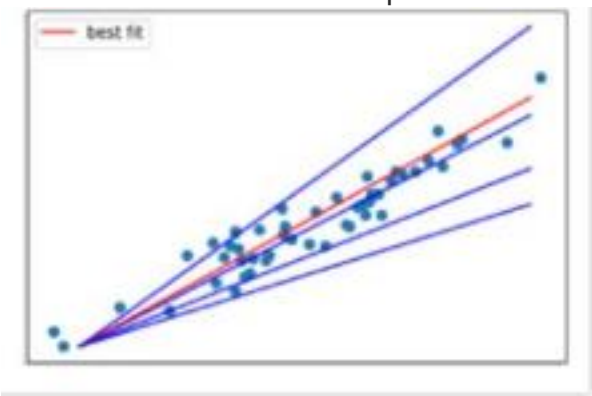
1. What is Linear regression?

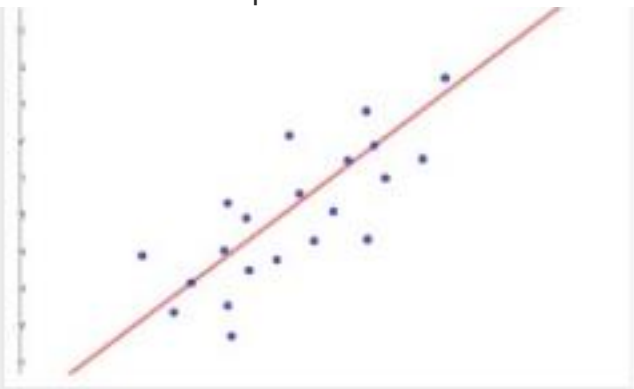
****

a. It tries to find out the best possible linear relationship between the input features and the target variable(y).

b. “It is a supervised machine learning algorithm that best fits the data which has the target variable(dependent variable) as a linear combination of the input features(independent variables). ”

 The target variable is also known as an independent variable or label. Input features are also known as dependent variables.





c. when you think of linear regression think of fitting a line such that the distance between the data points and the line is minimum. As shown above, the red line best fits that data than the other blue lines.

a. Linear equation is,

$$y = mx + c$$

(The goal of the linear regression is to find the best values for θ and b that represents the given data.)

2. How do you represent a simple linear regression?

********** a. Simple linear regression is used to estimate the relationship between two quantitative variables. You can use simple linear regression when you want to know:

b. How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).

c. The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

d. Assumptions:

i. No multicolinearity

ii. There should be linear relationship between input and output variable iii. Homoscedascity

iv. Normality of residual

Equation is,

$Y = mx + c$

Y – dependent variable

m - is the regression coefficient – how much we expect y to change as x increases.

c - is the intercept, the predicted value of y when the x is 0.

3. What is multiple linear regression?

********** a. Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. You can use multiple linear regression when you want to know:

b. How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).

c. The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

d. Assumptions:

i. No multicolinearity

ii. There should be linear relationship between input and output variable iii. Homoscedascity

iv. Normality of residual

e. Equation is,

$$y = m1x1 + m2x2+ \ldots. + mn\ xn + c$$

y - the predicted value of the dependent variable

c - the y-intercept (value of y when all other parameters are set to 0)

