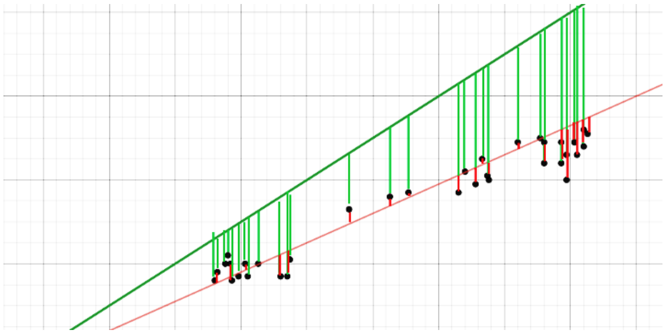
Machine Learning Algorithm

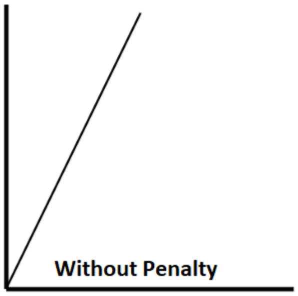
1. Linear Regression:

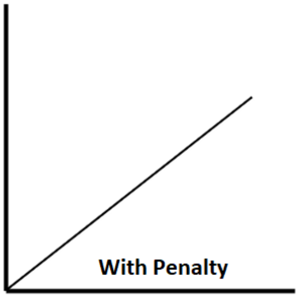
Linear Regression involves finding a ‘line of best fit’ that represents a dataset using the least squares method. Finding a linear equation that minimises the sum of squared residuals is the least squares method. A residual is the difference between the actual and projected values. To give an example, the red line is a better line of best fit than the green line because it is closer to the points, and thus, the residuals are smaller.



2. Ridge Regression

Ridge regression, also known as L2 Regularization, is a regression approach that reduces overfitting by introducing a little amount of bias. This is accomplished by reducing the sum of squared residuals plus a penalty equal to lambda times the slope squared. The term "lambda" relates to the severity of the punishment.





Without a penalty, the line of best fit has a steeper slope, which means that it is more sensitive to small changes in X. By introducing a penalty, the line of best fit becomes less sensitive to small changes in X. This is the idea behind ridge regression.

3. Lasso Regression

Lasso Regression, also known as L1 Regularization, is similar to Ridge regression. The only difference is that the penalty is calculated with the absolute value of the slope instead.

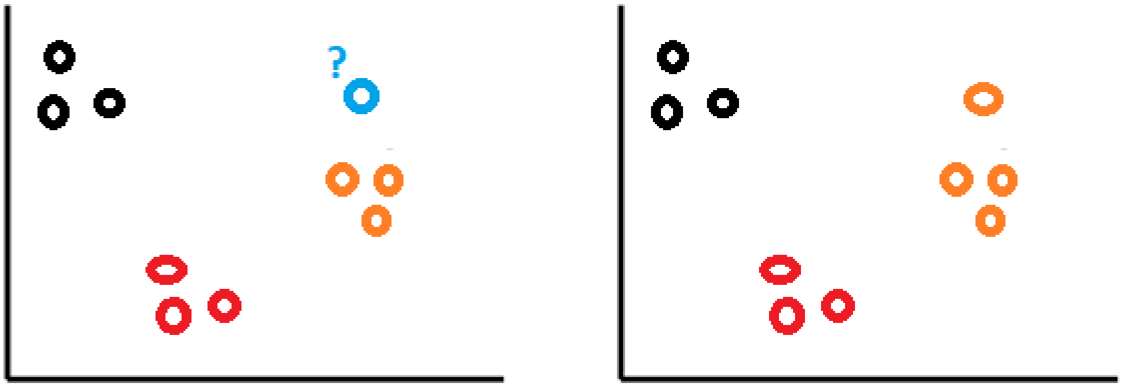
$$Minimizes\ Sum\ of\ Squared\ Residuals + \lambda * |slope|$$

4. Logistic Regression

Logistic regression is a classification method that also determines a "best fit line." Unlike linear regression, which uses least squares to find the best fit line, logistic regression uses maximum likelihood to get the best fit line (logistic curve). Because the y value can only be one or zero, this is done.

5. K-Nearest Neighbour

K-Nearest Neighbours is a classification technique that classifies a new sample by checking the closest classified points. If k=1, an unclassified point will be classed as a orange point in the example below.



If the value of k is too low, then it can be subject to outliers. However, if it’s too high, then it may overlook classes with only a few samples.

