Content Evaluator (CE) Tool for R – version 2

The Content Evaluator (CE) tool – version 2 is the latest version of the Merkle developed process for automating preliminary predictive modeling. It includes sampling, EDA and recoding, variable reduction, model training and model evaluation.

While CE can automate much of the “science” part of modeling, it cannot do the “art” of modeling. The list of variables that comes out of the CE process should be used as a starting point for you to use in building your final model. No set of code can replicate what a good analyst can intuit.

All code can be found on the Shared drive at /mnt/projects/shared/pst\_qmgisi/Modeling/RCE/. The files you will need are:

* CE Control Program v2.R
* contents.R
* DS Variable - R.xlsm

Copy these to your project directory to use as your working files that you will modify.

# Variable List Setup

The CE process needs to be told how to process each of your independent variables. This is done by specifying the variable type for each variable:

* Binary – these are variables with values of Y/N or 0/1. The code will recode all records with a value of Y, y or 1 to 1 and everything else to 0. These variables can be character or numeric.
* Nominal – these are variables where there is no inherent ordering to the values of a variable. Examples are occupation and Mosaic type. These variables can be character or numeric.
* Ordinal – these are variables with a limited number of values that have an inherent ordering. Examples are education level and number of children. These variables must be numeric.
* Continuous – these are variables with an unlimited number of values that have inherent ordering. Examples are raw income and total purchase dollars. These variables must be numeric.

A tool has been created to help you with creating the necessary lists.

1. Open the R program *contents.R* that you have copied to your project directory.

require(XLConnect)

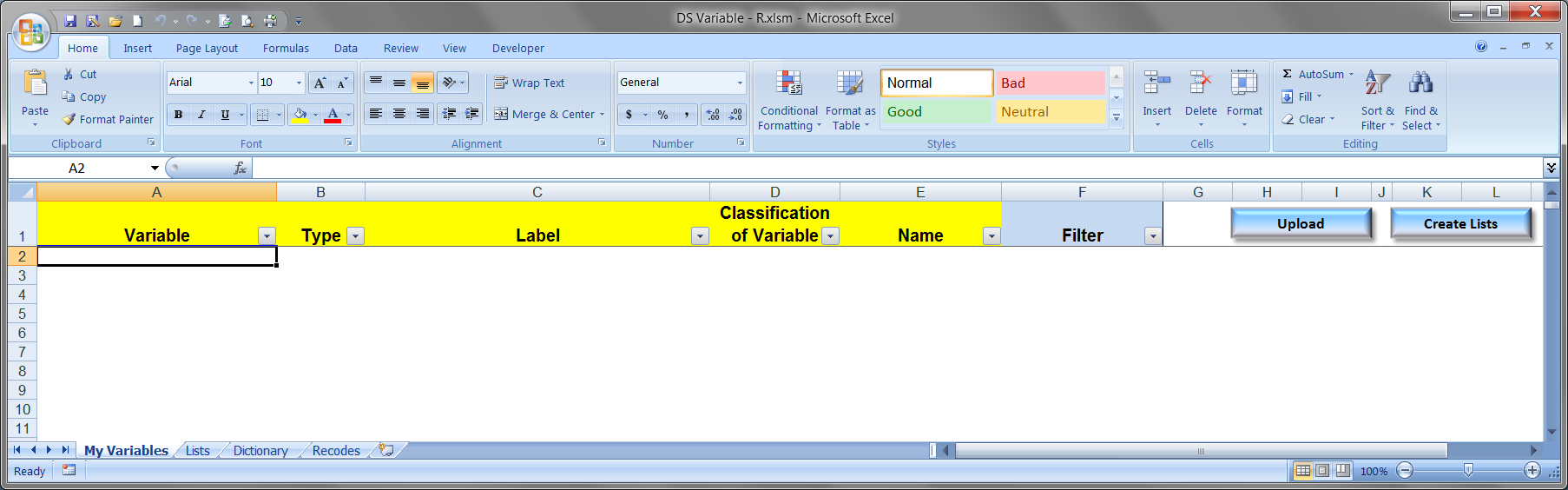
setwd("/mnt/projects/locked/pst\_qmgisi/Modeling/CE Rebuild/binary/ce\_output")

lib <- "/mnt/projects/locked/pst\_qmgisi/Modeling/CE Rebuild/sasdata"