m1x1 - the regression coefficient (B1) of the first independent variable

mnxn - the regression coefficient of the last independent variable

4. What are the assumptions made in the Linear regression model?

********** 1. Linearity:

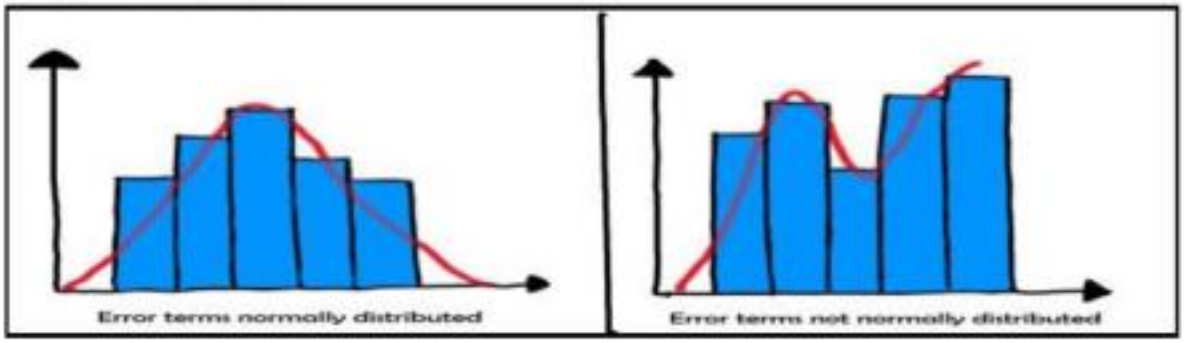
There should be linear relationship between dependent and independent variables.

2. No Multicolinearity:

There should not be linear relationship between independent variables. All independent variables should be independent of each other.

3. Normality of Residual:

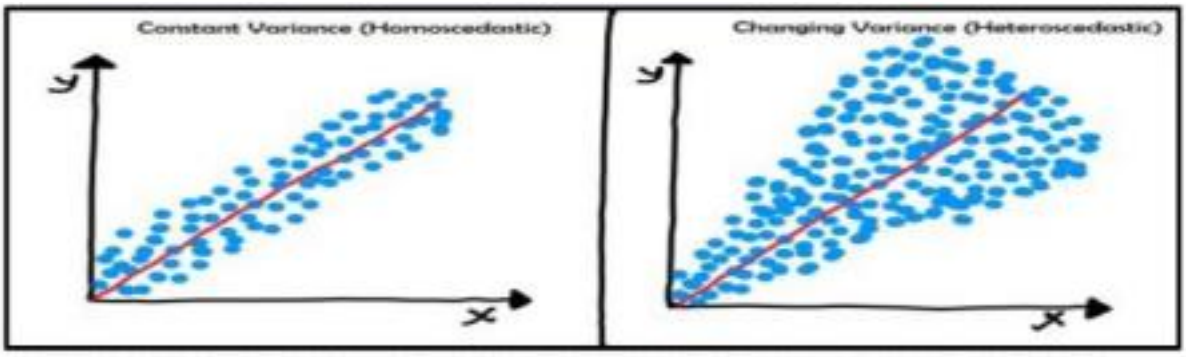
Underlying residuals are normally distributed. Residuals within the same range. (It uses density plot and Hypothesis testing for this)



If the residuals are not normally distributed, their randomness is lost, which implies that the model is not able to explain the relation in the data.

4. Homoscedascity of residual or no Heteroscedascity:

It is assumed that the residual terms have the same (but unknown) variance, σ2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.



5. What if these assumptions get violated?

****

1. Linear and Additive:
   - If you fit a linear model to a non-linear, non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model. Also, this will result in erroneous predictions on an unseen data set.

2. Autocorrelation:
   - The presence of correlation in error terms drastically reduces model's accuracy. curs in time series models where the next instant is dependent on previous instant. If the error terms are correlated, the

estimated standard errors tend to underestimate the true standard error.

- If this happens, it causes confidence intervals and prediction intervals to be narrower. Narrower confidence interval means that a 95% confidence interval would have lesser probability than 0.95 that it would contain the actual value of coefficients

3. Multicollinearity:

- This phenomenon exists when the independent variables are found to be moderately or highly correlated. In a model with correlated variables, it becomes a tough task to figure out the true relationship of a predictors with response variable. In other words, it becomes difficult to find out which variable is actually contributing to predict the response variable.
- Another point, with presence of correlated predictors, the standard errors tend to increase. And, with large standard errors, the confidence interval becomes wider leading to less precise estimates of slope parameters.
- Also, when predictors are correlated, the estimated regression coefficient of a correlated variable depends on which other predictors are available in the model. If this happens, you'll end up with an incorrect conclusion that a variable strongly / weakly affects target variable. Since, even if you drop one correlated variable from the model, its estimated regression coefficients would change. That's not good!

