Some Selection

Yuchen Hu

11/26/2018

importance based on earth (numerical gpa)

```
library(tidyr)
library(earth)
```

```
## Loading required package: plotmo
```

```
## Loading required package: plotrix
```

```
## Loading required package: TeachingDemos
```

```
library(data.table)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

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

```
library(ggplot2)
food <- read.csv("food_coded.csv",stringsAsFactors = FALSE)
food_numeric <- food %>% select_if(is.numeric)
food_numeric <- food_numeric[complete.cases(food_numeric), ]

food_numeric_gpa <- merge(as.numeric(food$GPA),
                          food_numeric,by="row.names",all.x=FALSE)[,-1]
```

```
## Warning in merge(as.numeric(food$GPA), food_numeric, by = "row.names",
## all.x = FALSE): NAs introduced by coercion
```

```
food_numeric_gpa <- food_numeric_gpa[complete.cases(food_numeric_gpa),]
earth.food_numeric <- earth(x ~ ., data=food_numeric_gpa)
earth.food_numeric
```

```
## Selected 8 of 21 terms, and 7 of 47 predictors
## Termination condition: RSq changed by less than 0.001 at 21 terms
## Importance: tortilla_calories, healthy_feeling, on_off_campus, ...
## Number of terms at each degree of interaction: 1 7 (additive model)
## GCV 0.08841846    RSS 2.901324    GRSq 0.2978685    RSq 0.5959196
```

```
importance <- evimp (earth.food_numeric)
importance
```

```
##                       nsubsets   gcv    rss
## tortilla_calories            7 100.0  100.0
## healthy_feeling              6  69.5   80.3
## on_off_campus                5  51.5   67.7
## life_rewarding               3  31.5   48.4
## ethnic_food                  3  26.6   47.7
## eating_changes_coded1        2  27.3   39.9
## fruit_day                    2  27.0   39.2
```

simple regression on variables selected

```
var_selected <- c("x",rownames(importance))
food_numeric_gpa_selected <- food_numeric_gpa[,var_selected]
fit.food_numeric_gpa_selected <- lm(x~ . ,data = food_numeric_gpa_selected)
summary(fit.food_numeric_gpa_selected)
```

```
##
## Call:
## lm(formula = x ~ ., data = food_numeric_gpa_selected)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52629 -0.21013  0.05374  0.18828  0.66363
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.2763492  0.2965222  11.049 3.85e-15 ***
## tortilla_calories       0.0004790  0.0001932   2.479 0.016510 *
## healthy_feeling        -0.0770633  0.0215776  -3.571 0.000785 ***
## on_off_campus           0.1053584  0.0581889   1.811 0.076091 .
## life_rewarding          0.0299411  0.0179021   1.672 0.100552
## ethnic_food             0.1006899  0.0324104   3.107 0.003089 **
## eating_changes_coded1   0.0038665  0.0180389   0.214 0.831133
## fruit_day              -0.1322609  0.0464768  -2.846 0.006366 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2852 on 51 degrees of freedom
## Multiple R-squared:  0.4221, Adjusted R-squared:  0.3427
## F-statistic: 5.321 on 7 and 51 DF,  p-value: 0.0001302
```

feature selection via random forest (numerical gpa)

```
library(party)
```

```
## Loading required package: grid
## Loading required package: mvtnorm
## Loading required package: modeltools
## Loading required package: stats4
## Loading required package: strucchange
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
## Loading required package: sandwich
```

```
RF.food_numeric <- cforest(x ~ . , data= food_numeric_gpa,
                           control=cforest_unbiased(mtry=2,ntree=50))
RF.food_numeric
```

```
##
##   Random Forest using Conditional Inference Trees
##
## Number of trees:  50
##
## Response:  x
## Inputs:  Gender, breakfast, calories_chicken, calories_day, calories_scone, coffee, comfort_food_reas
## Number of observations:  59
```

```
varimp(RF.food_numeric) %>% sort() #mean decrease in accuracy
```

```
##             mother_education                  pay_meal_out
##                -1.046069e-03                -9.133664e-04
##                 parents_cook               life_rewarding
##                -8.166851e-04                -4.987036e-04
##                   eating_out                        drink
##                -4.703719e-04                -3.963370e-04
##             father_education            tortilla_calories
##                -2.457975e-04                -1.615688e-04
##                         cook                      income
##                -1.315433e-04                -1.256219e-04
##                       coffee                   greek_food
##                -5.503738e-05                -2.865258e-05
##                       Gender                    breakfast
##                -7.948236e-06                 0.000000e+00
##                 calories_day   comfort_food_reasons_coded
##                 0.000000e+00                 0.000000e+00
##                      cuisine                       fries
##                 0.000000e+00                 0.000000e+00
##                  indian_food                        soup
```