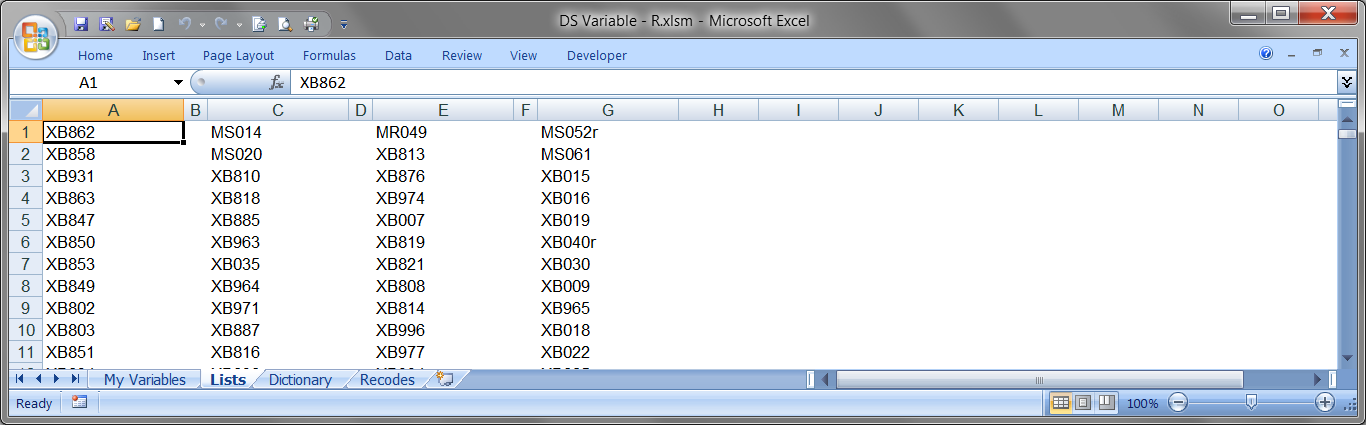
inData <- file.path(lib,"cont\_file.xdf")

* 1. Run the require statement to load the XLConnect package
  2. Set your working directory to the location you want to write out the Excel output. This should be the working directory you will be using in the RCE process.
  3. Set **lib** to the file path of the dataset you want to analyze
  4. Set **inData** to the dataset you want to analyze. This code assumes you are working with an xdf file rather than an R dataframe.
  5. Run the code. This will write out an Excel workbook called *var\_cntents.xlsx* to the working directory

1. Open the Excel workbook DS Variable - R.xlsm that you have copied to your project directory.
   1. Enable the macros



* 1. Click the button called **Upload** on the *My Variables* page. This will open a file selector box. Navigate to the folder **loc**, click the *var\_contents.xlsx* file and click open. A pop-up will inform you that the file has been successfully uploaded. Click OK to close.
  2. You will now see a table with all of your available variables. If the variable is a DataSource variable, it will pre-populate the Classification and Name. If the variable is in the standard DS recodes (available in step 1 of CE) then the Name field will use the name for the recoded variable rather than the raw variable. You can change the Classification of any of these variables. We have just tried to give you a good starting point and minimize the amount of work you have to do.
  3. If the variable is not in the DataSource variable list, the Classification and Name will be blank. You need to set these. The Classification field has a drop-down list of options for you to select from. The possible values are binary, nominal, ordinal, continuous, idvar, depvar and exclude. Idvar is used for the variable that identifies unique records. Depvar is used for your dependent variable. Exclude is used to exclude the variable from the list of independent variables. If you change the Classification to anything other than exclude, the Name field will populate with the variable name.  
       
     Column F is set up to help you filter the variable names if you have a lot of client specific variables. Change cell F1 to a string to look for in column A to identify a block of similar variables. For example, if all of your binary variables end in \_flg, change F1 to \_flg. Then filter for rows that have a blank for the classification (are not yet defined) and is not blank in column F. Set the first row to binary and copy to the rest of the selected rows. Continue with other strings that can be used to identify other blocks of similar variables. See the "Tips" document for a more detailed example.
  4. Once you have completed the classification of all of your variables, click the button **Create Lists** to populate the lists of variables by type in the sheet *Lists* and create the necessary text files. Files will be created in folder **loc**.



The standard document names are:

* + - binary – binvar\_list.txt
    - nominal – nomvar\_list.txt
    - ordinal – ordvar\_list.txt
    - continuous – contvar\_list.txt

The code will erase any existing files with these names before creating new ones.

* 1. Save your workbook so if you need to make changes you will not need to start from scratch unless the variables available changes.