4. Heteroskedasticity:

- The presence of non-constant variance in the error terms results in heteroskedasticity. Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model's performance. When this phenomenon occurs, the confidence interval for out of sample prediction tends to be unrealistically wide or narrow.

5. Normal Distribution of error terms:

- If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on

minimization of least squares. Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.

6. What is the assumption of homoscedasticity?

****

It is assumed that the residual terms have the same (but unknown) variance, σ2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.

7. What is the assumption of normality?

****

Underlying residuals are normally distributed. Residuals within the same range. (It uses density plot and Hypothesis testing for this)
If the residuals are not normally distributed, their randomness is lost, which implies that the model is not able to explain the relation in the data.

8. How to prevent heteroscedasticity?

****

1. Transform the dependent variable:
   - One way to fix heteroscedasticity is to transform the dependent variable in some way. One common transformation is to simply take the log of the dependent variable.

 (For example, if we are using population size (independent variable) to predict the number of flower shops in a city (dependent variable), we may instead try to use population size to predict the log of the number of flower shops in a city.
 Using the log of the dependent variable, rather than the original dependent variable, often causes heteroskedasticity to go away.)

2. Redefine the dependent variable:
 Another way to fix heteroscedasticity is to redefine the dependent variable. One common way to do so is to use a rate for the dependent variable, rather than the raw value.
 For example, instead of using the population size to predict the number of flower shops in a city, we may instead use population size to predict the number of flower shops per capita.

3. Use weighted regression:
 Another way to fix heteroscedasticity is to use weighted regression. This type of regression assigns a weight to each data point based on the variance of its fitted value.

9. What does multicollinearity mean?

****

 This means that an independent variable can be predicted from another independent variable in a regression model. For example, height and weight, household income and water consumption, mileage and price of a car, study time and leisure time, etc.
 Let me take a simple example from our everyday life to explain this. Colin loves watching television while munching on chips. The more television he watches, the more chips he eats and the happier he gets!
 Now, if we could quantify happiness and measure Colin's happiness while he's busy doing his favorite activity, which do you think would have a greater impact on his happiness? Having chips or watching television? That's difficult to determine because the moment we try to measure Colin's happiness from eating chips, he starts watching television. And the moment we try to measure his happiness from watching television, he starts eating chips.
 Eating chips and watching television are highly correlated in the case of Colin and we cannot individually determine the impact of the individual activities on his happiness. This is the multicollinearity problem!

10.What are feature selection and feature scaling?

********** 1. Feature selection:

- Feature selection methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable.
- The difference has to do with whether features are selected based on the target variable or not. Unsupervised feature selection techniques ignores the target variable, such as methods that remove redundant variables using correlation. Supervised feature selection techniques use the target variable, such as methods that remove irrelevant variables.
- Another way to consider the mechanism used to select features which may be divided into wrapper and filter methods
- Wrapper feature selection methods create many models with different subsets of input features and select those features that result in the best performing model according to a performance metric.
- Filter feature selection methods use statistical techniques to evaluate the relationship between each input variable and the target variable, and these scores are used as the basis to choose (filter) those input variables that will be used in the model
- Finally, there are some machine learning algorithms that perform feature selection automatically as part of learning the model. We might refer to these techniques as intrinsic feature selection methods.
- This includes algorithms such as penalized regression models like Lasso and decision trees, including ensembles of decision trees like random forest.
- Feature selection is also related to dimensionally reduction techniques in that both methods seek fewer input variables to a predictive model. The difference is that feature selection select features to keep or remove from the dataset, whereas dimensionality reduction create a projection of the data resulting in entirely new input features. As such, dimensionality reduction is an alternate to feature selection rather than a type of feature selection

2. Feature scaling:

- Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step
- Just to give you an example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively,

feature scaling would help them all to be in the same range, for example centered around 0 or in the range (0,1) depending on the scaling technique.

a. Normalization:

- Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as,

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

b. Standardization:

- Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

11.How to find the best fit line in a linear regression model?