6. Naïve Bayes

The naïve Bayes classification algorithm is inspired based on the Bayes theorem. The equation is given by

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

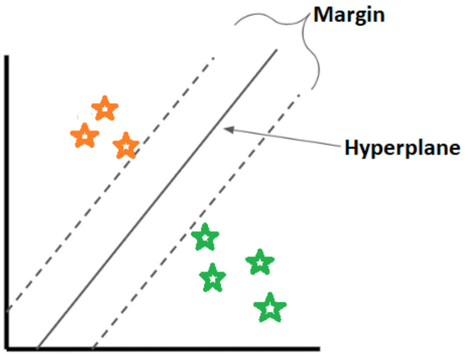
Because of the naive assumption that variables are independent given the class, we can rewrite P(X|y) as follows:

$$P(X|y) = P(x_1|y) * P(x_2|y) * \ldots * P(x_n|y)$$

Also, since we are solving for y, P(X) is a constant, which means that we can remove it from the equation and introduce a proportionality. Thus, the probability of each value of y is calculated as the product of the conditional probability of x_n given y.

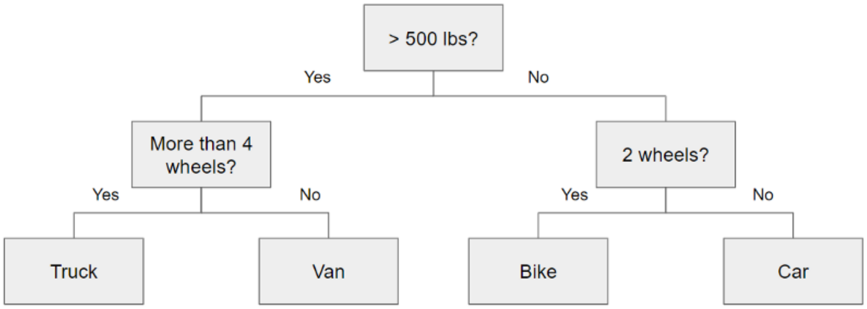
7. Support Vector Machine

Support Vector Machines are a classification method that identifies an ideal boundary, known as the hyperplane, for separating distinct classes. By increasing the margin between the classes, the hyperplane can be discovered.



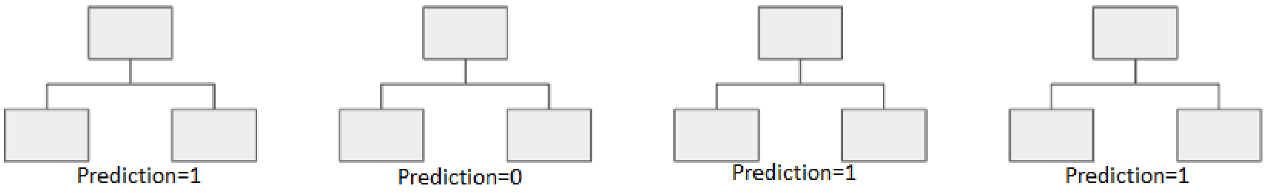
8. Decision Tree

A decision tree is essentially a series of conditional statements that determine what path a sample takes until it reaches the bottom. They are intuitive and easy to build but tend not to be accurate.



9. Random Forest

Random Forest is an ensemble technique, which means it integrates numerous models into a single model to boost predictive ability. It does it by leveraging bootstrapped datasets and random subsets of variables to create 1000s of smaller decision trees (also known as bagging). Random forests employ a ‘majority wins' strategy to estimate the value of the target variable by combining 1000s of smaller decision trees.



10. Ada Boost

AdaBoost is a boosted algorithm similar to Random Forests but it is differed in multiple ways

- AdaBoost typically makes a forest of stumps (a stump is a tree with only one node and two leaves) rather than a forest of trees.
- Each stump’s decision is not weighted equally in the final decision. Stumps with less total error (high accuracy) will have a higher say.
- The order in which the stumps are created is important, as each subsequent stump emphasizes the importance of the samples that were incorrectly classified in the previous stump.

11. Gradient Boost

Gradient Boost is similar to AdaBoost in that it creates many trees, each of which is built on top of the one before it. Gradient Boost, unlike AdaBoost, which creates stumps, creates trees with 8 to 32 leaves.

Gradient Boost is distinguished from AdaBoost by the way the decision trees are constructed. Gradient Boost begins with a baseline prediction, which is usually the average. The residuals of the samples are then used to build a decision tree. The process is repeated by taking the initial forecast + a learning rate times the residual tree's outcome to make a new prediction.

