MACHINE LEARNING

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

2 Decision trees are highly prone to overfitting

3 Decision tree

4 Precision

5 Model B

6 Ridge, and Lasso

7 Decision Tree

8 Pruning and Restricting the max depth of the tree

9 A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points

B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

10 Models with tons of predictiors tends to perform better in sample than when tested out of sample . The adjusted R2 penalizes you for adding the extra predictor variables that don't improve the existing model .It can helpful in model selection .adjusted R2 wil equal R2 for one predictor variable.

11 Similar to the lasso regression, ridge regression puts a similar constraint on the coefficients by introducing a penalty factor. However, **while lasso regression takes the magnitude of the coefficients, ridge regression takes the square**. Ridge regression is also referred to as L2 Regularization. Lasso regression is also as L1 Regularization

12 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.

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.

14 1 mean squared error ,2 mean absolute error 3 **R-squared**

15 sensitivity=0 .95, specificity=0 .82 , precision=0.80 , recall=0.95 and accuracy=0.88 .

SQL

1 A. Commit C. Rollback D. Savepoint

2 A. Create , C. Drop, D. Alter

3 B. SELECT NAME FROM SALES;

4 C. Authorizing Access and other control over Database

5 B. Column Alias

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. It is an optimization technique that is applied after normalization.

12 A database cursor is **an identifier associated with a group of rows**

13 A few of the more popular queries include:

- Single-Table Select query.
- Multiple-Table Select query.
- Range query.
- Complex query.
- Totals query.
- Action query.
- Parameter query.
- Crosstab query.

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. If there is any violation between the constraint and the data action, the action is aborted.

15 SQL is **a feature that is applied to a field so that it can automatically generate and provide a unique value to every record that you enter into an SQL table**. This field is often used as the PRIMARY KEY column, where you need to provide a unique value for every record you add.

STATISTICS

1 d) All of the mentioned

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 b) summarized

10 Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques.

Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.

11 The key point is to **choose metrics that clearly indicate where you are now in relation to your goals**. Good metrics can be improved. Good metrics measure progress, which means there needs to be room for improvement. For example, reducing churn by 0.8% or increasing your activation rate by 3%

12 To assess statistical significance, you would use hypothesis testing. The null hypothesis and alternate hypothesis would be stated first. Second, you'd calculate the p-value, which is the likelihood of getting the test's observed findings if the null hypothesis is true. Finally, you would select the threshold of significance (alpha) and reject the null hypothesis if the p-value is smaller than the alpha — in other words, the result is statistically significant.

13 Given example of data that does not have a Gaussion distribution ,nor log normal. Allocation of wealth among individuals values of oil reserve among oil fields.

14 **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 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 .