# RCE Code

Open the program CE Control Program v2.R. This is the program that contains all of the setup code and function calls for the CE process. Where possible, default values for the macro variables have been set to provide guidance.

Below is the code and an explanation of what needs to be entered for each section.

require(XLConnect)

require(plyr)

require(subselect)

require(car)

These four lines of code load the additional packages that will be needed by the process. These should be run without any changes.

source("/mnt/projects/shared/pst\_qmgisi/Modeling/RCE/CE Functions v2.R", echo = F)

This line will load the RCE functions and should not be changed. Just run it.

setwd("/mnt/projects/locked/pst\_qmgisi/Modeling/CE Rebuild/binary/ce\_output")

This line will set the working directory. Enter the path to the folder where you want to write all output from the CE process. Make sure your variable lists are in this folder.

## Step 1: Sampling

lib <- "/mnt/projects/locked/pst\_qmgisi/Modeling/CE Rebuild/sasdata" # Path to input file

f.sample(dat = file.path(lib, "bin\_file.xdf"),

outds = "CE1\_Resampled.xdf",

dep\_var = "resp",

binary\_dv = "Y",

include\_if = expression(!is.na(resp)),

split\_if = .7,

ds\_present = "Y",

bootstrap = "N",

oversampled\_rr = .20,

min\_num\_resp = 1000,

seed = 123456)

The first step in the Sampling process is to set **lib** to the path to access your input dataset.

The parameters for the sampling function are:

* dat – this identifies your input dataset. It combines the path you set in variable lib with the name of the dataset you set on this line. The dataset must be an xdf file.
* outds – this is the name of the dataset that will be written out by the function. It must be an xdf file. The standard name is **CE1\_Resampled.xdf**.
* dep\_var – this is the name of your dependent variable. Make sure you use the correct case for the letters in the variable name as R is case sensitive.
* binary\_dv – this is a flag to tell the process if your dependent variable is binary or not. Acceptable values are Y for binary and N for anything else. If you have an ordinal dependent variable, you can treat it as continuous for this process (set to N) and it will work. If you have a nominal dependent variable you should be using the PST Nominal modeling process as this has code to handle this situation.
* include\_if – this is a condition that will cause records from your input file to be included (the opposite of the SAS exclude\_if). Make sure you enclose your condition in **expression( )**. Common conditions are **expression(!is.na(resp))** and **expression(sales>0)**(replacing resp and sales with your dependent variable name). If you do not have an exclusion clause, set to **NULL**.
* split\_if – this is your split condition to identify the learning portion of your file. If you already have a flag on your file to identify this, use the default **expression(lrnval=="L")** (substituting your variable name for lrnval). If you do not already have a flag on your file to identify this, you would set split\_if to the percent of the file that you want to have in your learning portion. The normal split is 70% learning (**split\_if = .7**).
* ds\_Present – this tells the process whether or not you want to apply the standard DataSource recodes. Set to Y to apply.
* bootstrap – this is how you can request sampling to achieve a desired proportion between 0’s and 1’s. This only applies if your dependent variable is binary. Set to Y to invoke.
* oversampled\_rr – if you are doing sampling, this is the desired ratio of 1’s to 0’s. This value should be between .1 and .5.
* min\_num\_resp – if you are doing sampling, this is the minimum number of 1’s to be selected
* seed – if you are doing sampling, this is the seed to use when doing random selects

This step will perform the following tasks:

1. If *include\_if* is not NULL, any records that do not meet the condition will be discarded.
2. If *bootstrap* is set to N then the portion of the file identified by *split\_if* will be split in half as model and validation portions and the remainder of the file will be the test portion. If *bootstrap* is set to Y and you have a binary dependent variable, then the portion of the file identified by *split\_if* will be sampled to meet the proportion *oversampled\_rr* with the minimum number of 1’s equal to *min\_num\_resp*. It will then be split in half as model and validation portions and the remainder of the file will be the test portion.
3. If *ds\_Present* is set to Y, the standard DataSource recodes will be applied
4. A new dataset, *outds*, will be written to the output library. It will have a variable called *mod\_val\_test* that identifies the portions of the file.

