Part 3 - Feature Selection Methods

jean wills

07/07/2020

# Step 3 - Feature Selection Methods

Note that these models are NOT being tested for metrics, but rather as a ’one-off" to see if we can reduce the number of attributes

### 1: PCA

### 2. Recursive feature elimination (RFE) with Random Forest

### 3. Boruta

### 4. neural network LVQ method

### 5. Regression - ANOVA, AIC

### 6. Regression - AIC stepwise

## this file uses “BM\_mini\_num\_sc”

## Get the data in the proper format first:

### read in dataset - on “BM\_mini”

library(plyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

BM\_mini <- read.csv("/Users/jeanwills/Desktop/CKME136/1\_data/bank.csv", header=T, sep = ";", stringsAsFactors = T, na.strings = "NA")

### NUMERIC DATA Cleaning - change numeric data to 95% CI

BM<-BM\_mini  
  
Lq\_bal<- quantile(BM$balance, probs=c(0.025))  
Hq\_bal<- quantile(BM$balance, probs=c(0.975))  
Lq\_dur<- quantile(BM$duration, probs=c(0.025))  
Hq\_dur<- quantile(BM$duration, probs=c(0.975))  
Lq\_cam<- quantile(BM$campaign, probs=c(0.025))  
Hq\_cam<- quantile(BM$campaign, probs=c(0.975))  
Lq\_days<- quantile(BM$pdays, probs=c(0.025))  
Hq\_days<- quantile(BM$pdays, probs=c(0.975))  
Lq\_prv<- quantile(BM$previous, probs=c(0.025))  
Hq\_prv<- quantile(BM$previous, probs=c(0.975))  
BM$balance[BM$balance < Lq\_bal] <- Lq\_bal  
BM$balance[BM$balance > Hq\_bal] <- Hq\_bal  
BM$duration[BM$duration < Lq\_dur] <- Lq\_dur  
BM$duration[BM$duration > Hq\_dur] <- Hq\_dur  
BM$campaign[BM$campaign < Lq\_cam] <- Lq\_cam  
BM$campaign[BM$campaign > Hq\_cam] <- Hq\_cam  
BM$pdays[BM$pdays < Lq\_days] <- Lq\_days  
BM$pdays[BM$pdays > Hq\_days] <- Hq\_days  
BM$previous[BM$previous < Lq\_prv] <- Lq\_prv  
BM$previous[BM$previous > Hq\_prv] <- Hq\_prv  
  
# now have file BM ...

### now make minor adjsutments

# switch -1 -> 0 in 'pdays'  
BM$pdays<- ifelse(BM$pdays == -1, 0, BM$pdays)  
# switch duration in seconds to minutes for easier use  
BM$duration<- BM$duration/60

### now switch categorical to get numeric - RESULT is “BM\_mini\_num”

# numeric still NOT normalized/scaled,   
# actually ALL DATA not normalized/scaled EXCEPT default, housing, loan  
BM\_n <- BM  
BM\_n$job<- as.numeric(BM\_n$job) #12  
# marital: 1-single, 2-married, 3-divorced  
BM\_n$marital<- ifelse(BM\_n$marital == c("single"), 1,   
 ifelse(BM\_n$marital== c("married"), 2, 3))  
# education: 0:unknown, 1: primary, 2:secondary, 3:divorced  
BM\_n$education<- ifelse(BM\_n$education == c("unknown"), 0,   
 ifelse(BM\_n$education == c("primary"), 1,   
 ifelse(BM\_n$education == c("secondary"), 2, 3)))  
# default, housing, loan: if yes then 0 else 1  
# BM\_num$housing<- as.numeric(BM\_num$housing) #2  
BM\_n$default<- ifelse(BM\_n$default == c("yes"), 0, 1) #2  
BM\_n$housing<- ifelse(BM\_n$housing == c("yes"), 0, 1) #2  
BM\_n$loan<- ifelse(BM\_n$loan == c("yes"), 0, 1) #2  
BM\_n$contact<- as.numeric(BM\_n$contact) #3  
# month: jan:1, feb:2.....dec:12  
BM\_n$month<- ifelse(BM\_n$month == "jan", 1,   
 ifelse(BM\_n$month == "feb", 2,   
 ifelse(BM\_n$month == "mar", 3,  
 ifelse(BM\_n$month == "apr", 4,   
 ifelse(BM\_n$month == "may", 5,   
 ifelse(BM\_n$month == "jun", 6,  
 ifelse(BM\_n$month == "jul", 7,  
 ifelse(BM\_n$month == "aug", 8,  
 ifelse(BM\_n$month == "sep", 9,  
 ifelse(BM\_n$month == "oct", 10,  
 ifelse(BM\_n$month == "nov", 11, 12)))))))))))  
# poutcome: 0:unknown,other, 1:failure, 2: success  
BM\_n$poutcome<- ifelse(BM\_n$poutcome == c("failure"), 1, ifelse(BM\_n$poutcome== c("success"), 2, 0))   
# result: BM\_num with only numeric data (NOT scaled)  
# extra step for BM\_mini  
# BM\_n