****

A. Linear Regression models the relationship between a dependent variable (y) and one or more independent variables (X) using a best fit straight line (also known as regression line). The dependent variable is continuous. The independent variable(s) can be continuous or discrete, and the nature of the relationship is linear.

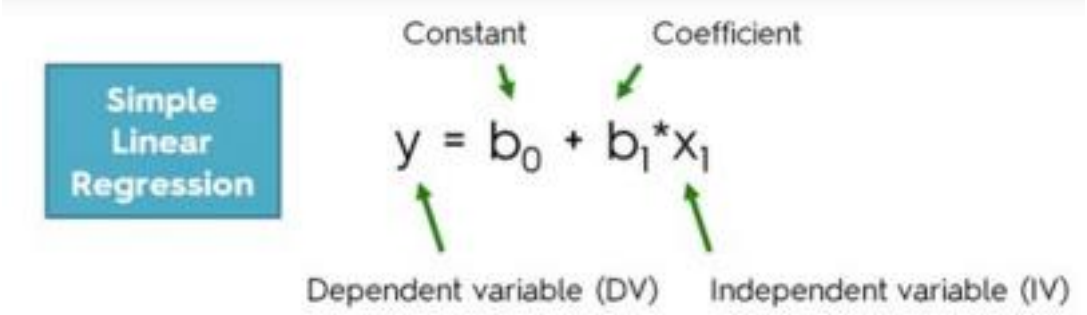
B. Linear relationships can either be positive or negative. A positive relationship between two variables basically means that an increase in the value of one variable also implies an increase in the value of the other variable. A negative relationship between two variables means that an increase in the value of one variable implies a decrease in the value of the other variable. (Correlation helps determine this relationship between variables)

C. Simple linear regression is an approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a

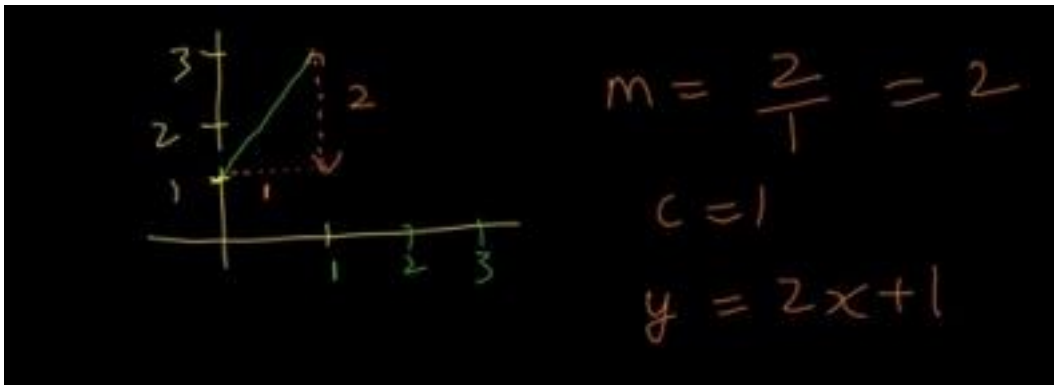
linear relationship between X and Y .

