Healthy Neighborhoods - Finding Variables

Andrew Cistola

6/30/2019

Prep Code

library(dplyr)  
library(randomForest)  
library(MASS)  
library(reshape)  
  
setwd("C:/Users/drewc/Documents/healthy\_neighborhoods")

Random Forest to Sort Variables

rf = read.csv("rf/rf\_master\_dmacs.csv")  
  
rf$Tract <- NULL  
  
rf = rf %>% mutate\_if(is.factor, as.numeric)  
  
of <- randomForest(  
 formula = Diabetes ~ .,   
 data = rf,   
 ntree = 1000,  
 importance=TRUE)  
  
rank = importance(of)  
  
write.csv(rank, "C:/Users/drewc/Documents/healthy\_neighborhoods/rf/rf\_results\_rank.csv") # clean and transpose in excel

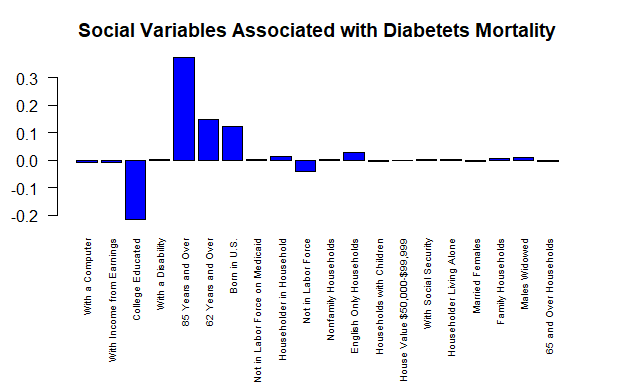
Bind Variables to Prepare Model

rank = read.csv("rf/rf\_results\_rank41.csv")  
rf = read.csv("rf/rf\_data\_dmacs.csv")  
  
bind = rbind.fill(rank, rf)  
  
write.csv(bind, "C:/Users/drewc/Documents/healthy\_neighborhoods/rf/rf\_results\_bind.csv") #remove NA and clean in excel  
mod = read.csv("rf/rf\_results\_bind.csv")  
  
frmla = as.formula(paste("Diabetes ~ ", paste(colnames(rank), collapse=" + "), sep = ""))  
  
fit = lm(frmla, data=mod)

Stepwise Multiple Regression Model

back <- stepAIC(fit, direction="backward")  
  
final <- data.frame(summary(back)$coefficients)  
  
write.csv(final, file = "C:/Users/drewc/Documents/healthy\_neighborhoods/rf/rf\_results\_dmacs.csv")

Build a Plot

finalcoef = c(-0.009209168, -0.009203037, -0.213959604, 0.004052260, 0.374193911, 0.150295698, 0.123187029, 0.004647710, 0.013869891, -0.038886763, 0.004852323, 0.027010880, -0.003714177, 0.001315302, 0.003673114, 0.003874608, -0.003758599, 0.006031646, 0.011160991, -0.004085458)  
  
finalvars = c("With a Computer", "With Income from Earnings", "College Educated", "With a Disability", "85 Years and Over", "62 Years and Over", "Born in U.S.", "Not in Labor Force on Medicaid", "Householder in Household", "Not in Labor Force", "Nonfamily Households", "English Only Households", "Households with Children", "House Value $50,000-$99,999", "With Social Security", "Householder Living Alone", "Married Females", "Family Households", "Males Widowed", "65 and Over Households")   
  
par(mar=c(9, 3, 3, 3))  
barplot(finalcoef, names.arg = finalvars, main = "Social Variables Associated with Diabetets Mortality", ylab = "Coefficient Value in Final Fit Model", col = "blue", las = 2, horiz = FALSE, cex.names = 0.6)