# 

### scale only original numeric data

# KEEP: age-1, balance-6, day-10, duration-12, campaign-13, pdays-14, previous-15  
BMS<- BM\_n  
# BMS<-BM\_mini  
normalize<- function(x) {  
 return ((x - min(x)) / (max(x) - min(x)))  
}  
BM\_s<- as.data.frame(lapply(BMS[,c(1,6,10,12:15)], normalize))  
# now recombine dataframes with the nominal components  
BM\_s$job<-BMS$job  
BM\_s$marital<-BMS$marital  
BM\_s$education<-BMS$education  
BM\_s$default<-BMS$default  
BM\_s$housing<-BMS$housing  
BM\_s$loan<-BMS$loan  
BM\_s$contact<-BMS$contact  
BM\_s$month<-BMS$month  
BM\_s$poutcome<-BMS$poutcome  
BM\_s$y<-BMS$y  
# convert  
BM\_mini\_sc<-BM\_s  
# result used BM\_num file but now BM\_num\_scale with normalized numeric data  
# and y is factor  
# to convert y to numeric use next line  
# BM\_scale$y<- ifelse(BM\_scale$y==c("yes"), 1, 0)  
rm(BMS)  
rm(BM\_s)  
rm(BM)  
rm(BM\_n)

## we are using SMOTE since it was found to be better so data will be rebalanced with SMOTE first

# step 2 - run training and test datasets FOR ALL CONFIGURATIONS

# str(BM\_mini\_sc)  
# rename datasets  
BM<-BM\_mini\_sc  
set.seed(30)  
# get train and test datasets  
# note that if we use cross validation, we can use the complete dataset for training or keep a portion for validation after  
#  
BM\_train\_index <- sample(nrow(BM), 0.7 \* nrow(BM))  
BM\_train<- BM[BM\_train\_index, ]  
BM\_test <- BM[-BM\_train\_index, ]  
BM\_train\_labels <- BM[BM\_train\_index, 17]  
BM\_test\_labels <- BM[-BM\_train\_index, 17]  
# note that we only balance the training sets  
# and leave test set as is.

# step 3a - run SPECIFIC Balancing step to get balanced data version:

## SMOTE SAMPLE MAJOR/MINOR - each for training

# install.packages("DMwR")  
library(DMwR)

## Loading required package: lattice

## Loading required package: grid

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

##   
## Attaching package: 'DMwR'

## The following object is masked from 'package:plyr':  
##   
## join

library(grid)  
# this one is slow  
set.seed(50)  
smote\_train <- SMOTE(y ~ ., data = BM\_train)   
table(smote\_train$y)

##   
## no yes   
## 1440 1080

# now we name datasets that are used in models below  
train=smote\_train  
  
BMX<- smote\_train  
data<- smote\_train  
RFEdata<-smote\_train  
BMX<-smote\_train  
dataset<- smote\_train  
# train<- smote\_train  
pca\_trainset = smote\_train   
BMY<- smote\_train

## #1 - PCA

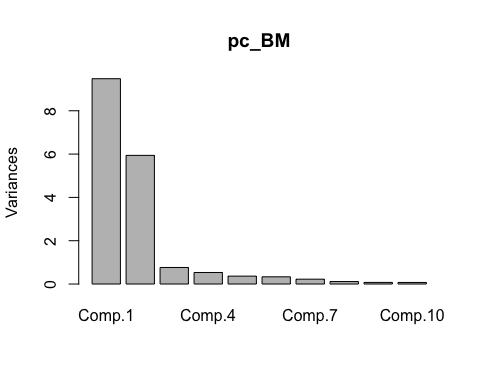
Principal component analysis (PCA) is a technique for reducing the dimensionality of datasets. It does so by creating new uncorrelated variables that maximize variance. There is no linearity or normality assumed in PCA. It seems that if the original correlation coeff’s between data < 0.3 then PCA won’t work well.

# data s/b numeric  
# do PCA on the independent variables only - take out $y  
BMX<- BMX[,-17]  
pc\_BM<- princomp(BMX)  
# pc\_BM<- princomp(BMX)  
# pc\_BM$scores  
summary(pc\_BM)

