1. Explain some methods used to select important variables from a data set

Answer: Remove the correlated variables prior to selecting important variables

Use linear regression and select variables based on p values

Use Forward Selection, Backward Selection, Stepwise Selection

Use Random Forest, Xgboost and plot variable importance chart

Measure information gain for the available set of features and select top n features accordingly.

1. Which technique is used to calculate correlation between Continuous and Categorical variable

Answer: Analysis of Covariance (ANCOVA)

1. Would you remove correlated variables first while running PCA on huge dataset which contains highly correlated variables?

Answer: Yes. Discarding correlated variables have a substantial effect on PCA because, in presence of correlated variables, the variance explained by a particular component gets inflated.

1. What is Gini Coefficient?

Answer: Gini coefficient is the ratio between area between the ROC curve and the diagonal line & the area of the above triangle

Gini = 2\*AUC – 1

1. What are the plots which can be used for identifying ARIMA models orders of AR and MA plots in ARIMA modelling?

Answer: ACF ( Auto Correlation Function) and PACF (Partial Auto Correlation Function)

1. The relation between Gini coefficient and Area under the curve:  
    a. 2\*AUC-1  
    b. 2\*AUC+1  
    c. AUC^2-1  
    d. They are not related.  
   ans: a
2. Which of the following are methods for dimension reduction  
    a. PCA  
    b. Factor Analysis  
    c. NA imputation  
    d. Multicollinearity analysis  
   ans: a,b,d
3. What is recall  
    a. Sensitivity  
    b. Specificity  
    c. 1-Sensitivity  
    d. 1-Specificity  
   ans: a
4. When we are dealing with small sample sizes and if our sample size is increased which of the following are likely to happen?
   1. Bias will decrease
   2. Variance will decrease
   3. Bias and Variance decrease
   4. Bias and variance are independent of sample size.

Ans: b

1. Which SQL statement is used to create a table in a database.
   1. Create DB
   2. Create Table
   3. Create Database table
   4. Create DB Tab

ANS: b

1. K nearest neighbour algorithm is
   1. Supervised
   2. Unsupervised
   3. Parametric
   4. Non-Parametric

ANS: a,d

1. Can a logistic regression model with AIC = 4136 be considered as good?

Ans: AIC value on its own cannot be used for model selection. It is a value which can be used as a relative measure. E.g. a logistic regression model with AIC=2134 is better than a logistic regression model with AIC = 4136

1. How can you convert a linear distribution that has been turned to logarithmic back to linear?

Ans: By raising the Log value to its equivalent base. If Log10 was used for log transformation, then to convert back to original value can be obtained by

original value =10^log10(original value)

1. How do you define a stationary time-series?

Ans: A series which has constant mean and constant variance

1. What are the ways in which a non-stationary time-series can be made stationary?

Ans: By differencing, seasonal or non-seasonal

1. What does bias-variance trade-off mean?

Ans: A way of optimising model complexity such that the it does not overfit on the current training data and fits poorly on unseen data(variance), also, at the same time it should not be a overly generalised model with very poor fit on any data(bias).

1. What is bootstrapping?

Ans: Choosing samples out of a sample of data for generating various statistics of that sample such a mean distribution, standard deviation, prediction error etc.

1. What is heteroscedasticity?

Ans: It is an indicator that the variance of a distribution is not constant

1. If A is a matrix and AT is A transposed, then A.AT = AT.A is (a. True or b. False) for all matrices.

Ans: False

1. How to get a list of all the packages installed in R ?

Answer: Use the command **installed.packages()**

1. Give the general expression to create a matrix in R.

Answer: The general expression to create a matrix in R is - matrix(data, nrow, ncol, byrow, dimnames). Using a matrix function

1. How to convert multiple columns from integer to factor in R? (avoid using for loop)

Ans: apply(dataset,2,function(x) as.factor(as.character()))

1. Difference between set.seed(100) and set.seed(123) in R

Ans: No difference

1. Packages used for connecting R to a database.

Ans: RODBC, RJDBC, RSQLServer, rsqlserver, RMySQL, ROracle (answering anyone is enough)

1. Function used for calling a .R file and executing it.

Ans: source()

1. Function for saving the current R workspace.

Ans: save.image()

1. Syntax for subsetting a dataframe by removing certain column names and not by column numbers.

Ans: data[,-which(names(data) %in% c(‘Column1’,’Column2’,….)]

data[,-c( ‘Column1’,’Column2’,….)] is wrong