More Understanding or Linear Regression problem:

If we know there exists a linear relationship between the independent variable, and the dependent variable. we can use linear regression to this kind of data. However, the goal is to find the best fitting line parameters that best suits our data



Understanding slope and intercept:



Cost Function:

Need to find optimal parameters so that line equation best fits the training

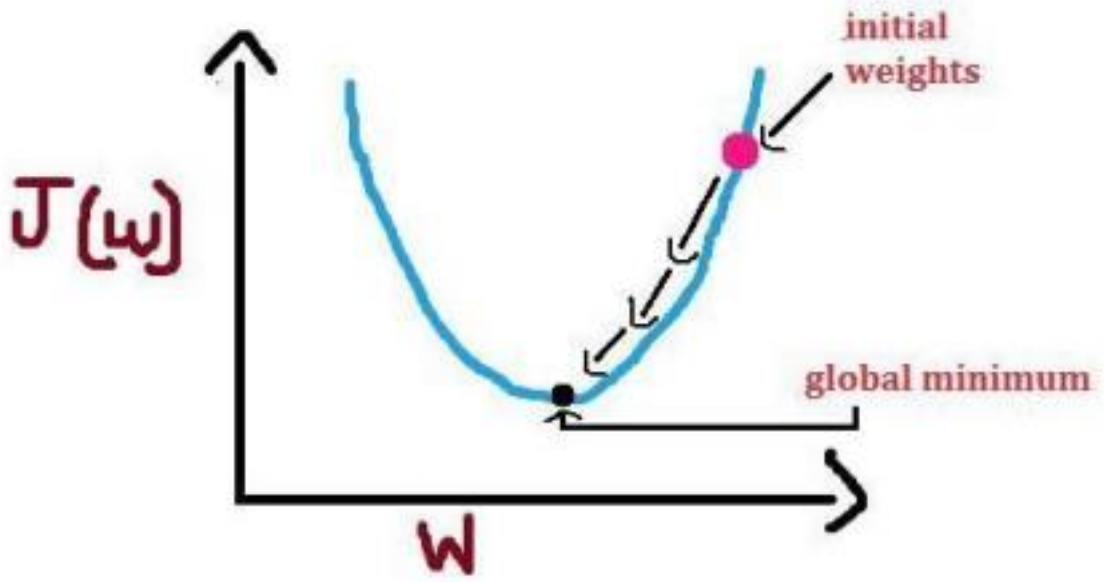
data. Residual = y_i – y_pred

Cost function is,

$$E = \frac{1}{n}\sum_{i=0}^{n}(y_i - \bar{y}_i)^2$$

Mean Squared Error Equation

It uses first order optimization algorithm to find best m and c values for which cost function is low.



And depending on best m and c values best fit line will be decided.

12.Why do we square the error instead of using modulus?

********** A. One reason apart from the “ease” of finding the Cost function using the squared error method over the absolute value method is because sometimes the latter method doesn’t have a differentiable point in it.

B. You may want to find the derivative of a cost function to find the most appropriate

value of bias and coefficient to fit the model. Should you use an absolute value, at that point (the non-differentiable point), the derivative will become not defined, and you can't proceed further.

13.What are techniques adopted to find the slope and the intercept of the linear regression?

****

- Gradient Descent
- Least Square Method / Normal Equation Method
- Adams Method
- Singular Value Decomposition (SVD)

14.What is cost Function in Linear Regression?

****

$$E = \frac{1}{n}\sum_{i=0}^{n}(y_i - \bar{y}_i)^2$$

Mean Squared Error Equation

Here y_i is the actual value and $\bar{y}_i$ is the predicted value. Lets substitute the value of $\bar{y}_i$:

$$E = \frac{1}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))^2$$

Substituting the value of $\bar{y}_i$

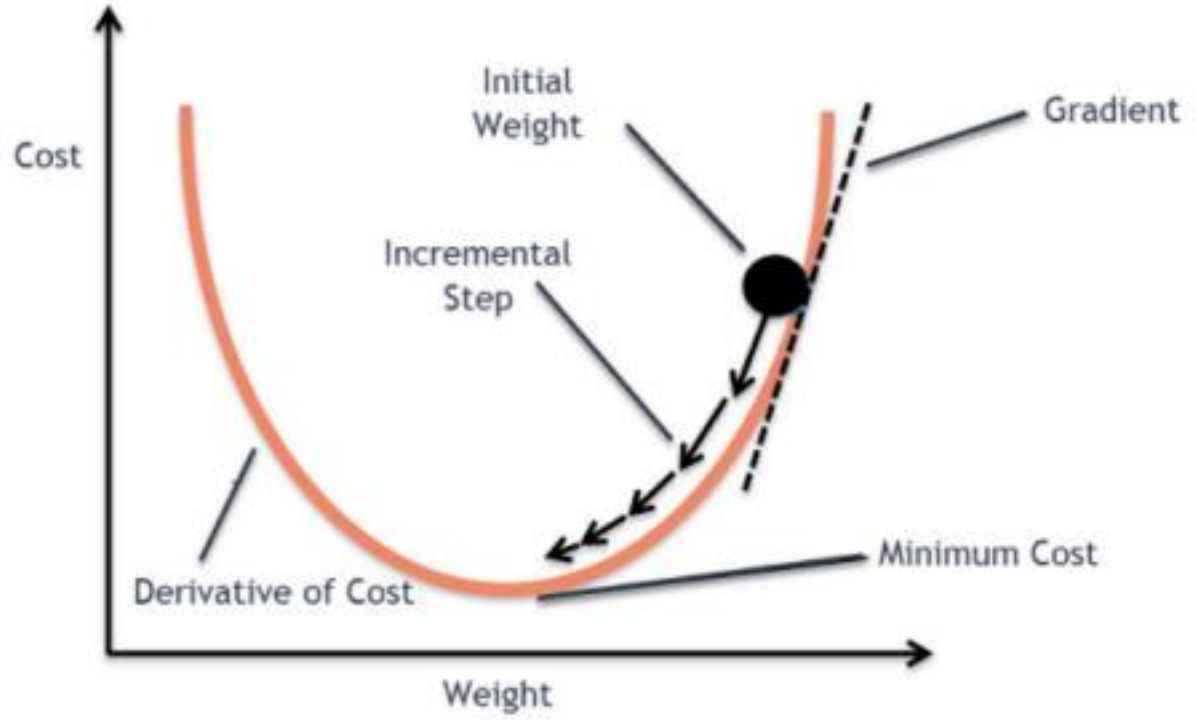
15.briefly explain gradient descent algorithm

****

- Let's say you are playing a game where the players are at the top of a mountain, and they are asked to reach the lowest point of the mountain. Additionally, they are blindfolded. So, what approach they would take to reach ground.
- The best way is to observe the ground and find where the land descends. From that position, take a step in the descending direction and iterate this process until we reach the lowest point.

 Gradient descent is an iterative optimization algorithm for finding the local minimum of a function.

 To find the local minimum of a function using gradient descent, we must take steps proportional to the negative of the gradient (move away from the gradient) of the function at the current point. If we take steps proportional to the positive of the gradient (moving towards the gradient), we will approach a local maximum of the function, and the procedure is called Gradient Ascent.



 The goal of the gradient descent algorithm is to minimize the given function (say cost function). To achieve this goal, it performs two steps iteratively:

1. Compute the gradient (slope), the first order derivative of the function at that point
2. Make a step (move) in the direction opposite to the gradient, opposite direction of slope increase from the current point by alpha times the gradient at that point

 Algorithm

1. Alpha is called Learning rate – a tuning parameter in the optimization process. It decides the length of the steps.
2. If there are two parameters, we can go with a 3-D plot, with cost on one axis and the two parameters (thetas) along the other two axes.
3. We have the direction we want to move in, now we must decide the size of the step we must take.(alpha – learning rate)

a) Learning rate is optimal, model converges to the minimum

b) Learning rate is too small, it takes more time but converges to the minimum

c) Learning rate is higher than the optimal value, it overshoots but converges ($1/C < \eta < 2/C$)

d) Learning rate is very large, it overshoots and diverges, moves away from the minima, performance decreases on learning

4. The cost function may consist of many minimum points. The gradient may settle on any one of the minima, which depends on the initial point (i.e initial parameters(theta)) and the learning rate. Therefore, the optimization may converge to different points with different starting points and learning rate.