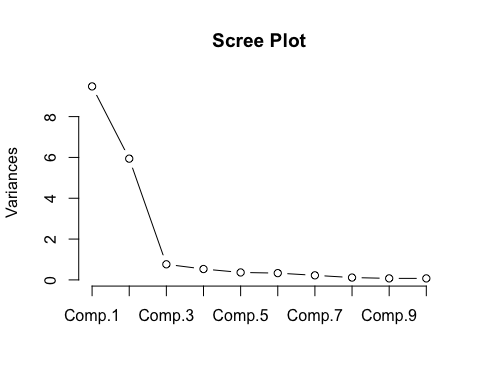
## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 3.0789106 2.4372571 0.87421278 0.72993324 0.60390389  
## Proportion of Variance 0.5242719 0.3285225 0.04226655 0.02946651 0.02016964  
## Cumulative Proportion 0.5242719 0.8527944 0.89506099 0.92452750 0.94469714  
## Comp.6 Comp.7 Comp.8 Comp.9 Comp.10  
## Standard deviation 0.57653209 0.4731591 0.338445204 0.272901727 0.270761617  
## Proportion of Variance 0.01838271 0.0123816 0.006334892 0.004118841 0.004054494  
## Cumulative Proportion 0.96307984 0.9754614 0.981796339 0.985915180 0.989969674  
## Comp.11 Comp.12 Comp.13 Comp.14  
## Standard deviation 0.223104478 0.202707055 0.196907992 0.140841635  
## Proportion of Variance 0.002752828 0.002272481 0.002144318 0.001097045  
## Cumulative Proportion 0.992722502 0.994994982 0.997139300 0.998236345  
## Comp.15 Comp.16  
## Standard deviation 0.1335479671 0.1185524809  
## Proportion of Variance 0.0009863636 0.0007772912  
## Cumulative Proportion 0.9992227088 1.0000000000

## visually see the data…

# can plot  
plot(pc\_BM)



# scree plot  
screeplot(pc\_BM, type="line", main="Scree Plot")



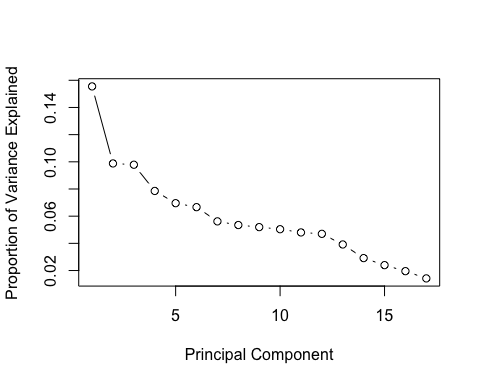
# see actual component values  
pc\_BM$loadings

##   
## Loadings:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9  
## age   
## balance   
## day -0.940  
## duration 0.184  
## campaign -0.246  
## pdays -0.119 0.103 0.282   
## previous -0.116 0.260   
## job -0.996   
## marital 0.143 0.216 0.202 0.937   
## education -0.466 -0.835 0.138 0.240   
## default   
## housing -0.113 0.129 -0.173 0.954   
## loan 0.986   
## contact 0.799 -0.397 0.375 -0.111 0.212   
## month 0.994   
## poutcome -0.282 0.241 0.784 -0.182 0.135   
## Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16  
## age 0.158 0.738 0.625   
## balance 0.242 0.913 0.230 -0.146 -0.123   
## day 0.184 0.129 -0.228   
## duration 0.975   
## campaign -0.232 0.935   
## pdays 0.611 -0.149 0.453 -0.525   
## previous 0.605 -0.139 -0.460 0.547   
## job   
## marital   
## education   
## default -0.993   
## housing 0.110   
## loan   
## contact   
## month   
## poutcome -0.422   
##   
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9  
## SS loadings 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000  
## Proportion Var 0.063 0.062 0.062 0.062 0.063 0.063 0.063 0.063 0.063  
## Cumulative Var 0.063 0.125 0.187 0.250 0.312 0.375 0.438 0.500 0.563  
## Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16  
## SS loadings 1.000 1.000 1.000 1.000 1.000 1.000 1.000  
## Proportion Var 0.062 0.062 0.062 0.062 0.063 0.063 0.062  
## Cumulative Var 0.625 0.688 0.750 0.813 0.875 0.938 1.000

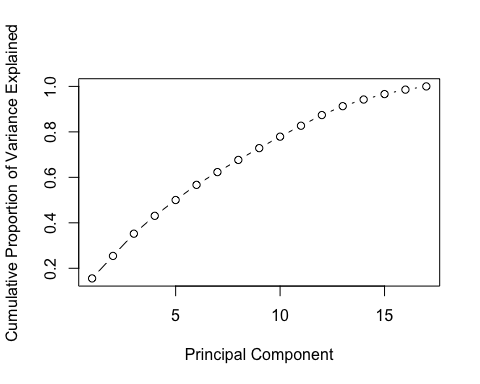
### PCA version 2 - This version uses “prcomp”

The code below shows that we still require ~10 princ.comp.’s to explain about 80% of the results. And each component is not responsible for a high percentageof the output. SO this PCA version not very useful on this dataset.

# PCA version 2  
# this example does not use test set....  
# convert y to numeric()  
# to convert y to numeric use next line  
pca\_trainset$y<- ifelse(pca\_trainset$y==c("yes"), 1, 0)  
pca = prcomp( pca\_trainset, scale = T )  
# variance  
pr\_var = ( pca$sdev )^2   
  
# % of variance  
prop\_varex = pr\_var / sum( pr\_var )  
  
# Plot  
plot( prop\_varex, xlab = "Principal Component",   
 ylab = "Proportion of Variance Explained", type = "b" )



# different scree plot  
# Scree Plot  
plot( cumsum( prop\_varex ), xlab = "Principal Component",   
 ylab = "Cumulative Proportion of Variance Explained", type = "b" )



