MACHINE LEARNING ANSWERS

1 - B) Low R-squared value for train-set and High R-squared value for test-set.

2 - B) Decision trees are highly prone to overfitting.

3 - C) Random Forest

4 - A) Accuracy

5 - B) Model B

6 – A and D

7 – B and C

8 - D) All of the above

9 - C) It is example of bagging technique

10 - The 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 only if 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 by chance.

11 - Lasso regression stands for Least Absolute Shrinkage and Selection Operator. It adds penalty term to the cost function. ... 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.

12 - Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

13 - To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

14 - Five metrics give us some hints about the goodness-of-fit of our model. The first two metrics, the Mean Absolute Error and the Root Mean Squared Error (also called Standard Error of the Regression), have the same unit as the original data.

15 – sensitivity = 0.8

Specificity = .96

Precision = .95

Recall= .8

Accuracy = .88

SQL ANSWERS

1 – A, C and D

2 – A, C and D

3 - D. SELECT # FROM SALES;

4 - C. Authorizing Access and other control over Database

5 - C. String

6 - B. COMMIT

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

8 - C. TABLE

9 - D. All of the mentioned

10 - A. ASC

11 - 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. ... In a traditional normalized database, we store data in separate logical tables and attempt to minimize redundant data.

12 - A database cursor is an identifier associated with a group of rows. It is, in a sense, a pointer to the current row in a buffer. You must use a cursor in the following cases: Statements that return more than one row of data from the database server: A SELECT statement requires a select cursor.

13 - Five types of SQL queries are 1) Data Definition Language (DDL) 2) Data Manipulation Language (DML) 3) Data Control Language(DCL) 4) Transaction Control Language(TCL) and, 5) Data Query Language (DQL)

14 - SQL constraints are used to specify rules for the data in a table. Constraints are used to limit the type of data that can go into a table. This ensures the accuracy and reliability of the data in the table.

15 - 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 ANSWERS

1 - d) All of the mentioned

2 - a) Discrete

3 - a) pdf

4 - b) median

5 - c) empirical mean

6 - b) standard deviation

7 - c) 0 and 1

8 - b) bootstrap

9 - b) summarized

10 - Histograms and box plots are graphical representations for the frequency of numeric data values. ... Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets.

11 - Good metrics are important to your company growth and objectives. Your key metrics should always be closely tied to your primary objective. ...

Good metrics can be improved. Good metrics measure progress, which means there needs to be room for improvement. ...

Good metrics inspire action.

12 - Create a null hypothesis.

Create an alternative hypothesis.

Determine the significance level.

Decide on the type of test you'll use.

Perform a power analysis to find out your sample size.

Calculate the standard deviation.

Use the standard error formula.

13 - There are many data types that follow a non-normal distribution by nature. Examples include: Weibull distribution, found with life data such as survival times of a product. Log-normal distribution, found with length data such as heights.

14 - In this case, analysts tend to use the mean because it includes all of the data in the calculations. However, if you have a skewed distribution, the median is often the best measure of central tendency. When you have ordinal data, the median or mode is usually the best choice.

15 - the chance that something will happen : probability There's very little likelihood of that happening.