Interview Questions.

① What is Normalization & Standardization and how is it helpful?

Normalization & Standardization are techniques used to scale numerical data. That can be makeit to machine learning model process the data effectively.

Normalization rescale of data fit range in specific wally between 0, 1

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

The dataset has values ranging 50 to 100, The normalization will scale these to the range [0, 1]

* The features is different unit height cm & weight ~~~~ kg

* The model sensitive is k-NN, k-mean, Neural Network.

Standardization :-

Standardization transforms data to have a mean of 0 & standard diviation of 1. This process convert the data to follow a standard normal distribution.

Formula

$$x_{std} = \frac{x - \mu}{\sigma}$$

* μ = mean of feature
* σ : Standard diviation of the featu

⊛ How is it helpful in Normalization & Standardization.

① Improve model performance :-

* Speed up gradient decent converge in optimization based on model (Ex :- logistic regression, Neural Network)

② Handle different Scale :-

* Avoid features with large range dominating the model

③ Reduce Bias :-

* Ensure all features contribute equally to the model's prediction

④ Stability & Accuracy :-

* Reduces numerical errors and improve the stability of computations

⑤ Better distance calculation :-

* Essential for distance-based models like kNN, K-mean, SVM.

---

②. What techniques can be used to address multicolinearity in multiple linear regression?

Multicolinearity occurs when independent variables in regression model are highly correlated with each other.

* unstable co-efficient
* Reduced interpretability
* Inflated standard errors.

①. Detecting Multicollinearity:

① Variance inflation Factor (VIF):

* Measures how much the variance of coefficient increases due to multilinearity

* Rule of Thumb: If VIF > 5 multicollinearity present.

Formula:-

$$VIF = \frac{1}{1 - R^2}$$

* Correlation matrix:-

* check the pair wise pearson correlation co-efficients

* correlations > 0.8 suggest multicollinearity.

②. Techniques to Handle Multicollinearity

A) Data driven solutions:

1) Remove one of the correlated features:-

Drop the variable with the least predictive power.

2) Combine correlated features

* Create a new features by combining correlated variables.

3) Pricipal component Analysis (PCA)

* Reduces dimensionality while keeping most of the variance.

B) Statistical Model Adjustments:-

4) Ridge Regression (L2 Regularization)

* Add a penalty term to the loss function co-efficients.

formula:-

$$\min \|y - X\beta\|^2 + \lambda \|B\|^2$$

5) LASSO Regression (L1 Regularization)

* shrinks some co-efficients to zero, effectively performing feature selection.

Formula:-

$$\min \|y - X\beta\|^2 + \lambda \Sigma |\beta|$$

6) Elastic Net regression :-

* A combine a Ridge & LASSO regularization.

C) model interpretation techniques :-

7) Partial Least Squares (PLS) :

* Similar of PCA but focuses on maximizing the explanation of the dependent variable.

8) Drop highly correlated Dummy variables :-

* If using categorical variables, drop one dummy variable to prevent dummy variable trap.