### #3 Boruta

# y is factor with 2 variables yes or no  
library(Boruta)  
library(mlbench)  
library(caret)

## Loading required package: ggplot2

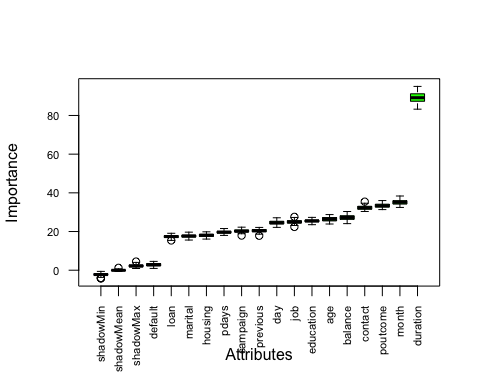
library(ranger)  
# build classification model  
set.seed(111)  
boruta<- Boruta(y ~ ., data = BMY, pValue = 0.01, mcAdj = TRUE, maxRuns = 200, doTrace = 0, holdHistory = TRUE, getImp = getImpRfZ)  
print(boruta)

## Boruta performed 116 iterations in 3.967153 mins.  
## 16 attributes confirmed important: age, balance, campaign, contact,  
## day and 11 more;  
## No attributes deemed unimportant.

# x=predictors, y is response vector, pValue is Conf.level, mcAdj=TRUE uses Bonferroni method,   
# maxRuns can be increased to resolve attributes left 'Tentative'.  
# getImp=getImpRfZ, which runs random forest from the ranger package and gathers Z-scores of mean decrease accuracy measure.  
# And is the function used to obtain attribute importance.

# Plot boruta

plot(boruta, las=2, cex.axis = 0.7)



# green are important attributes, red are not, and yellow are tentative  
# example: duration is very green, and job is red

## #4 LVQ method

# NEW second set: size=50, k=10  
# y stays as a category  
library(parallel)  
library(doMC)

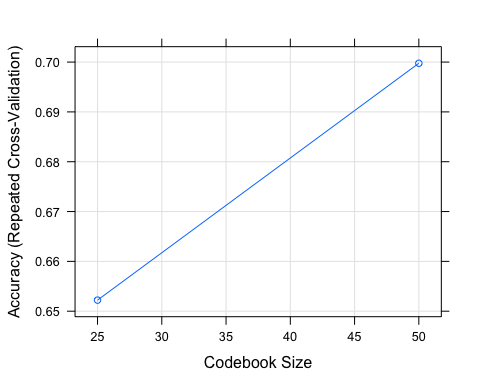
## Loading required package: foreach

## Loading required package: iterators

library(mlbench)  
library(caret)  
library(class)  
set.seed(7)  
# prepare training scheme  
control <- trainControl(method="repeatedcv", number=10, repeats=10)  
# manual example:   
# grid<- expand.grid(size=c(5,10,20,50), k=c(1,2,4,5,10))   
grid<- expand.grid(size=c(25,50), k=10)   
model <- train(y~., data=dataset, method="lvq", tuneGrid=grid, trControl=control)  
print(model)

## Learning Vector Quantization   
##   
## 2520 samples  
## 16 predictor  
## 2 classes: 'no', 'yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 10 times)   
## Summary of sample sizes: 2268, 2268, 2268, 2268, 2268, 2268, ...   
## Resampling results across tuning parameters:  
##   
## size Accuracy Kappa   
## 25 0.6522222 0.2837571  
## 50 0.6997619 0.3782678  
##   
## Tuning parameter 'k' was held constant at a value of 10  
## Accuracy was used to select the optimal model using the largest value.  
## The final values used for the model were size = 50 and k = 10.

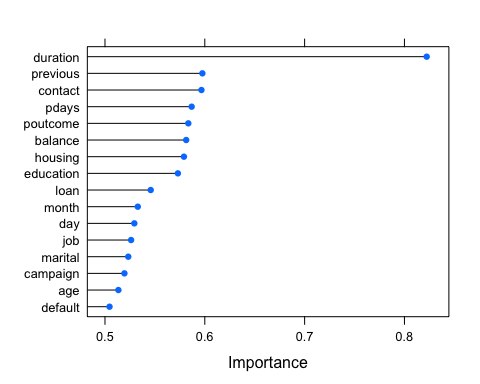
plot(model)



# estimate variable importance  
importance <- varImp(model, scale=FALSE)  
print(importance)

## ROC curve variable importance  
##   
## Importance  
## duration 0.8223  
## previous 0.5976  
## contact 0.5966  
## pdays 0.5868  
## poutcome 0.5835  
## balance 0.5814  
## housing 0.5791  
## education 0.5730  
## loan 0.5458  
## month 0.5329  
## day 0.5294  
## job 0.5262  
## marital 0.5233  
## campaign 0.5195  
## age 0.5134  
## default 0.5046

plot(importance)