16.How to evaluate regression models?

****

a. Mean/Median of prediction:

We can understand the bias in prediction between two models using the arithmetic mean of the predicted values.

(Disadvantage: Mean is affected by outliers. Use Median when you have outliers in your predicted values)

b. Mean/Median of prediction:

The standard deviation (SD) is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set,. In contrast, a high standard deviation indicates that the values are spread out over a broader range. The SD of predicted values helps in understanding the dispersion of values in different models.

$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{N - 1}}$$

c. Range of predictions:

- range of the prediction is the maximum and minimum value in the predicted values. Even range helps us to understand the dispersion between models.

d. Coefficient of Determination(R2)-

- R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.
- if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

R2 = 1 – (SSE / SST)

e. Mean Absolute Error:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

- It is thus an arithmetic average of the absolute errors, where yi is the prediction and xi the actual value. Note that alternative formulations may include relative frequencies as weight factors. The mean absolute error
  uses the same scale as the data being measured. This is known as a scale dependent accuracy measure and, therefore cannot be used to make
  comparisons between series using different scales.

f. Mean Squared Error (MSE):

- Mean Squared Error (MSE) or Mean Squared Deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and the actual value
- The MSE is a measure of the quality of an estimator — it is always non negative, and values closer to zero are better.

g. Root Mean Squared Error (RMSE):

- If the predicted responses are very close to the true responses the RMSE will be small. If the predicted and true responses differ substantially — at least for some observations — the RMSE will be

large. A value of zero would indicate a perfect fit to the data.

17. Which evaluation technique should you prefer to use for data having a lot of outliers in it?

****

Mean Absolute Error(MAE) is preferred when we have too many outliers present in the dataset because MAE is robust to outliers whereas MSE and RMSE are very susceptible to outliers and these start penalizing the outliers by squaring the error terms, commonly known as residuals.

18. What is residual? How is it computed?

****

- Residual is also called Error. It is the difference between the predicted y value and the actual y value.
- Residual = Actual y value – Predicted y value.
- It can be positive or negative.
- If residuals are always 0, then your model has a Perfect R square i.e. 1.

19. What are SSE, SSR, and SST? and What is the relationship between them?

********** a. SST:

The sum of squares total, denoted SST, is the squared differences between the observed dependent variable and its mean.
It is a measure of the total variability of the dataset.

b. SSR:

It is the sum of the differences between the predicted value and the mean of the dependent variable. Think of it as a measure that describes how well our line fits the data. If this value of SSR is equal to the sum of squares total, it means our regression model captures all the observed variability and is perfect.

c. SSE:

The error is the difference between the observed value and the predicted value. The smaller the error, the better the estimation power of the regression. Finally, I should add that it is also known as RSS or residual sum of squares.

20.What’s the intuition behind R-Squared?

********** R squared is a statistical measure that tells us how well the data fit in a regression model. It measures the proportion of variance in the dependent variable (Y) that can be explained by the independent variable (X). It can any value between 0 to 1 and is independent of the scale of Y. The formula for R squared is –

OR

R-squared = Explained variation / Total variation OR

a. R2 = +ve:

SSE = 0, Best score, All data points on BFL.

b. R2 = 0:

SSE = SST, Ypred = Ymean

(BFL line and mean line will be at same location)

c. R2 = -ve:

SSE > SST, not useful score.
(BFL is making more mistakes than mean line)

21.What does the coefficient of determination explain?

****

- The coefficient of determination is a measurement used to explain how much variability of one factor can be caused by its relationship to another related factor. This correlation, known as the "goodness of fit," is represented as a value between 0.0 and 1.0.
- how well the regression model fits the observed data. For example, a coefficient of determination of 60% shows that 60% of the data fit the regression model.(60% less variation around BFL than meal value)

22.Can R^2 be negative?

********** o A higher coefficient is an indicator of a better goodness of fit for the observations. The CoD can be negative, although this usually means that your model is a poor fit for your data. It can also become negative if you didn't set an intercept.

o It means regression line is making more mistakes than mean line.

23.What are the flaws in R-squared?

********** a. R-squared does not measure goodness of fit. It can be arbitrarily low when the model is completely correct. By making σ2σ2 large, we drive R-squared towards 0, even when every assumption of the simple linear regression model is correct in every particular. b. R-squared can be arbitrarily close to 1 when the model is totally wrong. c. R-squared says nothing about prediction error, even with σ2σ2 exactly the same, and no change in the coefficients. R-squared can

be anywhere between 0 and 1 just by changing the range of X. We're better off using Mean Square Error (MSE) as a measure of prediction error.

d. R-squared cannot be compared between a model with untransformed Y and one with transformed Y, or between different transformations of Y. R-squared can easily go down when the model assumptions are better fulfilled.

24.What is adjusted R^2?

********** Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

- The adjusted R-squared is a modified version of R-squared that adjusts for predictors that are not significant in a regression model.
- Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model.
- Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model.

25.What is the Coefficient of Correlation: Definition, Formula

****

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

26.What is the relationship between R-Squared and Adjusted R Squared?

****

1. R Squared is an econometric measure uses to explain the dependent and unconstrained variables where Adjusted R Squared is a value measuring that predicts the regression variables.
2. R Squared had Symbolized as R ^2 where Adjusted R Squared had written as Adjusted R ^2.
3. R squared is higher in getting the desired products, where Adjusted R Squared values are lower in measuring.
4. R Squared method had used to take the values originally where Adjusted R Squared values had been calculated mathematically.
5. Adjusted R Squared measurement requires the R Squared points for calculations.

27.What is the difference between overfitting and underfitting?

****

28.How to identify if the model is overfitting or underfitting?

****

 Overfitting is when the model's error on the training set (i.e. during training) is very low but then, the model's error on the test set (i.e. unseen samples) is large! Underfitting is when the model's error on both the training and test sets (i.e. during training and testing) is very high.

29.How to interpret a Q-Q plot in a Linear regression model?

****

 Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

 The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

30.What are the advantages and disadvantages of Linear Regression? **********

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

****

- In both L1 and L2 regularisation, the model is penalized for overfitting on train

data i.e. whenever the model tries to predict everything correctly on train data points, some penalty is added to the loss function in terms of the coefficients of the model.

- (we add a regularisation term(penalty) to the “loss function” so that the loss term does not becomes zero or close to zero for the train data.)
- The ML model will try to reduce the log loss to a very small value close to zero.

If the loss function is without the regularisation term, then the ML model will increase the weight parameter “m” to a very high value(ideally infinity) to make the overall loss close to zero. But this is something that will result in overfitting of the ML model.

L2 regularisation(Ridge):-

- To avoid overfitting we add a regularisation term
- The 2nd term in the loss function is the “L2” regularisation term. Here, the “squared magnitude of weight parameter” is added along with lambda (which is the hyperparameter to be tuned while building the model) to the logistic loss function.
- If the weight co_efficient “x” is made high, to reduce the 1st term in the loss function close to zero, then the second term will increase, thereby avoiding the overall loss function value from becoming zero. This way, the regularisation term penalizes the model for trying to make very accurate predictions on the training dataset points.

L1 Regularisation(Lasso):-

- Here, the “absolute value of weight parameter” is added along with lambda (which is the hyperparameter) to the loss function
- Similar to L2 regularisation, if the weight co_efficient “x” is made high, to reduce the value of 1st term in the loss function close to zero, then the second term i.e. the L1 regularisation term will increase, thereby avoiding the overall loss function from becoming zero.
- L1 regularisation penalizes the model less compared to L2 regularisation as it uses absolute values rather than the squared values of weight parameters in the loss function.