Machine learning-4

1. C (High R-squared value for train-set & low R-squared value for test-set)

2. B (Decision tree are highly prone to overfitting)

3. C (Random forest)

4. C (Precision)

5. B (Model-B)

6. A & D (Ridge & Lasso)

7. A & D (Adaboost & Xgboost)

8. A & C (Pruning & Restricting the max depth of the tree)

9. B (A tree in the ensembles focuses more on the data points on which the previous tree was not performing well)

10. Adjusted R-squared is an indicator of whether adding additional predictors improve a regression model or not. Adjusted R-squared penalize the presence of unnecessary predictors by reducing its value. Like if we add a predictor that does not improve the fit of the model, then the adjusted R-squared value will be lower than it was before the predictor was added.

11. Ridge(L2 regularization) and Lasso(L1 regularization) regression are powerful techniques generally used for creating ungenerous models in presence of a large number of features. These are shrinkage methods which are used to put a similar constraints on the coefficients by introducing a penalty factor. A only difference between ridge and lasso regression is that, ridge reduces the variance of model whereas lasso reduces the number of features selected in a model.

12. **VIF** : A VIF(variance inflation factor) is a measure of the amount of multicollinearity in regression analysis i.e, it can estimate how much the variance of a regression coefficient is inflated due to multicollinearity. It ia basically a tool to identify the degree of multicollinearity.

A suitable VIF value is three or below.

13. Scaling the data before feeding it to the model is important because it helps to ensure that all features are on a similar scale, if features are on different scales, then the algorithm may give more weight to the larger scale feature and it can lead to poor performance. Scaling the data can also helps to standardize the range of the input variables. This helps the model converge faster during training and can also prevent overfitting.

14. There are several metrics that can be used to evaluate the goodness of fit in a linear regression model i.e.,

a. R-squared- This is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

b. R-squared adjusted- It is the modification of r-squared.

c. Mean squared error(MSE)- This is the average of the squared differences between the predicted and actual values.

d. Root mean squared error(RMSE)- This is the square root of mean squared error, which is a more interpretable metrics.

e. Mean absolute error(MAE)- This is the average of the absolute difference between the predicted and actual values.

f. Mean absolute percentage error(MAPE)- This metrics expresses error as a percentage of the true value.

g. Correlation coefficient- This ia the measure of the strength of the linear relationship between two variables. It ranges between -1 and 1.

15. According to confusion metrix values are:

True positives-1000
False positives-250
True negatives-1200
False negatives-50

so,

Sensitivity = TP/ (TP+FN)= 1000/(1000+50)= 0.95
Specificity= TN/(TN+FP)= 1200/(1200+250)= 0.82

Precision= TP/(TP+FP)= 1000/(1000+250)= 0.8

Recall= TP/(TP+FN)= 1000/(1000+50)= 0.95

Accuracy= TP+TN/ (TP+TN+FP+FN)= 1000+1200/(1000+1200+250+50)= 0.88

SQL-Worksheet- 4

1. A,C,& D (Commit, Rollback,& Savepoint)

2. A,C,& D (Create, Drop, & Alter)

3. B (Select name from sales)

4. C (Authorizing access & other control over database)

5. B (Column Alias)

6. B (Commit)

7. A (Parenthesis-(...))

8. C (Table)

9. D (All of the mentioned i.e., Data types, primary keys and default values)

10. A (ASC)

11. **Denormalization :** Denormalization is a process of intentionally adding redundancy to a database by incorporating data from multiple tables into a single table, or by adding duplicated data to the same table. This is done to improve read performance of the database. However, it also increases the risk of data inconsistencies and can make it more difficult to update the data.

12. **Database cursor :** A cursor in database is a control structure that enables traversal over the records in a database. It allows for iterating through the rows of a result set one row at a

time, rather than loading all the rows into memory at once. Cursor can be used to retrieve,add,update, or delete rows one at a time.Cursors are mainly used in stored procedures and triggers in SQL.

13. There are five major types of queries are:

a. DDL (Data definition language)- Create, alter, drop, truncate,etc.

b. DML (Data manipulation language)- Insert, update, delete, call, etc.

c. DCL (Data control language)- Grant, revoke, etc.

d. DQL (Data query language)- Select.

e. TCL (Transaction control language)- Commit, savepoint, rollback, set constraints, etc.

14. **Constraint :** The constraint in a database is a rule that is used to limit the type of data that can be inserted or updated in a table. It is used to maintain the integrity and accuracy of data within a table and to prevent data from being entered that could cause errors in the database. Constraints are usually defined during table creation. There are several types of constraints i.e.,

a. Primary key : enforces unique values and can't be null for the column or set of columns on which it is defined.

b. Foreign key : enforces a link between data in two tables, ensuring that data in one table corresponds to data in another table.

c. Unique : ensures that tha data in a column or set of columns is unique and non-duplicate.

d. Not Null : ensures that a column can't contain null values.

e. Default : sets a default value for a column if no value is provided when a new row is inserted.

f. Trigger : it is a special type of constraint that is executed automatically when certain events occur in the database, such as insert, update or delete.

15. Auto increment is a feature in relational databases that automatically increases the value of a specified column (typically the primary key) by a predefined amount each time a new record is inserted into the table. It is used to ensure that each record has a unique identifier.

Statistics Worksheet-4

1. d (All of the mentioned i.e. the outcome from the roll of a die, the outcome of flip of a coin, the outcome of exam)

2. a (Discrete)

3. a (pdf)

4. c (Mean)

5. c (Empirical mean)

6. a (Variance)

7. c (0 and 1)

8. b (Bootstrap)

9. a (Frequency)

10. A boxplot and a histogram are both graphical representation of data. The only difference between them is the types of information they displays.
A boxplot is used to show the distribution of a dataset, including median, quartiles and outliers whereas a histogram shows the frequency distibution of a dataset and the shape of the data distribution.

11. Selecting the appropriate metrics to evaluate the performance of a model or system is very crucial. Some general guidelines are:

A. Define the problem: Understand the problem you are trying to solve and the goals you are trying to acheive. Thia will help to decide that which metrics are more relevant for evaluation of model performance.

B. Understand the data: Go through the characteristics of the data you are working with.

C. Choose appropriate metrics: Select metrics that align with your goals and the characteristics of data.

D. Use multiple metrics: To get a comprehensive view of your model's performance.

12. Assessing the statistical significance of an insight involves determining whether the observed results are likely or unlikely to have occured by chance. Some steps tp access are:

A. Define the null hypothesis: It is a statement of no difference or no effect.

B. Choose a test: Choose a statistical test that is appropriate for the type of data and the problem at hand.

C. Calculate the test statistics: It is calculated with the help of chosen statistical test, which summarizes the evidence against the null hypothesis.

D. Determine the p-value: It is a probability of observing a test statistic.

E. Make a conclusion: Based on p-value, make a conclusion about the statistical significance of the insight.

13. Examples of data that doesn't have a Gaussian distribution, nor log-normal are- Bernoulli distribution, Poisson distribution, Exponential distribution, Cauchy distribution, Pareto distribution, etc.

14. An example where the median is better measure than the mean is when the data has outliers or extreme values. The mean is sensitive to outliers or extreme value, as they can skew average while median is not affected by outliers.

15. Likelihood: In statistics, likelihood is a measure of how likely a set of observations or data is to have occured given a specific probability distribution or model. The likelihood function is a measure of how well the model fits the data.