## Step 2: EDA and variable recoding

f.recode(dat = "CE1\_Resampled.xdf",

keep\_list=c("KL\_id","resp","wgt","mod\_val\_test"),

dep\_var = "resp",

binary\_dv = "Y",

prefix = "R1\_",

missrate = .75,

concrate = .9,

valcnt = 50,

minbinnc = 500,

minbinnp = .05,

talpha = .05,

nom\_method = "INDEX",

min\_size = 500,

cap\_flrO = "Y",

transformationO = "N",

impmethodO = "MEDIAN",

stdmethodO = "NO",

p\_lo = .01,

p\_hi = .99,

cap\_flrC = "Y",

transformationC = "Y",

impmethodC = "MEAN",

stdmethodC = "STD",

profiling = "Y",

equal\_dist = "N",

num\_category = 10)

These are the common parameters that are used for the second step of the CE process – EDA and recoding:

* dat – this identifies your input dataset. The default is to use the xdf file created in the sampling step.
* keep\_list – this is the list of variables you want to keep in addition to your independent variables. You should always include your id variable, dependent variable, weight variable (if any) and mod\_val\_test. You can add any additional variables you like. Any variables in your input dataset that are not listed in one of the independent variable lists or keep\_list will be dropped from the datasets created by the process.
* dep\_var – this is the name of your dependent variable. Make sure you use the correct case for the letters in the variable name as R is case sensitive.
* binary\_dv – this is a flag to tell the process if your dependent variable is binary or not. Acceptable values are Y for binary and N for anything else.
* prefix – this is the prefix that will be used when creating the recoded independent variables. As an example with prefix set to R1\_, raw variable *sales* will be recoded to *R1\_sales*.
* missrate – this is the maximum missing percent allowed for any variable. If the missing rate is greater than this number, the variable will be discarded.
* concrate – this is the maximum percent of the file that can be in a single value. If an independent variable has more than this, it will be discarded because it is too concentrated to provide stable and useful information. This is used for binary, nominal and ordinal variables

These are the parameters used to process nominal variables:

* valcnt – this is the maximum number of unique values a nominal variable is allowed to have. Variables with more than this number of unique values will be discarded. If you do not want to restrict the number of values, set valcnt to zero.
* minbinnc and minbinnp – these are the minimum count and minimum percent of file in a bin for it to be usable. Nominal variables are grouped into bins based on similar performance. Minbinnc and minbinnp are used to determine if there are enough records in the bin for the evaluation to be valid. Either one can be set to zero to only use the other. For example, with a file of 50,000 records, setting minbinnc to 500 and minbinnp to 0 will force each bin to have at least 500 records in it. Setting minbinnc to 500 and minbinnp to .05 will force each bin to have at least 2,500 records in it (50,000 x .05 = 2,500 is greater than 500). Setting minbinnc to 1,000 and minbinnp to .01 will for each bin to have at least 1,000 records in it. (50,000 x .01 = 500 is less than 1,000).
* talpha – this is the t-test significance level for collapsing bins.
* nom\_method – this is the method to use for the recoded variables. The options are:
  + BINARY – this creates a new binary variable for each bin. This is the method used in the previous version of CE.
  + INDEX – this creates a single new variable with the index of the bin average dependent value to the overall average used for the values.
  + MEAN – this creates a single new variable with the bin average dependent value as the values.

Example: a variable is collapsed into four bins with the following values for average dependent and index

|  |  |  |
| --- | --- | --- |
| Bin | Average | Index |
| Overall | .16667 |  |
| 1 | .10959 | 65.75 |
| 2 | .11815 | 70.89 |
| 3 | .17002 | 102.01 |
| 4 | .21542 | 129.25 |

If you use nom\_method = BINARY you will get three new binary variables – one for bin 1, one for bin 2, and one for bin 3. If you use nom\_method = INDEX you will get one new variable with four possible values – 65.75, 70.89, 102.01, 129.25. If you use nom\_method = MEAN you will get one new variable with four possible values - .10959, .11815, .17002, .21542.

These are the parameters used to process ordinal and continuous variables:

* min\_size – if you choose Equal Response for the method of missing value imputation, min\_size gives the minimum number of missing records to use this method. If you have chosen Equal Response and the missing count is less than this number for a variable, the median will be used instead.

These are the parameters used to process ordinal variables:

* cap\_flrO – this controls whether capping and flooring are done to handle outliers before standardization is done. Set to Y to do.
* transformationO – this controls whether transformations are done. The transformations include square, square root, natural log, exponential and inverse. Set to Y to include evaluation of the transforms as well as the natural values of the variables.
* impmethodO – this is the method to use for missing value imputation. The options are ER for Equal Response, MEAN, MEDIAN, ZERO, MINIMUM or MIDRANGE.
* stdmethodO – this is the standardization method to use. Options are MEAN, MEDIAN, SUM, EUCLEN, USTD, STD, RANGE, MIDRANGE, MAXABS, IQR, MAD or NO to not do standardization.

