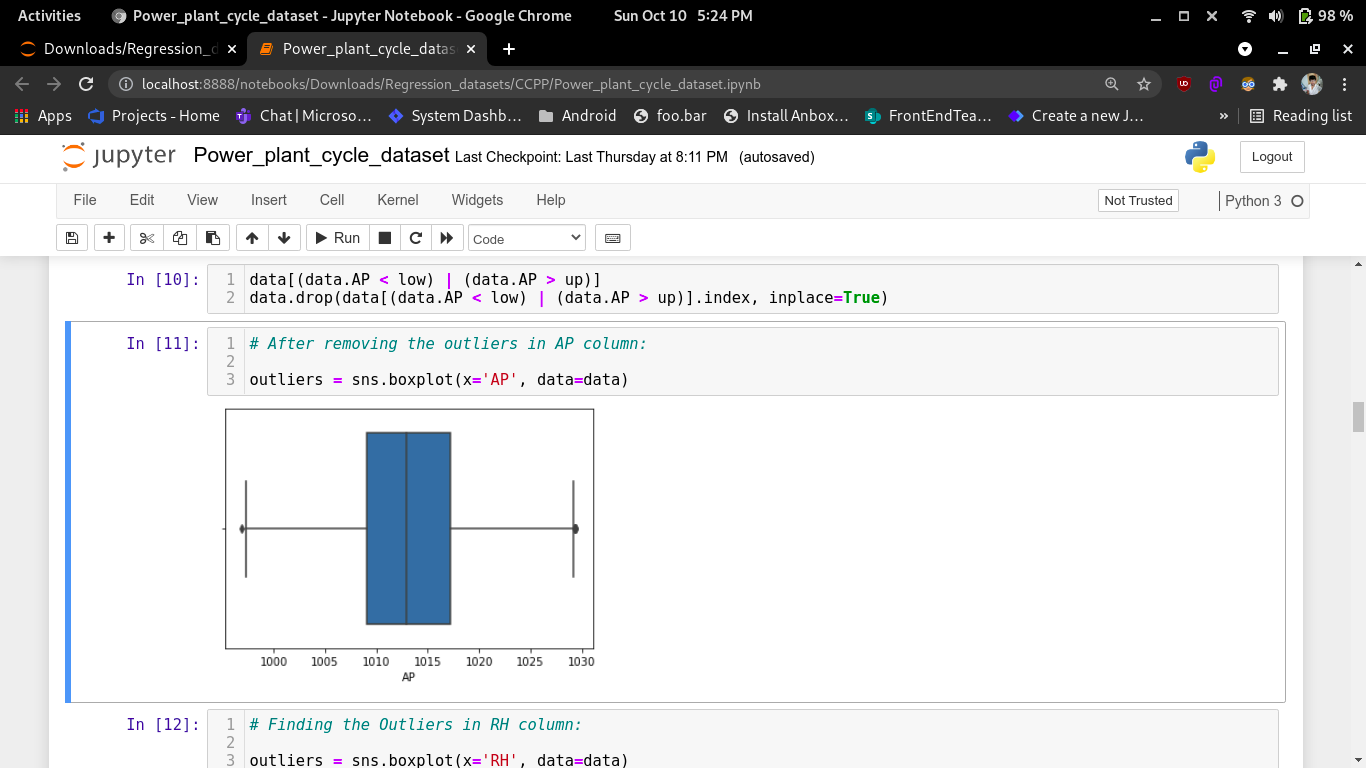
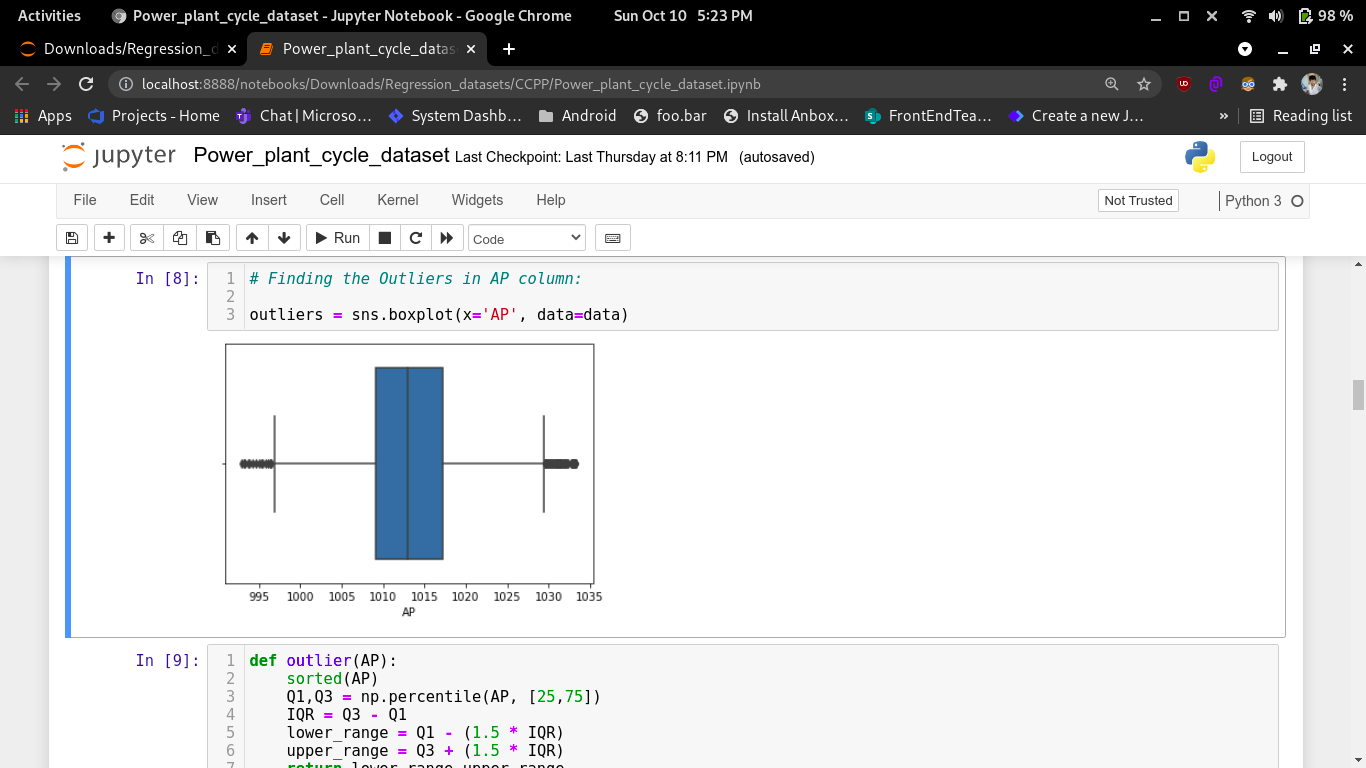
4. Find the lower\_bound = q1 – 1.5\*IQR

5. Find the upper\_bound = q3 + 1.5\*IQR

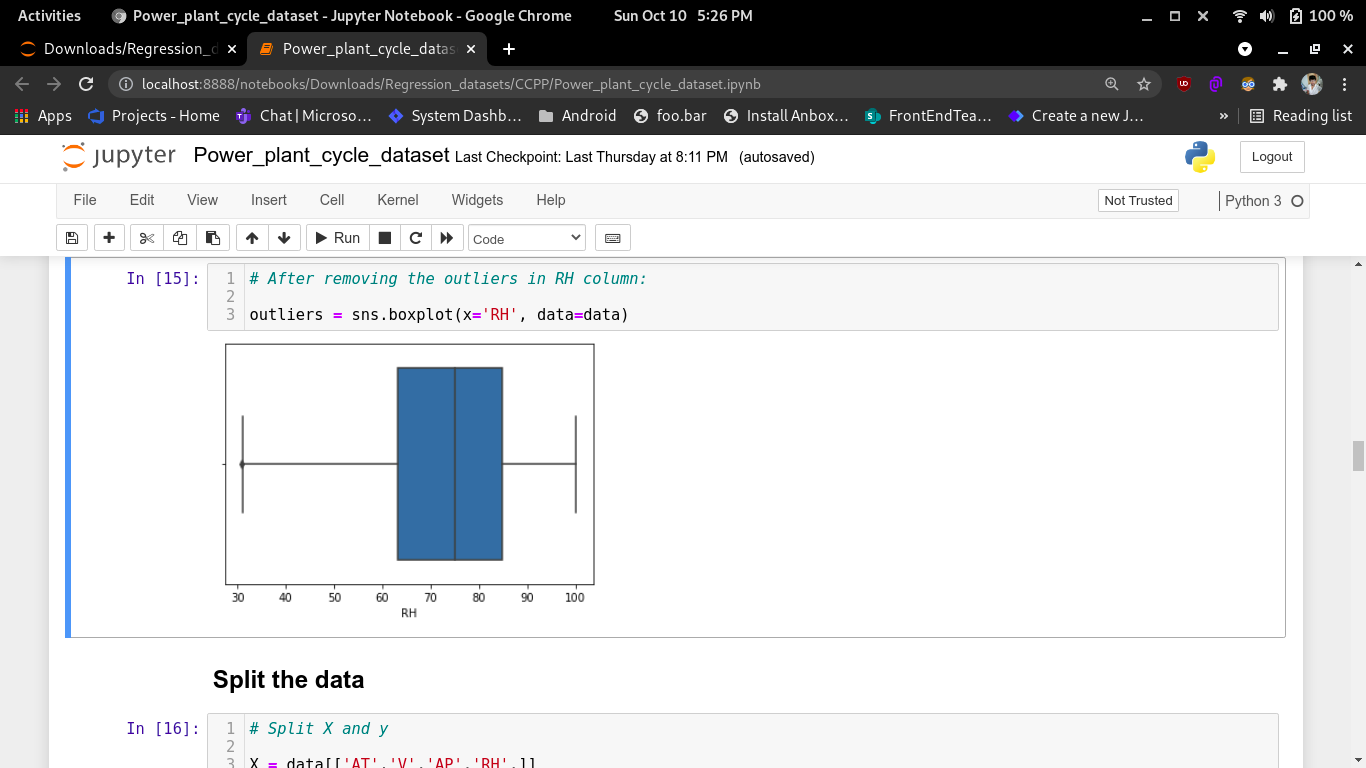
6. Points that lies above or below the IQR is an Outlier.

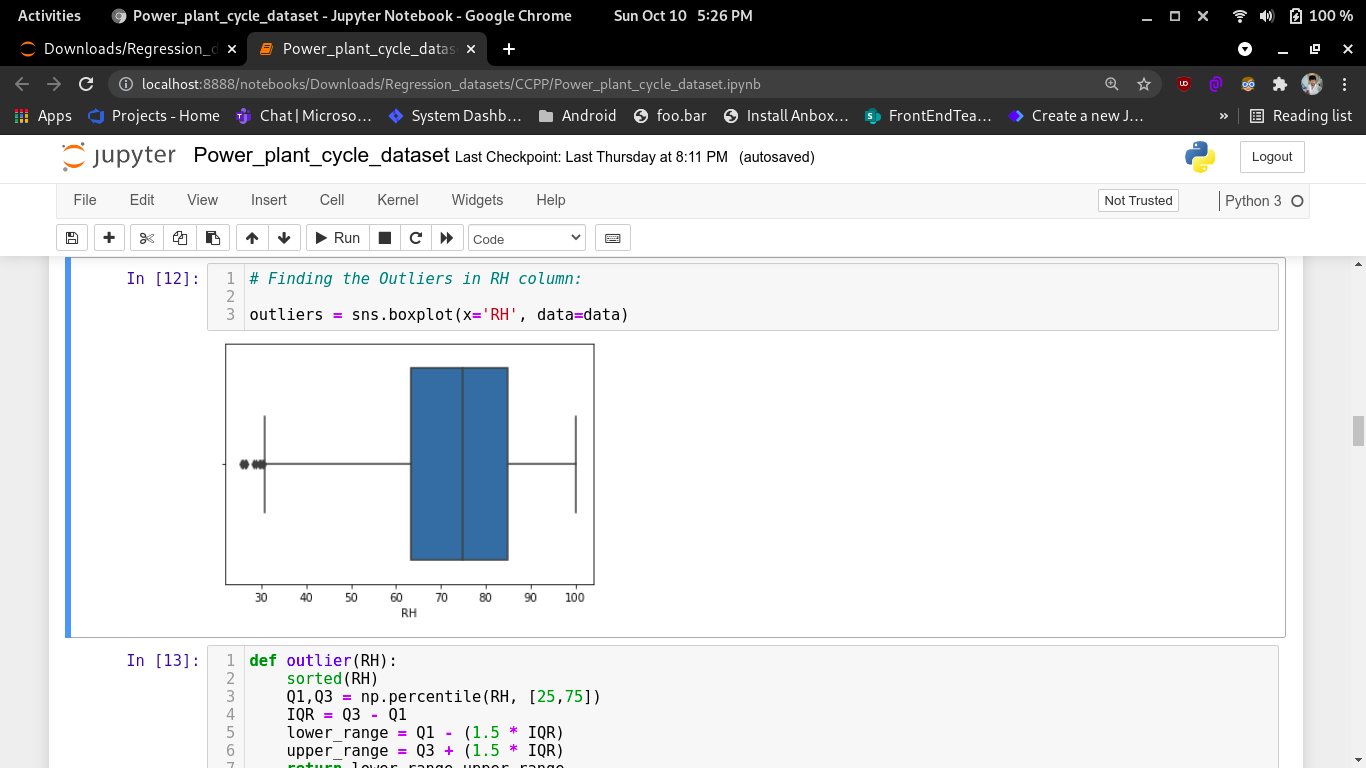
Box plots are useful as they provide a visual summary of the data enabling researchers to quickly identify mean values, the dispersion of the data set and signs of skewness.

