This set of Data Science Multiple Choice Questions & Answers (MCQs) focuses on â€œcaret â€“ 1â€�.

1. Which of the following can be used to generate balanced crossâ€“validation groupings from a set of data?
a) createFolds
b) createSample
c) createResample
d) none of the mentioned

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

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

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

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

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 subâ€“samples using a maximum dissimilarity approach?
a) minDissim
b) maxDissim
c) inmaxDissim
d) all of the mentioned

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

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

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

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

Answer: d
Explanation: There are many different modeling functions in R.
Answer: a
Explanation: The caret package is a set of functions that attempt to streamline the process for creating predictive models.
To practice all areas of Data Science, Here is complete set of 1000+ Multiple Choice Questions and Answers .
Participate in the Sanfoundry Certification contest to get free Certificate of Merit. Join our social networks below and stay updated with latest contests, videos, internships and jobs!

This set of Data Science MCQs focuses on â€œCaret â€“ 2â€�.

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

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

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

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

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

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

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

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

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

Answer: b
Explanation: By default, the distances are logged.
Answer: a
Explanation: Operations include centering and scaling.
To practice all areas of Data Science, Here is complete set of 1000+ Multiple Choice Questions and Answers .
Participate in the Sanfoundry Certification contest to get free Certificate of Merit. Join our social networks below and stay updated with latest contests, videos, internships and jobs!

This set of Data Science Multiple Choice Questions & Answers focuses on â€œCaret â€“ 3â€�.

1. varImp is a wrapper around the evimp function in the _______ package.
a) numpy
b) earth
c) plot
d) none of the mentioned

Answer: b
Explanation: The earth package is an implementation of Jerome Friedmanâ€™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

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

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

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

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

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

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

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

Answer: a
Explanation: All measures of importance are scaled to have a maximum value of 100.
Answer: a
Explanation: The exceptions are classification trees, bagged trees and boosted trees.
To practice all areas of Data Science, Here is complete set of 1000+ Multiple Choice Questions and Answers .
Participate in the Sanfoundry Certification contest to get free Certificate of Merit. Join our social networks below and stay updated with latest contests, videos, internships and jobs!

This set of Data Science Multiple Choice Questions & Answers (MCQs) focuses on â€œPrediction Motivationâ€�.

1. Which of the following is the valid component of the predictor?
a) data
b) question
c) algorithm
d) all of the mentioned

Answer: d
Explanation: A prediction is a statement about the future.

2. Point out the wrong statement.
a) In Sample Error is also called generalization error
b) Out of Sample Error is the error rate you get on the new dataset
c) In Sample Error is also called resubstitution error
d) All of the mentioned

Answer: a
Explanation: Out of Sample Error is also called generalization error.

3. Which of the following is correct order of working?
a) questions->input data ->algorithms
b) questions->evaluation ->algorithms
c) evaluation->input data ->algorithms
d) all of the mentioned

Answer: a
Explanation: Evaluation is done in the last.

4. Which of the following shows correct relative order of importance?
a) question->features->data->algorithms
b) question->data->features->algorithms
c) algorithms->data->features->question
d) none of the mentioned

Answer: b
Explanation: Garbage in should be equal to garbage out.

5. Point out the correct statement.
a) In Sample Error is the error rate you get on the same dataset used to model a predictor
b) Data have two parts-signal and noise
c) The goal of predictor is to find signal
d) None of the mentioned

Answer: d
Explanation: Perfect in sample prediction can be built.

6. Which of the following is characteristic of best machine learning method?
a) Fast
b) Accuracy
c) Scalable
d) All of the mentioned

Answer: d
Explanation: There is always a trade-off in prediction accuracy.

7. True positive means correctly rejected.
a) True
b) False

Answer: b
Explanation: True positive means correctly identified.

8. Which of the following trade-off occurs during prediction?
a) Speed vs Accuracy
b) Simplicity vs Accuracy
c) Scalability vs Accuracy
d) None of the mentioned

Answer: d
Explanation: Interpretability also matters during prediction.

9. Which of the following expression is true?
a) In sample error < out sample error
b) In sample error > out sample error
c) In sample error = out sample error
d) All of the mentioned

Answer: a
Explanation: Out of sample error is given more importance.
Answer: a
Explanation: Backtesting is the process of applying a trading strategy or analytical method to historical data to see how accurately the strategy or method would have predicted actual results.
To practice all areas of Data Science, Here is complete set of 1000+ Multiple Choice Questions and Answers .
Participate in the Sanfoundry Certification contest to get free Certificate of Merit. Join our social networks below and stay updated with latest contests, videos, internships and jobs!

This set of Data Science Multiple Choice Questions & Answers (MCQs) focuses on â€œCross Validationâ€�.

1. Which of the following is correct use of cross validation?
a) Selecting variables to include in a model
b) Comparing predictors
c) Selecting parameters in prediction function
d) All of the mentioned

Answer: d
Explanation: Cross-validation is also used to pick type of prediction function to be used.

2. Point out the wrong combination.
a) True negative=correctly rejected
b) False negative=correctly rejected
c) False positive=correctly identified
d) All of the mentioned

Answer: c
Explanation: False positive means incorrectly identified.

3. Which of the following is a common error measure?
a) Sensitivity
b) Median absolute deviation
c) Specificity
d) All of the mentioned

Answer: d
Explanation: Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function.

4. Which of the following is not a machine learning algorithm?
a) SVG
b) SVM
c) Random forest
d) None of the mentioned

Answer: a
Explanation: SVM stands for scalable vector machine.

5. Point out the wrong statement.
a) ROC curve stands for receiver operating characteristic
b) Foretime series, data must be in chunks
c) Random sampling must be done with replacement
d) None of the mentioned

Answer: d
Explanation: Random sampling with replacement is the bootstrap.