These are the parameters used to process continuous variables:

* p\_lo and p\_hi – these is the lower and upper percentiles to be used when checking a variable for constant value. The code will determine the value of the independent variable at the lower and upper percentiles. If the values are the same, the variable is considered constant and will be discarded.
* cap\_flrC – this controls whether capping and flooring are done to handle outliers before standardization is done. Set to Y to do.
* transformationC – this controls whether transformations are done. The transformations include square, square root, natural log, exponential and inverse. Set to Y to include evaluation of the transforms as well as the natural values of the variables.
* impmethodC – this is the method to use for missing value imputation. The options are ER for Equal Response, MEAN, MEDIAN, ZERO, MINIMUM or MIDRANGE.
* stdmethodC – this is the standardization method to use. Options are MEAN, MEDIAN, SUM, EUCLEN, USTD, STD, RANGE, MIDRANGE, MAXABS, IQR, MAD or NO to not do standardization.

These are the parameters used in profiling of the independent variables:

* profiling – this allows you to do profiling on all of the recoded variables. Profiling will automatically be done during step 5, final model selection, on the variables in the final model. If you want profiling on all variables, set this to Y.
* equal\_dist – if you choose to do profiling, you can choose equal distance for dividing variables into groups by setting this to Y. If set to N, then an equal number of records will be assigned to each group.
* num\_category – if you choose to do profiling, this is the maximum number of categories that will be used for ordinal and continuous variables.

This recoding step code goes through each variable type that is present and applies rules to recode the raw data into a more usable form. It does not use test portion of the file.

All variables:

* All variables are checked for missing rate. If the missing rate is greater than *missrate*, then the variable is discarded as unusable.

Binary variables:

* The code will recode all records with a value of Y, y or 1 to 1 and everything else to 0.
* The concentration rate is the percent of the file that is recoded to 1. If the concentration rate is greater than *concrate* or less than (1-*concrate*) then the variable is discarded.

Nominal variables:

* The concentration rate is the highest percent of the file that is in any single value. If the concentration rate is greater than *concrate* then the variable is discarded.
* If *valcnt* is greater than 0 and the number of unique values for a variable is greater than *valcnt* the variable is discarded.
* The file is summarized by value and sorted by mean dependent variable. The data is then collapsed to bins with at least the minimum number of records in each. The minimum number of records is calculated as the maximum of *minbinnc* and (*minbinnp* x the total number of observations).
* Bins are further collapsed based on a t-test of the difference in mean dependent variable between bins. The significance level for the t-test is *talpha*.

Ordinal variables:

* The concentration rate is the highest percent of the file that is in any single value. If the concentration rate is greater than *concrate* then the variable is discarded.
* The minimum value for the dependent variable on the non-missing portion of the file will be compared to the maximum. If they are the same, the variable will be discarded since there is no variance in the dependent variable.
* If *transformationO* is set to Y, transforms of the raw variable will be created.
* Missing values will be assigned a value based in *impmethodO*.
* If *stdmethodO* is not set to NO, then the data will be standardized using the method specified in *stdmethodO*.
* The best form of the variable will be selected using the highest absolute correlation to the dependent variable.

Continuous variables:

* The value for the variable at the *p\_lo* percentile will be compared to the value at the *p\_hi* percentile. If they are the same, then the variable is discarded as having a constant value.
* The minimum value for the dependent variable on the non-missing portion of the file will be compared to the maximum. If they are the same, the variable will be discarded since there is no variance in the dependent variable.
* If *transformationC* is set to Y, transforms of the raw variable will be created.
* Missing values will be assigned a value based in *impmethodC*.
* If *stdmethodC* is not set to NO, then the data will be standardized using the method specified in *stdmethodC*.
* The best form of the variable will be selected using the highest absolute correlation to the dependent variable.

A new dataset, **CE2\_Recoded**, will be written to the output library. It will contain the variables listed in *keep\_list* as well as all of the new recoded variables.

Once all of the recoding is done, the following reports will be written out: EDA, Correlation, and if profiling is set to Y, Profile. There is an Excel tool to read and format the profiling report. It is called CE\_Profiling.xlsm. Simply open the tool, enable macros, click the "Upload Data" button and select the file just created by the R process.

You will also have text files for the recodes for each variable type.

