Buskirk: ASA Webinar, September 19, 2019 Example for Classification Trees

The ASA2019LabData.RData is an R workspace that contains the data objects we will need to work with for these examples. To load the workspace, choose "Load Workspace" from the File menu in the basic R console or "Load Workspace" under the Session tab in R studio. The ASA2019LabData.R contains R code that can be used to generate the various models of interest and can be loaded into your R session by opening a new Script window within R.

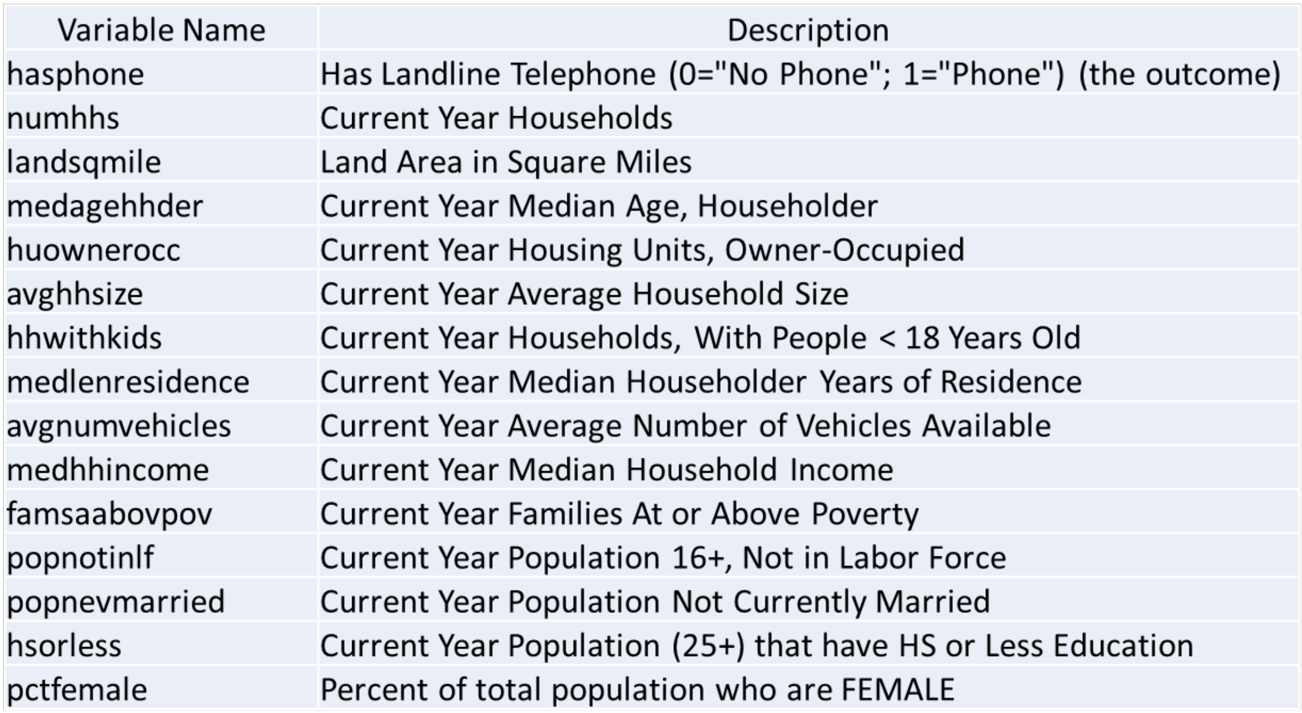
KNN Example: The knntrain.x and knntrain.y, knntest.x and knntest.y data frames included in this workspace represent random splits of a larger random sample of residential addresses from across the United States along with appended census block group level information related to various covariates of interest.

The knntrain.x and knntest.x data frames in the workspace represent the x/predictors/covariates for the training and test sets, respectively.

The knntrain.y and knntest.y vectors also represent the Y/outcome of interest stored as a factor where 0 indicates no landline phone and 1 indicates has a landline phone.

The goal is to determine the optimal value of "k" for a K-nearest neighbor model that can predict whether or not addresses from a new ABS sample will have a phone number available.

VARIABLE NAMES and DESCRIPTIONS:



Variable Name	Description
hasphone	Has Landline Telephone (0="No Phone"; 1="Phone") (the outcome)
numhhs	Current Year Households
landsqmile	Land Area in Square Miles
medagehhder	Current Year Median Age, Householder
huownerocc	Current Year Housing Units, Owner-Occupied
avghhsize	Current Year Average Household Size
hhwithkids	Current Year Households, With People < 18 Years Old
medlenresidence	Current Year Median Householder Years of Residence
avgnumvehicles	Current Year Average Number of Vehicles Available
medhhincome	Current Year Median Household Income
famsaabovpov	Current Year Families At or Above Poverty
popnotinlf	Current Year Population 16+, Not in Labor Force
popnevmarried	Current Year Population Not Currently Married
hsorless	Current Year Population (25+) that have HS or Less Education
pctfemale	Percent of total population who are FEMALE

Task A: Using the knntrain.x and knntain.y dataframes, determine the optimal value of K to be used for k-nearest neighbors based on a 10 fold cross validation with possible values of k ranging from 1 to 21 (odd values only). Provide a plot illustrating the cross validation results (accuracy and error bars) and the final value of k.