12. XG Boost

The key difference between XGBoost and Gradient Boost is the way the residual trees are constructed. The residual trees are constructed using XGBoost by computing similarity scores between leaves and preceding nodes to determine which variables are used as the roots and nodes.

1. What is linear and non-linear regression?

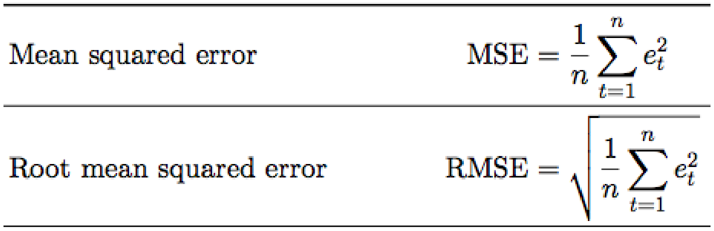
A supervised statistical technique for determining the relationship between a dependent variable and a set of independent variables is regression analysis.

- The linear regression model assumes that the dependent and independent variables have a linear relationship. Y = a+bx, where x is the independent variable and Y is the dependent variable.
- It's simple to use and interpret linear regression.
- The Y = a+bx equation is not followed in non-linear regression. Curve fitting is substantially more flexible with non-linear regression. The polynomial of k-degrees can be used to express it.

2. What are MSE and RMSE?

The mean squared error (MSE) is the sum of all squared errors or it's the average of the squared differences between the expected and actual value.

The square root of the average of squared differences between the expected and actual value is the RMSE (Root Mean Squared Error).

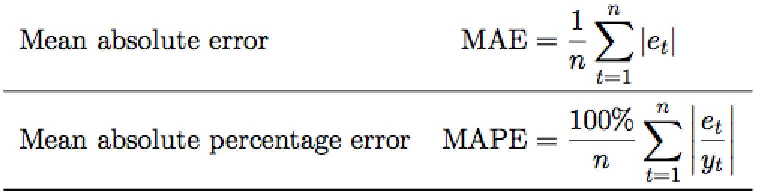


Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$

3. What is MAE and MAPE?

The mean absolute error (MAE) is the sum of all absolute or positive errors or it's the average of absolute or positive differences between the expected and actual value.

The average absolute error in percentage terms is calculated using MAPE (Mean Absolute Percent Error). It can be defined as the percentage average of absolute or positive errors.



Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

4. What is the difference between R-square and Adjusted R-square?

R-square is the measure the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model.

$$\mathit{Coefficient\ of\ Determination} \rightarrow \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The main problem with the R-squared is that it will always the same or increase with adding more variables. Here Adjusted R square can help. Adjusted R-square penalizes you for adding variables that do not improve your existing model.

5. What is the difference between Correlation and Regression?

The strength or degree of association between two variables is measured by correlation. It fails to account for causality. A single point is used to represent it.

Regression is a method of determining how one variable influence another. Model fitting is what regression is all about. It depicts cause and consequence and captures causality. It is represented by a line.

6. What is Multicollinearity?

Collinearity is another name for multicollinearity. It's a situation in which two or more independent variables are highly correlated, such that one variable may be predicted linearly from the others. It assesses the interrelationships and associations between independent variables.

Multicollinearity is caused by the incorrect use of dummy variables or any variable in the data that is computed from another variable.

It has an effect on regression coefficients and results in large standard errors. The correlation coefficient, Variance inflation factor (VIF), and Eigenvalues can all be used to detect this.

7. What is VIF? How do you calculate it?

The variance inflation factors (VIF) assess how much collinearity increases the variance of an estimated regression coefficient. It calculates the amount of multicollinearity in a regression study.

It runs an ordinary least square regression with Xi as a function of all other explanatory or independent variables, and then computes VIF using the formula:

VIF=1/(1-R^2)

8. What is heteroscedasticity?

Heteroscedasticity refers to the situation where the variability of a variable is unequal across the range of values of a second variable that predicts it. It can be detected using graphs or statistical tests such as Breush-Pagan test and NCV test.

9. What is a Box-Cox Transformation?

A mathematical transformation of a variable to approximate a normal distribution is known as the Box-Cox transformation. Skewed data is transformed into normally distributed data using the Box-Cox method.

10. What are the basic assumptions of Linear Regression?

The assumptions of linear regression are

- The relationship between the features and target are **linear**.
- The error term has constant variance (**Homoscedasticity**).
- There is no **multicollinearity** between the features.
- Observations are **independent** of each other.
- The error(residuals) follows a **normal** distribution.

11. List down some of the metrics used to evaluate a Regression Model.

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-Squared (Coefficient of Determination)
- Adjusted R-Squared

12. How do we interpret a Q-Q plot for a linear regression model?

The Q-Q plot is a graphical representation of the quantiles of two distributions plotted against one another. We should focus on the 'y = x' line, which corresponds to a normal distribution, whenever we analyse a Q-Q plot. In statistics, this line is also known as the 45-degree line.

It implies that the quantiles of each distribution are the same. If you see a deviation from this line, it means that one of the distributions is skewed as compared to the other, which is the normal distribution.

13. What is Ordinary Least Squares (OLS)?

In linear regression model the objective is to find coefficients α and β by minimizing the error. The process of linear model tries to minimize the sum of squared error between the observed and predicted values is called ordinary least squares.

14. What are dummy variables?

A categorical independent variable is referred to as a dummy variable. Dummy variables are variables that are used in regression analysis. It's also called a qualitative variable, a category variable, a binary variable, or an indicator variable. There are always n-1 dummy variables in a column with n categories.

15. How random forest regressor works?

Random forest is a bagging algorithm that parallelizes the execution of several decision trees. We will take few samples from the data set and create one decision tree for each sample. Performs majority vote on final predicted values of many trees in a classification problem. Finds the mean of the final predicted values from many decision trees in a regression problem.

16. What are the disadvantages of linear regression?

The assumption of linearity is the major disadvantage of linear regression. It fails to fit complex issues because it assumes a linear relationship between the input and output variables. It is vulnerable to outliers and noise. Multicollinearity has an impact on it.

17. What is the use of regularisation? Explain L1 and L2 regularisations.

Regularization is a technique for dealing with the problem of overfitting. It attempts to find the right balance between bias and variation. It streamlines the learning process for more complex and adaptable models. Regularization techniques L1 and L2 are usually utilized. The L1 or LASSO (Least Absolute Shrinkage and Selection Operator) regression adds a penalty term to the loss function that is equal to the absolute magnitude of the coefficient. The squared magnitude of the coefficient is added to the loss as a penalty term in L2 or Ridge regression.

18. Is a random forest a better model than a decision tree?

Yes, a random forest is better than a decision tree because it is an ensemble bagging method that combines multiple decision trees to create a more robust and accurate combined classifier.

19. What is a kernel? Explain the kernel trick?

To differentiate the classes, SVM employs a kernel function. The kernel transforms the lower-dimensional space into the higher-dimensional space that is required. It can be used to solve both linear and nonlinear separation problems. We do have option of using a linear, polynomial, and RBF kernel.

20. What is XGBoost?

eXtreme Gradient Boosting is the abbreviation for XGBoost. It trains models independently of one another and in a sequential order. Using Gradient Descent, each newly trained model corrects the errors made by the preceding ones. It is more adaptable and quicker. It has a higher risk of overfitting. Regularization is used to penalise the model.

21. What is AdaBoost?

One of the ensembles boosting classifiers is Ada-boost, or Adaptive Boosting. AdaBoost is an iterative ensemble method that combines several low-performing classifiers to produce a high-accuracy, powerful classifier.

22. What is the difference between Bagging and Boosting?

A Bootstrap Aggregation is a Bagging. It works in parallel to create multiple models. Bagging algorithms reduce variance, making them ideal for models with high volume and low bias. Random forest and the Extra tree method are two examples of bagging.

Boosting is a method of creating models in a sequential order. Boosting algorithms reduce bias, making them suited for models with low variance and high bias. XGBoost AdaBoost is an example of boosting.

23. What do you mean by overfitting and underfitting?

Models that are underfitted have a large bias, less complex, and have less variance.The problem can be solved by raising the model's complexity and adding more parameters.

Overfitted models have a lower bias, more complex, and have a higher variance. Reduce complexity and introduce regularization to overcome it.

24. Which algorithm is better random forests or SVM and why?

Yes, when compared to a support vector machine, a random forest is a better option.

- In comparison to SVM, building a random forest is faster.
- SVM has a lower scalability and uses more memory.
- Random forests are used to determine the relevance of features.

25. What is logistic regression?

Logistic regression is one of the most extensively used binary classification techniques since it is simple to use and apply. It's utilised in spam identification, churn prediction, and diabetes prediction, among other things. It employs a sigmoid function to forecast the probability of a binary event using the log of odds as the dependent variable.