## Step 3: Variable reduction

f.var\_redu(dat = "CE2\_Recoded.xdf",

dep\_var = "resp",

binary\_dv = "Y",

samplesize = 50000,

redu\_weight = "N",

weight = "wgt",

sources = 3,

maxnum = 1000,

maxcorr = .7,

ind\_dv\_corr = "Y",

max\_dv\_corr = .7,

univ\_reg = "N",

maxpuni = .05,

correlation = "Y",

corrcut = .01,

factor = "Y",

nfact = 20,

minfact = .5,

regression = "Y",

alphareg = .05,

logistic = "Y",

alphalog = .05,

information = "Y",

decile = 20,

infvcut = .01)

These are the common parameters that are used for the third step of the CE process – variable reduction:

* dat – this identifies your input dataset. The default is to use the xdf file created in the recoding step.
* dep\_var – this is the name of your dependent variable. Make sure you use the correct case for the letters in the variable name as R is case sensitive.
* binary\_dv – this is a flag to tell the process if your dependent variable is binary or not. Acceptable values are Y for binary and N for anything else.
* samplesize – this is the maximum number of records that will be used in the variable reduction processes. If you have fewer records than this number then all records will be used. If you have more records than this number than a sample of this size will be taken. The larger this number is, the longer this process will take so it is recommended that you use at most 50,000. This is generally sufficient to evaluate the usefulness of your independent variables.
* redu\_weight – this allows you to use weights in this step. Set to Y to use weights. Set to N to not use weights.
* weight – this it the name of the variable to be used as weights if redu\_weight is set to Y.
* sources – this is the minimum number of sources to survive the variable reduction phase. There are multiple methods available for variable evaluation. Each has a criterion for passing that method. *Sources* is the number of passing grades a variable must have or it will be discarded. Make sure that this number is less than or equal to the number of methods selected.
* maxnum – this is the maximum number of variables to be kept. Only use this if you want to restrict the number of variables going into step 4. This is generally only needed if you are having problems with logistic model convergence.
* maxcorr – this is the maximum correlation allowed between the independent variables. If a correlation is too high the variable with the highest correlation to the dependent variable will be kept and the other discarded.

These are the parameters used for evaluating the maximum correlation to the dependent variable. If a correlation is too high the independent variable will be discarded. This will override the rest of the process and the additional tests will not be performed.

* ind\_dv\_corr – this tells the process whether you want to use this check. Set to Y to include; set to N to skip.
* max\_dv\_corr – this is the maximum correlation allowed to the dependent variable.

These are the parameters used for the univariate regression method of variable evaluation. This method evaluates each independent variable by itself for a relationship with the dependent variable.

* univ\_reg – this tells the process whether you want to use this method. Set to Y to include; set to N to skip.
* maxpuni – this is the maximum p value for passing this evaluation method.

These are the parameters needed for the correlation method of variable evaluation. This method evaluates each independent variable based on its correlation with the dependent variable.

* correlation – this tells the process whether you want to use this method. Set to Y to include; set to N to skip.
* corrcut – this is the minimum correlation for passing this evaluation method.

These are the parameters used for the factor method of variable evaluation. This method evaluates each independent variable based on its maximum loading with the factors created. If you get an error message that it is unable to optimize, you will need to skip this test.

* factor – this tells the process whether you want to use this method. Set to Y to include; set to N to skip.
* nfact – this is the number of factors that should be used.
* minfact – this is the minimum factor loading for passing this evaluation method.

These are the parameters used for the multivariate linear regression method of variable evaluation. This method builds a linear regression model using all independent variables together. If a variable enters the model, it passes the evaluation.

* regression – this tells the process whether you want to use this method. Set to Y to include; set to N to skip.
* alphareg – this is the maximum alpha value for a variable to enter the model using forward selection.

These are the parameters used for the multivariate logistic regression method of variable evaluation. This method is only available if you have a binary dependent variable. This method builds a logistic regression model using all independent variables together. If a variable enters the model, it passes the evaluation.

* logistic – this tells the process whether you want to use this method. Set to Y to include; set to N to skip.
* alphalog – this is the maximum alpha value for a variable to enter the model using forward selection.

These are the parameters used for the information value method of variable evaluation. This method evaluates each independent variable based on its information value.

* information – this tells the process whether you want to use this method. Set to Y to include; set to N to skip.
* decile – this is the number of groups the variables should be split into to calculate the information value of the variable.
* infvcut – this is the minimum information value for passing this evaluation method.