Here , is the screenshots of outliers that are present and removed in Combined Cycle Power Plant dataset:



a) outliers in Ambient Pressure a1) after removing outliers of Ambient Pressure





b) outliers in Relative Humidity b1) after removing outliers of Relative Humidity

**Significance of Outliers:**

• Outliers badly affect mean and standard deviation of the dataset. These may statistically give erroneous results.

• Most machine learning algorithms do not work well in the presence of outlier. So it is desirable to detect and remove outliers.

• Outliers are highly useful in anomaly detection like fraud detection where the fraud transactions are very different from normal transactions.

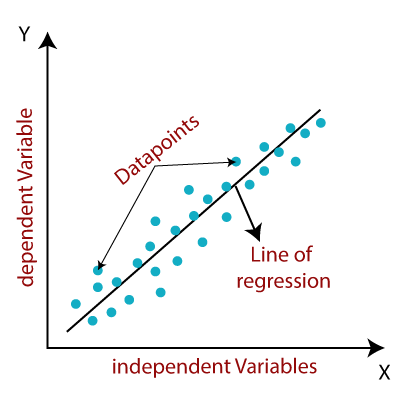
**Machine Learning Model:**

**Linear Regression Algorithm:**

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. If there is only one input variable (x), then such linear regression is called simple linear regression. And if there is more than one input variable, then such linear regression is called multiple linear regression. The relationship between variables in the linear regression model can be explained using the below image.



Formula: **y = a 0 +a 1 x+ ε**

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error. The different values for weights or the coefficient of lines (a 0 , a 1 ) gives a different line of regression, so we need to calculate the best values for a 0 and a 1 to find the best fit line, so to calculate this we use cost function.

**Cost function:**

o The different values for weights or coefficient of lines (a 0 , a 1 ) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.

o Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.

o We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.

**Best Fit Algorithm:**

The algorithm that gives maximum accuracy is Gradient Booster Regression and hence it is chosen as best fit algorithm for Combined Cycle Power Plant Dataset.

**Analysis of the dataset:**

By using the various Regression Algorithms we have analysed the accuracy as R2 Score for each algorithm and they are tabulated below:

|  |  |
| --- | --- |
| **Algorithm** | **R2 Score** |
| Linear Regression | 0.9245715944203565 |
| Ridge Algorithm | 0.9245716568030239 |
| Lasso Algorithm | 0.9244312271147285 |
| Elastic Net Regression | 0.9242972578980744 |
| Support Vector Algorithm | 0.9264914975194778 |
| Random Forest Regression | 0.9348176435769412 |
| Gradient Booster | 0.9448055229581702 |

This table shows the accuracy for each algorithm for Combined Cycle Power Plant Dataset. Here, Gradient Booster Regression Algorithm gives the maximum accuracy of 94%.

**Conclusion:**

From all the above discussions, we have concluded that the performance will be more accurate only after all the process have done. In this Combined Cycle Power Plant dataset the Estimated Relative Performance accuracy has been predicted. Hence, for this Combined Cycle Power Plant dataset **Gradient Booster Regression** Algorithm gives the maximum accuracy of **94%.**