NOTE: PLEASE USE THE set.seed(2019) before you perform the 10-fold cross validation!

Task B: Now using the optimal value of k generate a knn model using the training data and then determine the overall accuracy of the model using the knntest.y data.

NOTE: Please use the set.seed(711) before you generate the KNN model.

Hint 1: **predobject<-knn(training X's, testing X's, training Y's, k=value from task A)**

Hint 2: overall accuracy computed as: **mean(predobject==testing Ys)**

Task C: What is the confusion matrix for this model using the test set of the y's and their predicted values? From this, what are the estimates of the sensitivity and specificity for the knn model?

Hint: confusionMatrix(predicted values first, actual values second)

Task D: Are the variables in knntrain.x on the same range?

Hint: You can see the min and max values by typing **summary(knntrain.x)** or computing the ranges by typing: **diff(apply(knntrain.x,2,FUN=range))**

Hint 1: Note – you don't need to standardize the y variables here! Why?

Hint 2: to standardize the variables in the training/test sets use code like this:

knntrainxSTD<-apply(TRAIN X's, 2, FUN=range01)

knntestxSTD <- (same thing as above, but you will need TEST X's)

Task E: Repeat Task A using a set of covariates that have been standardized. How does your optimal value of k compare to Task A.

Note: PLEASE USE set.seed(72019) before you perform the 10-fold cross validation here!

Hint: you will use the syntax from task A, but reset the seed to be 72019 by typing: > **set.seed(72019)**

Task F: Repeat task B using the set of covariates that have been standardized. This time, use seed 107.

Hint: you will use the syntax from task B, but reset the seed to be 107 by typing: > **set.seed(107)**

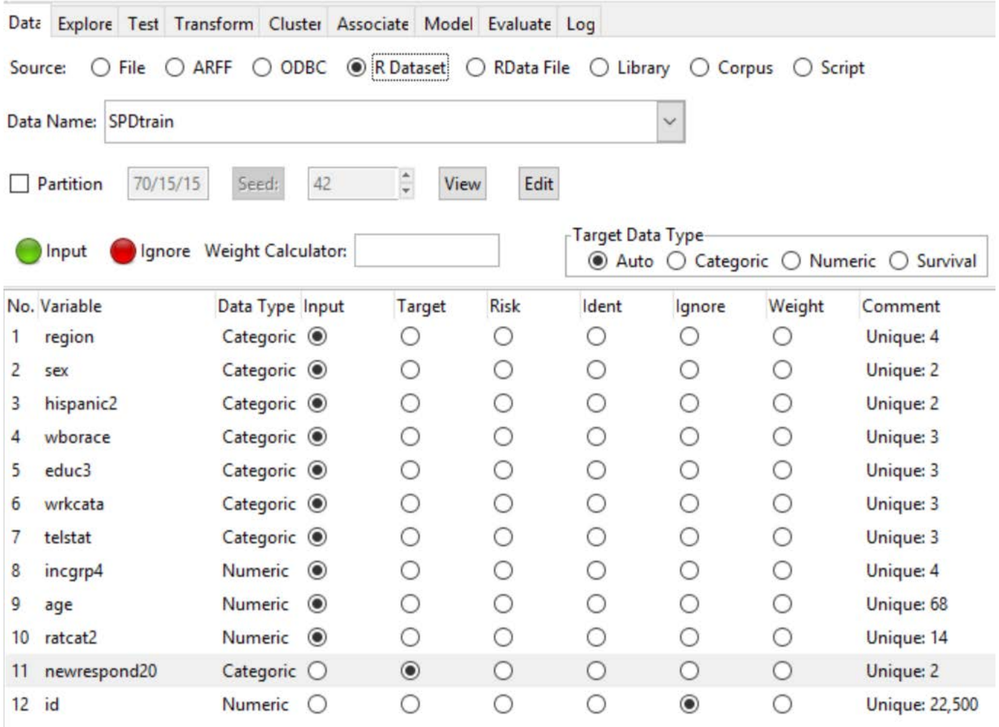
Task G: Repeat Task C using the standardize set of covariates.

CART Example: This example leverages the power of rattle but the corresponding R code to execute similar models can be found in the ASA2019LabData.R script file. The ASA2019TreeData.RData also contains **the data SPDtrain and SPDtest which we will use for this part to build a tree model to predict response status from a collection of covariate values that have been described in this article:** https://www.surveypractice.org/article/2718-an-introduction-to-machine-learning-methods-for-survey-researchers.

The SPDtrain and SPDtest data sets were created from a larger file SPData2 also contained in the R workspace file and details of the creation of these data sets and pre-processing are available in the file ASA2019TreeData.R.

The outcome of interest is newrespond20 and the predictors we want to use are all of the demographic variables including age, income, race, etc. that are listed in the table in the article referenced above. The id number is not a predictor, but is included in the file just for reference.