The purpose of this variable reduction code is to reduce the number of candidate variables based on various statistical tests before building a model. This is done to reduce the amount of resources and length of time needed to build a model. It does not use the test portion of the file.

The first step is to test for correlation between the independent variables. If any pair has an unsigned correlation greater than *maxcorr*, the variable with the lower correlation to the dependent variable will be removed. This step will always be done to ensure a well conditioned set of data.

There are seven evaluation methods that you turn on or off with a control macro variable:

* Correlation to the dependent variable. This is turned on by setting *ind\_dv\_corr* to Y. If any variable has an unsigned correlation greater than *max\_dv\_corr*, the variable will be removed. Any variables failing this test will not be evaluated by the remaining tests.
* Univariate regression – evaluates each variable in isolation by building a logistic (binary dependent) or linear regression model with a single independent variable. It is turned on by setting *univ\_reg* to Y. The p-value of the appropriate test statistic (binary is chi-square, otherwise f test) is captured. Variables with a p-value less than or equal to *maxpuni* pass this test.
* Correlation – evaluates each variable by calculating the correlation with the dependent variable. It is turned on by setting *correlation* to Y. Variables with an unsigned correlation greater than or equal to *corrcut* pass this test.
* Factor – evaluates variables using factor analysis. It is turned on by setting *factor* to Y. *nfact* factors are created from the correlation matrix. The maximum unsigned loading for each variable is then determined. All variables with a maximum unsigned loading greater than or equal to *minfact* pass this test.
* Linear regression – evaluates variables in combination by building a linear regression model using forward selection and significance level to enter of *alphareg*. It is turned on by setting *regression* to Y. In the first pass, the variables are broken into groups to make the analysis more manageable. A model is built for each group. All variables that enter a model are then combined and a final model is built. All variables that enter the second model pass this test.
* Logistic regression – evaluates variables in combination by building a logistic regression model using forward selection and significance level to enter of *alphalog*. It is only available if your dependent variable is binary. It is turned on by setting *logistic* to Y. In the first pass, the variables are broken into groups to make the analysis more manageable. A model is built for each group. All variables that enter a model are then combined and a final model is built. All variables that enter the second model pass this test.
* Information value – evaluates variables using its information value. It is turned on by setting *information* to Y. Each variable is summarized to bins using *decile* as the maximum number of bins to use. The variable information value is then calculated. Variables with a information value greater than or equal to *infvcut* pass this test.

A variable must pass *sources* number of tests to be kept.

The output of this step is a report of the statistics of all variables and a text file listing the reduced list of variables.

## Step 4: Model selection and tuning

source("CE3\_Varlist\_redu.txt") # Load variable list generated in step 3

f.model\_val(insdn = "CE2\_Recoded.xdf",

varlist = varlist\_redu,

dep\_var = "resp",

binary\_dv = "Y",

weight = NULL,

sel\_alpha = .05,

refit = F,

includelist = NA,

startlist = NA,

excludelist = NA,

criteria = "c",

threshold = 0,

SQL\_join = "union",

minimp = .01,

graph\_plot = "Y")

The fourth step of the CE process is model selection and tuning. The first step in this process is to load the list of variables selected in step 3 by sourcing the text file created.

The parameters used in this step are:

* insdn – this identifies your input dataset. The default is to use the xdf file created in the recoding step.
* varlist – this is the list of variables you want to consider. The default is to use the list of variables created in step 3.
* dep\_var – this is the name of your dependent variable. Make sure you use the correct case for the letters in the variable name as R is case sensitive.
* binary\_dv – this is a flag to tell the process if your dependent variable is binary or not. Acceptable values are Y for binary and N for anything else.
* weight – this is the name of the variable that has the weight for each observation. If you do not want to include weights in this step, set this parameter to **NULL**.
* sel\_alpha – this is the alpha level to use for addition and removal of variables from the model
* refit – this is a logical flag that tells the modeling process whether you want to refit at each step. This is only used with a binary dependent variable. It will add significantly to the run time and should only be used if you are having convergence issues and you have at most 100 independent variables.
* includelist – this is the list of variables that must be in the model. Set to **NA** if you do not want to use.
* startlist – this is the list of variables that must be in the first model. With backwards selection, these can be removed so are not guaranteed to be in the final model. Set to **NA** if you do not want to use.
* excludelist – this is a list of variables that should not be included in the model. Set to **NA** if you do not want to use.
* criteria – this is the metric to use during variable tuning. The options for a binary dependent variables are: AIC, SC, LogL2, Rsquare, SomersD, Gamma, TauA, c, Concord, Discon, LackFit, Lift\_Index, infv, ks. The options for a continuous dependent variables are: AIC, SBC, JP, Rsquare, AdjRsq, RMSE, CoeffVar, PC, Lift\_Index, gini, infv, ks. The default for binary is c and for continuous is AdjRsq. Variable tuning looks at this measure to determine if a variable should be discarded because it appears to be causing over fitting.
* threshold – this is the minimum change in the evaluation metric to include a variable
* SQL\_join – this is the type of join to use between file portions in variable tuning. Variable tuning will select a set of variables based on each portion of the learning file independently. It will then join the lists using this variable. If you use union (recommended), you will get all variables selected in either portion. If you use intersect, you will only get the variables that are in both portions.
* minimp – this is the minimum relative importance to keep a variable.
* graph\_plot – this controls whether the process creates an Excel report on the variables selected by this step. Set to Y to create.