## model 5 - MULTIPLE REGRESSION - running only once and not using test dataset

# let's build a set of models of various combo's of attributes and use ANOVA or AIC to choose best model  
# now convert y to 0 or 1   
train$y<- ifelse(train$y==c("yes"), 1, 0)  
  
m1<- lm(y~ contact, data=train)  
m2<- lm(y~ contact + poutcome, data=train)  
m3<- lm(y~ duration + poutcome, data=train)  
m4<- lm(y~ poutcome + month, data=train)  
m5<- lm(y~ poutcome, data=train)  
  
m6<- lm(y~ poutcome + month + job, data=train)  
m7<- lm(y~ month, data=train)  
m8<- lm(y~ education, data=train)  
m9<- lm(y~ job, data=train)  
m10<- lm(y~ poutcome + month + contact + default + education + job + duration, data=train)  
  
m11<- lm(y~ month + contact + default + education + job + duration, data=train)  
m12<- lm(y~ poutcome + month + contact + default + job + duration, data=train)  
m13<- lm(y~ poutcome + month + contact + job + duration, data=train)  
m14<- lm(y~ poutcome + month + contact + duration, data=train)  
m15<- lm(y~ poutcome + month + job + duration, data=train)  
  
m16<- lm(y~ poutcome + month + default + duration, data=train)  
m17<- lm(y~ poutcome + month + duration, data=train)  
m18<- lm(y~ poutcome + job + duration, data=train)  
m19<- lm(y~ poutcome + contact + duration, data=train)  
m20<- lm(y~ poutcome + month + contact + default + education + job, data=train)  
  
m21<- lm(y~ poutcome + month + contact + default + education, data=train)  
m22<- lm(y~ poutcome + month + contact + default, data=train)  
m23<- lm(y~ poutcome + month + contact, data=train)  
m24<- lm(y~ duration, data=train)  
m25<- lm(y~ contact + default + education, data=train)  
  
m26<- lm(y~ contact + default, data=train)  
m27<- lm(y~ poutcome + month + contact, data=train)  
m28<- lm(y~ poutcome + contact + job, data=train)  
#  
# run anova on all the models and compare RSS and pick lowest (of those that are significant)  
anova(m1,m2,m3,m4,m5,m6,m7,m8,m9,m10,m11,m12,m13,m14,m15,m16,m17,m18,m19,m20,m21,m22,m23,m24,m25,m26,m27,m28)

## Analysis of Variance Table  
##   
## Model 1: y ~ contact  
## Model 2: y ~ contact + poutcome  
## Model 3: y ~ duration + poutcome  
## Model 4: y ~ poutcome + month  
## Model 5: y ~ poutcome  
## Model 6: y ~ poutcome + month + job  
## Model 7: y ~ month  
## Model 8: y ~ education  
## Model 9: y ~ job  
## Model 10: y ~ poutcome + month + contact + default + education + job +   
## duration  
## Model 11: y ~ month + contact + default + education + job + duration  
## Model 12: y ~ poutcome + month + contact + default + job + duration  
## Model 13: y ~ poutcome + month + contact + job + duration  
## Model 14: y ~ poutcome + month + contact + duration  
## Model 15: y ~ poutcome + month + job + duration  
## Model 16: y ~ poutcome + month + default + duration  
## Model 17: y ~ poutcome + month + duration  
## Model 18: y ~ poutcome + job + duration  
## Model 19: y ~ poutcome + contact + duration  
## Model 20: y ~ poutcome + month + contact + default + education + job  
## Model 21: y ~ poutcome + month + contact + default + education  
## Model 22: y ~ poutcome + month + contact + default  
## Model 23: y ~ poutcome + month + contact  
## Model 24: y ~ duration  
## Model 25: y ~ contact + default + education  
## Model 26: y ~ contact + default  
## Model 27: y ~ poutcome + month + contact  
## Model 28: y ~ poutcome + contact + job  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 2518 583.24   
## 2 2517 560.58 1 22.657 146.0043 < 2.2e-16 \*\*\*  
## 3 2517 410.68 0 149.901   
## 4 2517 580.30 0 -169.627   
## 5 2518 581.93 -1 -1.623 10.4558 0.0012385 \*\*   
## 6 2516 579.51 2 2.413 7.7746 0.0004305 \*\*\*  
## 7 2518 615.33 -2 -35.812 115.3890 < 2.2e-16 \*\*\*  
## 8 2518 606.86 0 8.462   
## 9 2518 616.22 0 -9.359   
## 10 2512 389.82 6 226.408 243.1657 < 2.2e-16 \*\*\*  
## 11 2513 416.78 -1 -26.962 173.7442 < 2.2e-16 \*\*\*  
## 12 2513 392.84 0 23.932   
## 13 2514 393.00 -1 -0.153 0.9884 0.3202328   
## 14 2515 393.77 -1 -0.776 4.9982 0.0254606 \*   
## 15 2515 408.95 0 -15.175   
## 16 2515 409.88 0 -0.927   
## 17 2516 410.13 -1 -0.255 1.6418 0.2001963   
## 18 2516 409.44 0 0.692   
## 19 2516 393.79 0 15.650   
## 20 2513 554.68 3 -160.890   
## 21 2514 554.86 -1 -0.183 1.1767 0.2781399   
## 22 2515 559.72 -1 -4.861 31.3220 2.423e-08 \*\*\*  
## 23 2516 560.25 -1 -0.527 3.3985 0.0653726 .   
## 24 2518 450.40 -2 109.853   
## 25 2516 577.33 2 -126.939   
## 26 2517 582.37 -1 -5.036 32.4520 1.364e-08 \*\*\*  
## 27 2516 560.25 1 22.122 142.5544 < 2.2e-16 \*\*\*  
## 28 2516 560.12 0 0.129   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Run AIC model version

