**Data Science Questions and Answers – caret-1**

**This set of Data Science Multiple Choice Questions & Answers (MCQs) focuses on “caret-1”.**

1. Which of the following can be used to generate balanced cross–validation groupings from a set of data ?  
a) createFolds  
b) createSample  
**c) createResample**  
d) none of the Mentioned  
View Answer

Answer: a  
Explanation: createResample can be used to make simple bootstrap samples.

2. Point out the wrong statement:  
a) Simple random sampling of time series is probably the best way to resample times series data.  
b) Three parameters are used for time series splitting  
c) Horizon parameter is the number of consecutive values in test set sample  
d) All of the Mentioned  
View Answer

Answer: a  
Explanation: Simple random sampling of time series is probably not the best way to resample times series data.

3. Which of the following function can be used to maximize the minimum dissimilarities ?  
a) sumDiss  
b) minDiss  
c) avgDiss  
**d) all of the Mentioned**  
View Answer

Answer: d  
Explanation: sumDiss can be used to maximize the total dissimilarities.

4. Which of the following function can create the indices for time series type of splitting ?  
a) newTimeSlices  
**b) createTimeSlices**  
c) binTimeSlices  
d) none of the Mentioned  
View Answer

Answer: b  
Explanation: Rolling forecasting origin techniques are associated with time series type of splitting.

5. Point out the correct statement:  
a) Asymptotics are used for inference usually  
b) caret includes several functions to pre-process the predictor data  
c) The function dummyVars can be used to generate a complete set of dummy variables from one or more factors  
d**) All of the Mentioned**  
View Answer

Answer: d  
Explanation: The function dummyVars takes a formula and a data set and outputs an object that can be used to create the dummy variables using the predict method.

6. Which of the following can be used to create sub–samples using a maximum dissimilarity approach ?  
a) minDissim  
b) maxDissim  
c) inmaxDissim  
**d) all of the Mentioned**  
View Answer

Answer: b  
Explanation: Splitting is based on the predictors.

7. caret does not use the proxy package.  
a) True  
b**) False**  
View Answer

Answer: b  
Explanation: caret uses the proxy package.

8. Which of the following function can be used to create balanced splits of the data ?  
a) newDataPartition  
**b) createDataPartition**  
c) renameDataPartition  
d) none of the Mentioned  
View Answer

Answer: b  
Explanation: If the y argument to this function is a factor, the random sampling occurs within each class and should preserve the overall class distribution of the data.

9. Which of the following package tools are present in caret ?  
a) pre-processing  
b) feature selection  
c) model tuning  
**d) all of the Mentioned**  
View Answer

Answer: d  
Explanation: There are many different modeling functions in R.

10. caret stands for classification And regression training.  
**a) True**  
b) False  
View Answer

Answer: a  
Explanation: The caret package is a set of functions that attempt to streamline the process for creating predictive models.

**Data Science Questions and Answers – caret-2**

**This set of Data Science MCQs focuses on “Caret”.**

1. Which of the following function is a wrapper for different lattice plots to visualize the data ?  
a) levelplot  
**b) featurePlot**  
c) plotsample  
d) none of the Mentioned  
View Answer

Answer: b  
Explanation: featurePlot is used for data visualization in caret.

2. Point out the wrong statement:  
**a) In every situation, the data generating mechanism can create predictors that only have a single unique value**  
b) Predictors might have only a handful of unique values that occur with very low frequencies  
c) The function findLinearCombos uses the QR decomposition of a matrix to enumerate sets of linear combinations  
d) All of the Mentioned  
View Answer

Answer: a  
Explanation: In some situations, the data generating mechanism can create predictors that only have a single unique value.

3. Which of the following function can be used to identify near zero-variance variables ?  
a) zeroVar  
b) nearVar  
**c) nearZeroVar**  
d) all of the Mentioned  
View Answer

**Answer: c  
Explanation: The saveMetrics argument can be used to show the details and usually defaults to FALSE.**

4. Which of the following function can be used to flag predictors for removal ?  
a) searchCorrelation  
b) findCausation  
**c) findCorrelation**  
d) none of the Mentioned  
View Answer

Answer: c  
Explanation: Some models thrive on correlated predictors.