The model selection and tuning runs the process on the model portion of the file, then the validation portion of the file, then compares the variables selected by each.

1. Create the list of independent variables to use by using *varlist\_redu* (from step 3), *include\_list*, *start\_list*, and *exclude\_list*.
2. Create a regression model (logistic if a binary dependent, otherwise linear) using forward selection.
3. Create a regression model (logistic if a binary dependent, otherwise linear) using backward selection.
4. Combine all variables selected by either model into the candidate variable list.
5. Calculate statistics for the full model using all candidate variables.
6. Calculate statistics for a series of models using all but one of the candidate variables in each model. As an example, if you have 10 candidate models the first model would use variables 2 through 10, the second variables 1 and 3 through 10, etc.
7. Subtract the statistics for the subset models from the statistics for the full model. This will show what the impact of each individual variable is.
8. Accumulate all of the statistics into a dataset.

Once this process has been run on both portions of the file, the next step is variable tuning. This will select variables from each portion based on *criteria*, *threshold* and *minimp*. If you specify an invalid criteria the code will change it to the default for your dependent variable type – c for binary, AdjRsq for linear. The two lists will be joined based on *SQL\_join*.

The output of this step is a report of the statistics for the models built, a text file listing the selected variables and, if *graph\_plot* is set to Y, an Excel report showing basic information about the selected variables.

## Step 5: Final Model build and validation on test sample

source("CE4\_Varlist\_Final.txt") # Load variable list generated in step 4

f.model\_lift(insdn = "CE2\_Recoded.xdf",

varlist = varlist\_final,

dep\_var = "resp",

binary\_dv = "Y",

weight = NULL,

fin\_alpha = .05,

refit = F,

method = "stepwise",

fin\_num\_category = 10,

fin\_equal\_dist = "N")

The fifth step of the CE process is final model selection. The first step in this process is to load the list of variables selected in step 4 by sourcing the text file created.

The parameters used in this step are:

* insdn – this identifies your input dataset. The default is to use the xdf file created in the recoding step.
* varlist – this is the list of variables you want to consider. The default is to use the list of variables created in step 4.
* dep\_var – this is the name of your dependent variable. Make sure you use the correct case for the letters in the variable name as R is case sensitive.
* binary\_dv – this is a flag to tell the process if your dependent variable is binary or not. Acceptable values are Y for binary and N for anything else.
* weight – this is the name of the variable that has the weight for each observation. If you do not want to include weights in this step, set this parameter to **NULL**.
* fin\_alpha – this is the alpha level to use for addition and removal of variables from the model
* refit – this is a logical flag that tells the modeling process whether you want to refit at each step. This is only used with a binary dependent variable. It will add significantly to the run time and should only be used if you are having convergence issues and you have at most 100 independent variables.
* method – this is the selection method to use when building the final model. Options are forward, backward and stepwise.
* fin\_num\_category – this is the maximum number of categories that will be used in profiling of the variables in the final model.
* fin\_equal\_dist – this controls whether equal distance will be used for dividing variables into groups. Set to Y to use. If set to N, then an equal number of records will be assigned to each group.

The steps in final model selection are:

1. A model will be built using everything but the test portion of the file – logistic if a binary dependent, otherwise linear. The selection method is controlled by *method*.
2. Statistics will be calculated in the same way as Step 4.
3. Gains table are created.
4. Final variables are profiled.

The output of this step is a report of the model built, a text file listing the final variables selected and a text file with the scoring equation for the final model.