AIC(m1,m2,m3,m4,m5,m6,m7,m8,m9,m10,m11,m12,m13,m14,m15,m16,m17,m18,m19,m20,m21,m22,m23,m24,m25,m26,m27,m28)

## df AIC  
## m1 3 3469.627  
## m2 4 3371.780  
## m3 4 2587.654  
## m4 4 3458.928  
## m5 3 3463.964  
## m6 5 3457.493  
## m7 3 3604.600  
## m8 3 3569.706  
## m9 3 3608.271  
## m10 9 2466.269  
## m11 8 2632.803  
## m12 8 2483.777  
## m13 7 2482.761  
## m14 6 2485.729  
## m15 6 2581.017  
## m16 6 2586.723  
## m17 5 2586.289  
## m18 5 2582.033  
## m19 5 2483.821  
## m20 8 3353.109  
## m21 7 3351.939  
## m22 6 3371.918  
## m23 5 3372.291  
## m24 3 2818.290  
## m25 5 3447.995  
## m26 4 3467.881  
## m27 5 3372.291  
## m28 5 3371.711

## Run stepwise model

nullModel<-lm(y~1,data=train) # start with 0  
fullModel<-lm(y~., data=train) # try ALL  
houseStep<- step(nullModel, scope=list(lower=nullModel, upper=fullModel), direction = "both")