Task A: Process these data in Rattle to make sure we ignore the id number and select the newrespon20 as the outcome variable of interest. Also, we will NOT create a partition of these data for our model creation.



Source: ○ File ○ ARFF ○ ODBC ● R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Data Name: SPDtrain

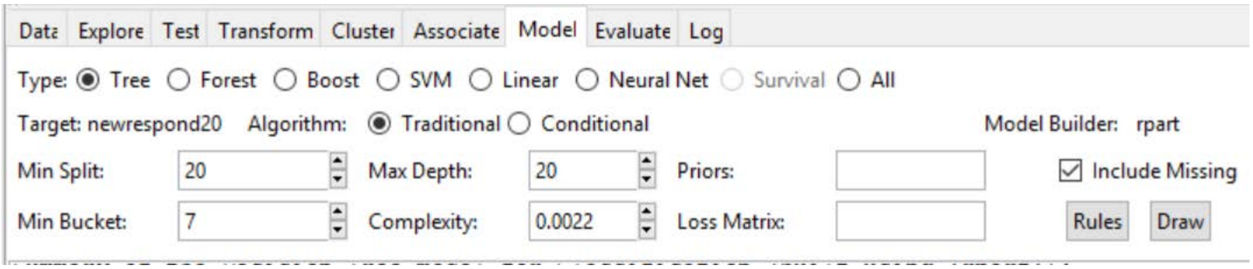
□ Partition 70/15/15 Seed: 42

● Input ● Ignore Weight Calculator:

Target Data Type ● Auto ○ Categoric ○ Numeric ○ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	region	Categoric	●	○	○	○	○	○	Unique: 4
2	sex	Categoric	●	○	○	○	○	○	Unique: 2
3	hispanic2	Categoric	●	○	○	○	○	○	Unique: 2
4	wborace	Categoric	●	○	○	○	○	○	Unique: 3
5	educ3	Categoric	●	○	○	○	○	○	Unique: 3
6	wrkcata	Categoric	●	○	○	○	○	○	Unique: 3
7	telstat	Categoric	●	○	○	○	○	○	Unique: 3
8	incgrp4	Numeric	●	○	○	○	○	○	Unique: 4
9	age	Numeric	●	○	○	○	○	○	Unique: 68
10	ratcat2	Numeric	●	○	○	○	○	○	Unique: 14
11	newrespond20	Categoric	○	●	○	○	○	○	Unique: 2
12	id	Numeric	○	○	○	○	●	○	Unique: 22,500

Task B: We will use the data file SPDtrain to create a tree model with the following parameters set: (we have already determined the value of the complexity parameter for this exercise.



Type: ● Tree ○ Forest ○ Boost ○ SVM ○ Linear ○ Neural Net ○ Survival ○ All

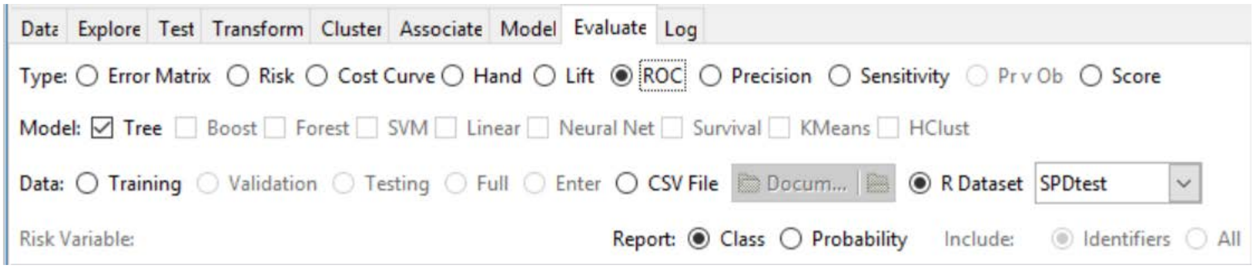
Target: newrespond20 Algorithm: ● Traditional ○ Conditional Model Builder: rpart

Min Split: 20 Max Depth: 20 Priors: ☑ Include Missing

Min Bucket: 7 Complexity: 0.0022 Loss Matrix:

Task C: We will compute this model and then plot it.

Task D: Using the data in SPDtest, we will compute the AUC and the confusion matrix from which we can compute the model's estimated accuracy, sensitivity and specificity.



Type: ○ Error Matrix ○ Risk ○ Cost Curve ○ Hand ○ Lift ● ROC ○ Precision ○ Sensitivity ○ Pr v Ob ○ Score

Model: ☑ Tree □ Boost □ Forest □ SVM □ Linear □ Neural Net □ Survival □ KMeans □ HClust

Data: ○ Training ○ Validation ○ Testing ○ Full ○ Enter ○ CSV File ● R Dataset SPDtest

Risk Variable: Report: ● Class ○ Probability Include: ● Identifiers ○ All

Task E: Repeat task B, C and D except with the following values of the Complexity parameter: .0009, .001, 0.0250. For each of these values (and the one we used initially, .0022) compare prediction measures of accuracy including overall Accuracy, Sensitivity, Specificity and the AUC.