5. Point out the correct statement:  
a) findLinearColumns will also return a vector of column positions can be removed to eliminate the **linear dependencies  
b) findLinearCombos will return a list that enumerates dependencies**  
c) The function findLinearRows can be used to generate a complete set of row variables from one factor  
d) none of the Mentioned  
View Answer

Answer: b  
Explanation: For each linear combination, it will incrementally remove columns from the matrix and test to see if the dependencies have been resolved.

6. Which of the following can be used to impute data sets based only on information in the training set. ?  
a) postProcess  
**b) preProcess**  
c) process  
d) all of the Mentioned  
View Answer

**Answer: b  
Explanation: This can be done with K-nearest neighbors.**

7. The function preProcess estimates the required parameters for each operation.  
a) True  
b) False  
**View Answer**

**Answer: a  
Explanation: predict.preProcess is used to apply them to specific data sets.**

8. Which of the following can also be used to find new variables that are linear combinations of the original set with independent components ?  
**a) ICA**  
b) SCA  
c) PCA  
d) None of the Mentioned  
View Answer

Answer: a  
Explanation: ICA stands for independent component analysis.

9. Which of the following function is used to generate the class distances ?  
a) preprocess.classDist  
**b) predict.classDist**  
c) predict.classDistance  
d) all of the Mentioned  
View Answer

Answer: b  
Explanation: By default, the distances are logged.

10. The preProcess class can be used for many operations on predictors.  
**a) True**  
b) False  
View Answer

Answer: a  
Explanation: Operations include centering and scaling.

**Data Science Questions and Answers – caret-3**

**This set of Data Science Multiple Choice Questions & Answers focuses on “Caret”.**

1. varImp is a wrapper around the evimp function in the \_\_\_\_\_\_\_ package.  
a) numpy  
b**) earth**c) plot  
d) none of the Mentioned  
View Answer

Answer: b  
Explanation: The earth package is an implementation of Jerome Friedman’s Multivariate Adaptive Regression Splines.

2. Point out the wrong statement:  
a) The trapezoidal rule is used to compute the area under the ROC curve  
b) For regression, the relationship between each predictor and the outcome is evaluated  
c) An argument, para, is used to pick the model fitting technique  
d) All of the Mentioned  
View Answer

Answer: c  
Explanation: An argument, nonpara, is used to pick the model fitting technique.

3. Which of the following curve analysis is conducted on each predictor for classification ?  
a) NOC  
**b) ROC**  
c) COC  
d) All of the Mentioned  
View Answer

Answer: b  
Explanation: For two class problems, a series of cutoffs is applied to the predictor data to predict the class.

4. Which of the following function tracks the changes in model statistics ?  
**a) varImp**  
b) varImpTrack  
c) findTrack  
d) none of the Mentioned  
View Answer

Answer: a  
Explanation: GCV change value can also be tracked.

5. Point out the correct statement:  
**a) The difference between the class centroids and the overall centroid is used to measure the variable influence**  
b) The Bagged Trees output contains variable usage statistics  
c) Boosted Trees uses different approach as a single tree  
d) None of the Mentioned  
View Answer

Answer: a  
Explanation: The larger the difference between the class centroid and the overall center of the data, the larger the separation between the classes.

6. Which of the following model model include a backwards elimination feature selection routine ?  
a) MCV  
**b) MARS**c) MCRS  
d) All of the Mentioned  
View Answer

Answer: b  
Explanation: MARS stands for Multivariate Adaptive Regression Splines.

7. The advantage of using a model-based approach is that is more closely tied to the model performance.  
**a) True**  
b) False  
View Answer

Answer: a  
Explanation: Model-based approach is able to incorporate the correlation structure between the predictors into the importance calculation.

**8. Which of the following model sums the importance over each boosting iteration ?  
a) Boosted Trees**  
b) Bagged Trees  
c) Partial Least Squares  
d) None of the Mentioned  
View Answer

Answer: a  
Explanation: gbm package can be used here.

9. Which of the following argument is used to set importance values ?  
**a) scale**  
b) set  
c) value  
d) all of the Mentioned  
View Answer

Answer: a  
Explanation: All measures of importance are scaled to have a maximum value of 100.

10. For most classification models, each predictor will have a separate variable importance for each class.  
**a) True**  
b) False  
View Answer

Answer: a  
Explanation: The exceptions are classification trees, bagged trees and boosted trees.