## Start: AIC=-3543.42  
## y ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + duration 1 166.748 450.40 -4335.2  
## + poutcome 1 35.216 581.93 -3689.5  
## + contact 1 33.907 583.24 -3683.8  
## + previous 1 23.897 593.25 -3640.9  
## + housing 1 15.203 601.94 -3604.3  
## + loan 1 12.615 604.53 -3593.5  
## + pdays 1 10.528 606.61 -3584.8  
## + education 1 10.278 606.86 -3583.7  
## + campaign 1 5.737 611.41 -3565.0  
## + month 1 1.816 615.33 -3548.9  
## + day 1 1.560 615.58 -3547.8  
## + default 1 1.388 615.75 -3547.1  
## + marital 1 1.196 615.95 -3546.3  
## + job 1 0.919 616.22 -3545.2  
## + balance 1 0.669 616.47 -3544.2  
## + age 1 0.607 616.54 -3543.9  
## <none> 617.14 -3543.4  
##   
## Step: AIC=-4335.16  
## y ~ duration  
##   
## Df Sum of Sq RSS AIC  
## + poutcome 1 39.717 410.68 -4565.8  
## + contact 1 29.297 421.10 -4502.7  
## + previous 1 24.566 425.83 -4474.5  
## + housing 1 16.942 433.45 -4429.8  
## + pdays 1 11.005 439.39 -4395.5  
## + loan 1 9.176 441.22 -4385.0  
## + education 1 7.640 442.76 -4376.3  
## + campaign 1 6.193 444.20 -4368.1  
## + day 1 1.655 448.74 -4342.4  
## + balance 1 1.348 449.05 -4340.7  
## + job 1 1.293 449.10 -4340.4  
## + default 1 0.700 449.70 -4337.1  
## + month 1 0.682 449.71 -4337.0  
## + marital 1 0.493 449.90 -4335.9  
## <none> 450.40 -4335.2  
## + age 1 0.268 450.13 -4334.7  
## - duration 1 166.748 617.14 -3543.4  
##   
## Step: AIC=-4565.8  
## y ~ duration + poutcome  
##   
## Df Sum of Sq RSS AIC  
## + contact 1 16.890 393.79 -4669.6  
## + housing 1 14.059 396.62 -4651.6  
## + loan 1 7.310 403.37 -4609.1  
## + education 1 6.087 404.59 -4601.4  
## + campaign 1 2.554 408.12 -4579.5  
## + previous 1 1.257 409.42 -4571.5  
## + job 1 1.240 409.44 -4571.4  
## + pdays 1 1.099 409.58 -4570.6  
## + balance 1 0.910 409.77 -4569.4  
## + month 1 0.548 410.13 -4567.2  
## + day 1 0.447 410.23 -4566.5  
## + marital 1 0.436 410.24 -4566.5  
## <none> 410.68 -4565.8  
## + default 1 0.246 410.43 -4565.3  
## + age 1 0.103 410.57 -4564.4  
## - poutcome 1 39.717 450.40 -4335.2  
## - duration 1 171.249 581.93 -3689.5  
##   
## Step: AIC=-4669.63  
## y ~ duration + poutcome + contact  
##   
## Df Sum of Sq RSS AIC  
## + housing 1 8.502 385.29 -4722.6  
## + loan 1 7.000 386.79 -4712.8  
## + education 1 3.400 390.39 -4689.5  
## + campaign 1 3.346 390.44 -4689.1  
## + pdays 1 2.379 391.41 -4682.9  
## + balance 1 0.883 392.90 -4673.3  
## + job 1 0.781 393.01 -4672.6  
## + day 1 0.608 393.18 -4671.5  
## + previous 1 0.399 393.39 -4670.2  
## <none> 393.79 -4669.6  
## + age 1 0.296 393.49 -4669.5  
## + marital 1 0.165 393.62 -4668.7  
## + default 1 0.137 393.65 -4668.5  
## + month 1 0.014 393.77 -4667.7  
## - contact 1 16.890 410.68 -4565.8  
## - poutcome 1 27.310 421.10 -4502.7  
## - duration 1 166.791 560.58 -3781.7  
##   
## Step: AIC=-4722.63  
## y ~ duration + poutcome + contact + housing  
##   
## Df Sum of Sq RSS AIC  
## + loan 1 6.631 378.65 -4764.4  
## + education 1 3.780 381.51 -4745.5  
## + campaign 1 3.239 382.05 -4741.9  
## + previous 1 0.859 384.43 -4726.3  
## + day 1 0.685 384.60 -4725.1  
## + balance 1 0.635 384.65 -4724.8  
## + pdays 1 0.608 384.68 -4724.6  
## <none> 385.29 -4722.6  
## + job 1 0.239 385.05 -4722.2  
## + marital 1 0.168 385.12 -4721.7  
## + month 1 0.126 385.16 -4721.5  
## + default 1 0.125 385.16 -4721.4  
## + age 1 0.017 385.27 -4720.7  
## - housing 1 8.502 393.79 -4669.6  
## - contact 1 11.333 396.62 -4651.6  
## - poutcome 1 27.039 412.32 -4553.7  
## - duration 1 168.529 553.82 -3810.3  
##   
## Step: AIC=-4764.38  
## y ~ duration + poutcome + contact + housing + loan  
##   
## Df Sum of Sq RSS AIC  
## + education 1 3.824 374.83 -4788.0  
## + campaign 1 3.159 375.50 -4783.5  
## + previous 1 0.980 377.68 -4768.9  
## + day 1 0.569 378.09 -4766.2  
## + pdays 1 0.440 378.21 -4765.3  
## + balance 1 0.332 378.32 -4764.6  
## <none> 378.65 -4764.4  
## + job 1 0.144 378.51 -4763.3  
## + marital 1 0.100 378.55 -4763.0  
## + month 1 0.061 378.59 -4762.8  
## + age 1 0.034 378.62 -4762.6  
## + default 1 0.017 378.64 -4762.5  
## - loan 1 6.631 385.29 -4722.6  
## - housing 1 8.133 386.79 -4712.8  
## - contact 1 11.187 389.84 -4693.0  
## - poutcome 1 25.720 404.37 -4600.8  
## - duration 1 165.423 544.08 -3853.0  
##   
## Step: AIC=-4787.96  
## y ~ duration + poutcome + contact + housing + loan + education  
##   
## Df Sum of Sq RSS AIC  
## + campaign 1 3.258 371.57 -4808.0  
## + previous 1 0.880 373.95 -4791.9  
## + day 1 0.624 374.21 -4790.2  
## + balance 1 0.429 374.40 -4788.8  
## + pdays 1 0.315 374.52 -4788.1  
## <none> 374.83 -4788.0  
## + age 1 0.099 374.73 -4786.6  
## + month 1 0.062 374.77 -4786.4  
## + job 1 0.021 374.81 -4786.1  
## + default 1 0.003 374.83 -4786.0  
## + marital 1 0.002 374.83 -4786.0  
## - education 1 3.824 378.65 -4764.4  
## - loan 1 6.675 381.51 -4745.5  
## - housing 1 8.506 383.34 -4733.4  
## - contact 1 8.921 383.75 -4730.7  
## - poutcome 1 25.480 400.31 -4624.2  
## - duration 1 163.815 538.65 -3876.3  
##   
## Step: AIC=-4807.96  
## y ~ duration + poutcome + contact + housing + loan + education +   
## campaign  
##   
## Df Sum of Sq RSS AIC  
## + previous 1 0.805 370.77 -4811.4  
## + balance 1 0.372 371.20 -4808.5  
## + pdays 1 0.339 371.23 -4808.3  
## <none> 371.57 -4808.0  
## + day 1 0.247 371.33 -4807.6  
## + age 1 0.068 371.50 -4806.4  
## + month 1 0.023 371.55 -4806.1  
## + job 1 0.019 371.55 -4806.1  
## + marital 1 0.010 371.56 -4806.0  
## + default 1 0.001 371.57 -4806.0  
## - campaign 1 3.258 374.83 -4788.0  
## - education 1 3.923 375.50 -4783.5  
## - loan 1 6.594 378.17 -4765.6  
## - housing 1 8.406 379.98 -4753.6  
## - contact 1 9.481 381.05 -4746.5  
## - poutcome 1 22.225 393.80 -4663.6  
## - duration 1 163.847 535.42 -3889.4  
##   
## Step: AIC=-4811.43  
## y ~ duration + poutcome + contact + housing + loan + education +   
## campaign + previous  
##   
## Df Sum of Sq RSS AIC  
## + pdays 1 1.528 369.24 -4819.8  
## + balance 1 0.330 370.44 -4811.7  
## <none> 370.77 -4811.4  
## + day 1 0.246 370.52 -4811.1  
## + age 1 0.080 370.69 -4810.0  
## + month 1 0.015 370.75 -4809.5  
## + job 1 0.011 370.76 -4809.5  
## + marital 1 0.007 370.76 -4809.5  
## + default 1 0.002 370.77 -4809.4  
## - previous 1 0.805 371.57 -4808.0  
## - campaign 1 3.183 373.95 -4791.9  
## - education 1 3.825 374.59 -4787.6  
## - loan 1 6.703 377.47 -4768.3  
## - contact 1 8.586 379.35 -4755.7  
## - housing 1 8.841 379.61 -4754.0  
## - poutcome 1 9.620 380.39 -4748.9  
## - duration 1 163.585 534.35 -3892.4  
##   
## Step: AIC=-4819.83  
## y ~ duration + poutcome + contact + housing + loan + education +   
## campaign + previous + pdays  
##   
## Df Sum of Sq RSS AIC  
## + balance 1 0.355 368.88 -4820.3  
## + day 1 0.324 368.92 -4820.0  
## <none> 369.24 -4819.8  
## + age 1 0.117 369.12 -4818.6  
## + month 1 0.060 369.18 -4818.2  
## + marital 1 0.007 369.23 -4817.9  
## + job 1 0.003 369.24 -4817.9  
## + default 1 0.001 369.24 -4817.8  
## - pdays 1 1.528 370.77 -4811.4  
## - previous 1 1.994 371.23 -4808.3  
## - campaign 1 3.181 372.42 -4800.2  
## - education 1 3.437 372.68 -4798.5  
## - loan 1 6.447 375.69 -4778.2  
## - housing 1 6.487 375.73 -4777.9  
## - contact 1 9.432 378.67 -4758.3  
## - poutcome 1 11.146 380.39 -4746.9  
## - duration 1 163.727 532.97 -3897.0  
##   
## Step: AIC=-4820.26  
## y ~ duration + poutcome + contact + housing + loan + education +   
## campaign + previous + pdays + balance  
##   
## Df Sum of Sq RSS AIC  
## + day 1 0.305 368.58 -4820.3  
## <none> 368.88 -4820.3  
## - balance 1 0.355 369.24 -4819.8  
## + month 1 0.077 368.81 -4818.8  
## + age 1 0.067 368.82 -4818.7  
## + marital 1 0.010 368.87 -4818.3  
## + job 1 0.004 368.88 -4818.3  
## + default 1 0.001 368.88 -4818.3  
## - pdays 1 1.553 370.44 -4811.7  
## - previous 1 1.950 370.83 -4809.0  
## - campaign 1 3.128 372.01 -4801.0  
## - education 1 3.519 372.40 -4798.3  
## - loan 1 6.135 375.02 -4780.7  
## - housing 1 6.312 375.20 -4779.5  
## - contact 1 9.461 378.35 -4758.4  
## - poutcome 1 11.195 380.08 -4746.9  
## - duration 1 164.023 532.91 -3895.2  
##   
## Step: AIC=-4820.34  
## y ~ duration + poutcome + contact + housing + loan + education +   
## campaign + previous + pdays + balance + day  
##   
## Df Sum of Sq RSS AIC  
## <none> 368.58 -4820.3  
## - day 1 0.305 368.88 -4820.3  
## - balance 1 0.335 368.92 -4820.0  
## + month 1 0.069 368.51 -4818.8  
## + age 1 0.046 368.53 -4818.6  
## + marital 1 0.011 368.57 -4818.4  
## + job 1 0.002 368.58 -4818.4  
## + default 1 0.000 368.58 -4818.3  
## - pdays 1 1.629 370.21 -4811.2  
## - previous 1 1.995 370.58 -4808.7  
## - campaign 1 2.732 371.31 -4803.7  
## - education 1 3.540 372.12 -4798.3  
## - loan 1 6.057 374.64 -4781.3  
## - housing 1 6.325 374.91 -4779.5  
## - contact 1 9.522 378.10 -4758.1  
## - poutcome 1 11.059 379.64 -4747.8  
## - duration 1 164.016 532.60 -3894.7

### use command to see step results

houseStep

##   
## Call:  
## lm(formula = y ~ duration + poutcome + contact + housing + loan +   
## education + campaign + previous + pdays + balance + day,   
## data = train)  
##   
## Coefficients:  
## (Intercept) duration poutcome contact housing loan   
## -0.05814 0.94045 0.16269 -0.08243 0.10744 0.14624   
## education campaign previous pdays balance day   
## 0.05040 -0.16392 0.16791 -0.14789 0.05701 -0.04153