Machine Learning
Answer Sheet

1. C
2. B
3. C
4. B
5. B
6. A, D
7. B, C
8. A, C
9. A, B

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 that accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not.

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.

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

A variance inflation factor (VIF) is a measure of the amount of **multicollinearity** in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results.

In general, a VIF above 10 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above. Sometimes a high VIF is no cause for concern at all. For example, you can get a high VIF by including products or powers from other variables in your regression, like x and x^2.

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

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 will help the gradient descent converge more quickly towards the minima.

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

Three statistics are used in Ordinary Least Squares (OLS) regression to evaluate model fit: R-squared, the overall F-test, and the Root Mean Square Error (RMSE). All three are based on two sums of squares: Sum of Squares Total (SST) and Sum of Squares Error (SSE).

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

Actual/Predicted	**True**	**False**
True	**1000**	**50**
1000	**250**	**1200**

Sensitivity = 1000/1050 = 0.95
Specificity = 1200/1450 = 0.83
Precision = 1000/1250 = 0.8
Recall = 1000/1050 = 0.95
Accuracy = 2200/2500 = 0.8

SQL
Answer Sheet

1. A, C, D
2. A, B, D
3. B
4. C
5. B
6. B
7. A
8. C
9. D
10. A

11. What is denormalization?

Denormalization is the process of adding precomputed redundant data to an otherwise normalized relational database to improve read performance of the database. Normalizing a database involves removing redundancy so only a single copy exists of each piece of information.

12. What is a database cursor?

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. What are the different types of the queries?

Five types of SQL queries are mentioned below.

1) Data Definition Language (DDL)
2) Data Manipulation Language (DML)
3) Data Control Language(DCL)
4) Transaction Control Language(TCL)
5) Data Query Language (DQL)

Data Definition Language(DDL) helps you to define the database structure or schema.
Data Manipulation Language (DML) allows you to modify the database instance by inserting, modifying, and deleting its data.
DCL (Data Control Language) includes commands like GRANT and REVOKE, which are useful to give "rights & permissions."
Transaction control language or TCL commands deal with the transaction within the database.
Data Query Language (DQL) is used to fetch the data from the database.

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.

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
Answer Sheet

1. D
2. A
3. A
4. C
5. C
6. A
7. C
8. B
9. B

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

Histograms and box plots are very similar in that they both help to visualize and describe numeric data. Although histograms are better in determining the underlying distribution of the data, box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space.

11. How to select metrics?

Metrics need to be chosen with care because when poorly chosen, they create an illusion of control. The metric data might deliver green dashboards while in reality, our organisation is not performing very well. Following are some important point how to select best metrics:-

Ensure your metrics are connected to your vision and mission
Take complexity into account (do not assume linear cause and effect)
Embed the metrics in the work so they do not become a separate goal
Measure “outside-in”
Focus on outcomes (impact) rather than output
Build up and revise your metrics as you go
How do you assess the statistical significance of an insight?
Statistical significance is often calculated with statistical hypothesis testing, which tests the validity of a hypothesis by figuring out the probability that your results have happened by chance.

12. Give examples of data that does not have a Gaussian distribution, nor log-normal.

Allocation of wealth among individuals
Values of oil reserves among oil fields (many small ones, a small number of large ones)

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

Income is the classic example of when to use the median instead of the mean because its distribution tends to be skewed. The median indicates that half of all incomes fall below 27581, and half are above it. For these data, the mean overestimates where most household incomes fall.

15. What is the Likelihood?

The likelihood is the probability that a particular outcome is observed when the true value of the parameter is , equivalent to the probability mass on ; it is not a probability density over the parameter . The likelihood, , should not be confused with , which is the posterior probability of given the data .