**MACHINE LEARNING**

10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

The adjusted R-squared is a modified version of R-squared. The adjusted R-squared shows whether adding additional predictors improve a regression model or not. R-squared comes with an inherent problem – additional input variables will make the R-squared stay the same or increase (this is due to how the R-squared is calculated mathematically). Therefore, even if the additional input variables show no relationship with the output variables, the R-squared will increase. The adjusted R-squared looks at whether additional input variables are contributing to the model

11. Differentiate between Ridge and Lasso Regression.

The difference between ridge and lasso regression is that it tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero. Thus, Ridge regression decreases the complexity of a model but does not reduce the number of variables since it never leads to a coefficient been zero rather only minimizes it. Hence, this model is not good for feature reduction.

12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable. R2 value is determined to find out how well an independent variable is described by the other independent variables. A high value of R2 means that the variable is highly correlated with the other variables. This is captured by the VIF. So, the closer the R2 value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable. VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others.

13. Why do we need to scale the data before feeding it to the train the model?

Feature scaling is needed to bring every feature in the same footing without any upfront importance. Another reason why feature scaling is applied is that few algorithms like Neural network gradient descent converge much faster with feature scaling than without it. One more reason is saturation, like in the case of sigmoid activation in Neural Network, scaling would help not to saturate too fast.

14. What are the different metrics which are used to check the goodness of fit in linear regression?

Mean Absolute Error

 Root Mean Squared Error

Relative Absolute Error

Relative Squared Error

R2

15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.

TP=1000

FN=50

TN=1200

FP=250

Accuracy=(TP+TN)/TOTAL=2200/2500

Sensitivity/Recall=1000/1000+50

Specificity=1200/1200+250

Precision=1000/1000+250

**SQL**

11. What is denormalization?

Denormalization is a database optimization technique in which we add redundant data to one or more tables. This can help us avoid costly joins in a relational database.

12. What is a database cursor?

Cursor is a Temporary Memory or Temporary Work Station. Cursors are used to store Database Tables. There are 2 types of Cursors: Implicit Cursors, and Explicit Cursors.

13. What are the different types of the queries?

Data Definition Language (DDL) Statements

Data Manipulation Language (DML) Statements

Transaction Control Statements

Session Control Statements

System Control Statement

Embedded SQL Statements

14. Define constraint?

Constraints in SQL Server are predefined rules and restrictions that are enforced in a single column or multiple columns, regarding the values allowed in the columns, to maintain the integrity, accuracy, and reliability of that column’s data. In other words, if the inserted data meets the constraint rule, it will be inserted successfully. If the inserted data violates the defined constraint, the insert operation will be aborted.

15. What is auto increment?

Auto-increment allows a unique number to be generated automatically when a new record is inserted into a table. Often this is the primary key field that we would like to be created automatically every time a new record is inserted.

**STATISTICS**

10. What is the difference between a boxplot and histogram?

Histogram tells us about the frequency distribution whereas a boxplot is a rather detailed version of the data distribution giving an idea regarding centrality as well as spread of data.

11. How to select metrics?

Depending upon the need we need to perform descriptive or inferential tools of stats, be it to see the data centrality, the spread of data, correlation or the multicollinearity .On the other hand, deploying hypothesis test,etc.

12. How do you assess the statistical significance of an insight?

Statistical significance can be accessed using hypothesis testing:

– Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B)

– Then, we choose a suitable statistical test and statistics used to reject the null hypothesis

– Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)

– We calculate the observed test statistics from the data and check whether it lies in the critical region

Common tests:

– One sample Z test

– Two-sample Z test

– One sample t-test

– paired t-test

– Two sample pooled equal variances t-test

– Two sample unpooled unequal variances t-test and unequal sample sizes (Welch’s t-test)

– Chi-squared test for variances

– Chi-squared test for goodness of fit

– Anova

– Regression F-test

14. Give an example where the median is a better measure than the mean.

A dataset having outliers would present a distorted figure in terms of mean . In such cases, median would be a better alternative.

15. What is the Likelihood?

In statistics, the likelihood function measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters