Statistics Worksheet – 4

Q1) D → All of the mentioned

Q2) A → Discrete

Q3) A → Probability Density Function (pdf)

Q4) C → Mean

Q5) C → Empirical Mean

Q6) A → Variance

Q7) C → 0 & 1

Q8) B → Bootstrap

Q9) B → Summarized

Q10) Histograms are used to determine underlying probability distribution of data and its skewness whereas Boxplot can be used for comparing different datasets and also gives the idea of outliers and statistical data like Quartiles, max, min, median.

Q11)

- Choose metrics with clear owners
- Measure rates instead of totals

Q12) By performing hypothesis testing. p-value or the probability value gives the statistical significance. The range of p value lies between 0 to 1. p-value less than 0.05 is statistically significant

Q13) Any type of categorical data, exponential distribution.

Q14) Calculating salary of employee or rates of property where higher price in certain area of package of MD/CEO is used in calculation this might lead us to outliers which might tend to increase the Mean hence Median should be considered in such a case.

Q15) Likelihood is how well a sample provides support for particular values of a parameter in model.

Machine Learning Worksheet – 4

Q1) C → High R-Squared Value of train set and low R-Squared value of test set

Q2) B → Decision trees are prone to overfitting

Q3) C → Random Forest

Q4) A → Accuracy

Q5) B → Model B

Q6) A → Ridge Regression

Q7) C → Random Forest

Q8) A → Pruning

Q9) A

Q10) Additional input variables will make the R Squared increase even if there is no relationship between input variables and output variables. Adjusted R Squared looks at whether additional input variables are contributing to model or not

Q11) Ridge regression reduces the model complexity by coefficient shrinkage i.e. magnitude of coefficient decreases but does not attain value of zero, whereas in case of lasso regression our coefficient reduces to absolute 0. This property of lasso is known as feature selection.

Lasso is used when we have more no of features

Q12) Variance Inflation factor provides a measure of multicollinearity among the independent variables in a multi regression model. VIF value 1.

Q13) Scaling is required in cases where the data for a variable are in different units like kb and gb. In such cases when we bring the value in a single unit there might be huge differences in terms of values which might lead to skewness and increased outliers hence scaling is required.

Q14) R Square, Adjusted R Square, MSE, RMSE

Q15) TP = 1000, TN = 1200, FP = 50, FN = 250

Sensitivity = TP / (TP+FN) = 0.80

Specificity = TN / (TN+FP) = 0.96

Precision = TP / (TP+FP) = 0.95

Recall = TP/ (TP+FN) = 0.80

Accuracy = (TP+TN)/(TP+TN+FP+FN) = 0.88

SQL Worksheet – 4

Q1) B → Select

Q2) B → Select

Q3) B → SELECT NAME FROM SALES;

Q4) C → Authorizing Access and Other Control over Database

Q5) B → Column Alias

Q6) B → Commit

Q7) A → Parenthesis

Q8) C → Table

Q9) D → All

Q10) A → ASC

Q11) Denormalization is a database optimization technique where redundant data is added to one or more tables to get rid of complex joint operations. This is done to speed up data base access speed.

Q12) Database Cursor is used to pinpoint records in database. It shows the specific record in database that is being worked upon.

Q13) Types of queries = select, action, parameter, aggregate.

Q14) SQL Constraints are used to specify rules for the data in a table. Constraints are used to limit type of data that can go into the table ensuring accuracy and reliability of data in table.

Q15) Auto increment feature automatically generates a numerical primary key value for every new record inserted.