6. Which of the following is a categorical outcome?
a) RMSE
b) RSquared
c) Accuracy
d) All of the mentioned

Answer: c
Explanation: RMSE stands for Root Mean Squared Error.

7. For k cross-validation, larger k value implies more bias.
a) True
b) False

Answer: b
Explanation: For k cross-validation, larger k value implies less bias.

8. Which of the following method is used for trainControl resampling?
a) repeatedcv
b) svm
c) bag32
d) none of the mentioned

Answer: a
Explanation: repeatedcv stands for repeated cross-validation.

9. Which of the following can be used to create the most common graph types?
a) qplot
b) quickplot
c) plot
d) all of the mentioned

Answer: a
Explanation: qplot() is short for a quick plot.
Answer: a
Explanation: Larger k value implies more variance.
To practice all areas of Data Science, Here is complete set of 1000+ Multiple Choice Questions and Answers .
Participate in the Sanfoundry Certification contest to get free Certificate of Merit. Join our social networks below and stay updated with latest contests, videos, internships and jobs!

This set of Data Science Multiple Choice Questions & Answers (MCQs) focuses on â€œPredicting with Regressionâ€�.

1. Predicting with trees evaluate _____________ within each group of data.
a) equality
b) homogeneity
c) heterogeneity
d) all of the mentioned

Answer: b
Explanation: Predicting with trees is easy to interpret.

2. Point out the wrong statement.
a) Training and testing data must be processed in different way
b) Test transformation would mostly be imperfect
c) The first goal is statistical and second is data compression in PCA
d) All of the mentioned

Answer: a
Explanation: Training and testing data must be processed in same way.

3. Which of the following method options is provided by train function for bagging?
a) bagEarth
b) treebag
c) bagFDA
d) all of the mentioned

Answer: d
Explanation: Bagging can be done using bag function as well.

4. Which of the following is correct with respect to random forest?
a) Random forest are difficult to interpret but often very accurate
b) Random forest are easy to interpret but often very accurate
c) Random forest are difficult to interpret but very less accurate
d) None of the mentioned

Answer: a
Explanation: Random forest is top performing algorithm in prediction.

5. Point out the correct statement.
a) Prediction with regression is easy to implement
b) Prediction with regression is easy to interpret
c) Prediction with regression performs well when linear model is correct
d) All of the mentioned

Answer: d
Explanation: Prediction with regression gives poor performance in non linear settings.

6. Which of the following library is used for boosting generalized additive models?
a) gamBoost
b) gbm
c) ada
d) all of the mentioned

Answer: a
Explanation: Boosting can be used with any subset of classifier.

7. The principal components are equal to left singular values if you first scale the variables.
a) True
b) False

Answer: b
Explanation: The principal components are equal to left singular values if you first scale the variables.

8. Which of the following is statistical boosting based on additive logistic regression?
a) gamBoost
b) gbm
c) ada
d) mboost

Answer: a
Explanation: mboost is used for model based boosting.

9. Which of the following is one of the largest boost subclass in boosting?
a) variance boosting
b) gradient boosting
c) mean boosting
d) all of the mentioned

Answer: b
Explanation: R has multiple boosting libraries.
Answer: b
Explanation: PCA is most useful for linear type models.
To practice all areas of Data Science, Here is complete set of 1000+ Multiple Choice Questions and Answers .
Participate in the Sanfoundry Certification contest to get free Certificate of Merit. Join our social networks below and stay updated with latest contests, videos, internships and jobs!

This set of Data Science Multiple Choice Questions & Answers (MCQs) focuses on â€œModel Based Predictionâ€�.

1. Which of the following is correct about regularized regression?
a) Can help with bias trade-off
b) Cannot help with model selection
c) Cannot help with variance trade-off
d) All of the mentioned

Answer: a
Explanation: Regularized regression does not perform as well as random forest.

2. Point out the wrong statement.
a) Model based approach may be computationally convenient
b) Model based approach use Bayes theorem
c) Model based approach are reasonably inaccurate on real problems
d) All of the mentioned

Answer: c
Explanation: Model based approach are reasonably accurate on real problems.

3. Which of the following methods are present in caret for regularized regression?
a) ridge
b) lasso
c) relaxo
d) all of the mentioned

Answer: d
Explanation: In caret one can tune over the no of predictors to retain instead of defined values for penalty.

4. Which of the following method can be used to combine different classifiers?
a) Model stacking
b) Model combining
c) Model structuring
d) None of the mentioned

Answer: a
Explanation: Model ensembling is also used for combining different classifiers.

5. Point out the correct statement.
a) Combining classifiers improves interpretability
b) Combining classifiers reduces accuracy
c) Combining classifiers improves accuracy
d) All of the mentioned

Answer: c
Explanation: You can combine classifier by averaging.

6. Which of the following function provides unsupervised prediction?
a) cl_forecast
b) cl_nowcast
c) cl_precast
d) none of the mentioned

Answer: d
Explanation: cl_predict function is clue package provides unsupervised prediction.

7. Model based prediction considers relatively easy version for covariance matrix.
a) True
b) False

Answer: b
Explanation: Model based prediction considers relatively easy version for covariance matrix.

8. Which of the following is used to assist the quantitative trader in the development?
a) quantmod
b) quantile
c) quantity
d) mboost

Answer: a
Explanation: Quandl package is similar to quantmod.

9. Which of the following function can be used for forecasting?
a) predict
b) forecast
c) ets
d) all of the mentioned

Answer: b
Explanation: Forecasting is the process of making predictions of the future based on past and present data and analysis of trends.
Answer: b
Explanation: Predictive analytics goes beyond forecasting.
To practice all areas of Data Science, Here is complete set of 1000+ Multiple Choice Questions and Answers .