```
##                 0.000000e+00                 0.000000e+00
##                    thai_food                     vitamins
##                 0.000000e+00                 0.000000e+00
##                 persian_food                  veggies_day
##                 5.118713e-06                 2.218725e-05
##              turkey_calories comfort_food_reasons_coded.1
##                 2.705305e-05                 6.834321e-05
##         eating_changes_coded               calories_scone
##                 7.764502e-05                 1.532474e-04
##                   employment                  grade_level
##                 1.679142e-04                 2.282766e-04
##                     exercise              calories_chicken
##                 2.363530e-04                 2.637404e-04
##            nutritional_check               waffle_calories
##                 2.824234e-04                 3.259049e-04
##        eating_changes_coded1                marital_status
##                 3.716495e-04                 3.793353e-04
##                 italian_food                on_off_campus
##                 4.257912e-04                 4.818360e-04
##               healthy_feeling              ideal_diet_coded
##                 5.160172e-04                 5.351037e-04
##            fav_cuisine_coded                    fruit_day
##                 5.416653e-04                 5.545717e-04
##                     fav_food       self_perception_weight
##                 6.309459e-04                 8.790286e-04
##                       sports           diet_current_coded
##                 9.011096e-04                 1.022908e-03
##                  ethnic_food
##                 1.422699e-03
```

```
varimp(RF.food_numeric, conditional=TRUE) %>% sort()  #adjusts for correlations
```

```
##                 parents_cook                 pay_meal_out
##                -1.086139e-03                -8.924099e-04
##              mother_education                        drink
##                -6.143825e-04                -6.053831e-04
##                  ethnic_food                   greek_food
##                -4.668987e-04                -2.823450e-04
##                       Gender         eating_changes_coded1
##                -2.656501e-04                -2.453452e-04
##                   eating_out                       coffee
##                -1.994691e-04                -1.707529e-04
##                  veggies_day                calories_scone
##                -1.196832e-04                -7.400371e-05
##                  persian_food                    breakfast
##                -1.206424e-05                 0.000000e+00
##                  calories_day   comfort_food_reasons_coded
##                 0.000000e+00                 0.000000e+00
##                       cuisine                       fries
##                 0.000000e+00                 0.000000e+00
##                  indian_food                        soup
##                 0.000000e+00                 0.000000e+00
##                    thai_food                     vitamins
##                 0.000000e+00                 0.000000e+00
##                   employment                       income
```

```
##                 5.572645e-05                 8.894033e-05
##               life_rewarding       self_perception_weight
##                 9.837327e-05                 1.286683e-04
##                     exercise             father_education
##                 1.410065e-04                 2.134583e-04
##                  grade_level                 italian_food
##                 2.314497e-04                 2.422875e-04
##                         cook comfort_food_reasons_coded.1
##                 2.543069e-04                 2.673401e-04
##             calories_chicken             ideal_diet_coded
##                 3.081935e-04                 3.613923e-04
##            fav_cuisine_coded         eating_changes_coded
##                 3.817466e-04                 4.115186e-04
##           diet_current_coded              turkey_calories
##                 4.211474e-04                 4.689219e-04
##                     fav_food                on_off_campus
##                 5.290688e-04                 5.808755e-04
##               marital_status                    fruit_day
##                 5.832718e-04                 7.716169e-04
##            nutritional_check                       sports
##                 8.113917e-04                 8.844828e-04
##              healthy_feeling            tortilla_calories
##                 8.869927e-04                 9.067031e-04
##              waffle_calories
##                 1.088802e-03
```

regression on variables with positive decrease

```
RF.selected <- varimp(RF.food_numeric, conditional=TRUE) %>% sort()
var_RFselected <- c("x",names(RF.selected[RF.selected>0]))
food_numeric_gpa_RFselected <- food_numeric_gpa[,var_RFselected]
fit.food_numeric_gpa_RFselected <- lm(x~ . ,data = food_numeric_gpa_RFselected)
summary(fit.food_numeric_gpa_RFselected)
```

```
##
## Call:
## lm(formula = x ~ ., data = food_numeric_gpa_RFselected)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5386 -0.1709  0.0216  0.1441  0.4207
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.298e+00  6.581e-01   5.011 1.45e-05 ***
## eating_out                  -5.612e-03  4.682e-02  -0.120  0.90525
## coffee                      -1.338e-01  1.015e-01  -1.319  0.19565
## life_rewarding               3.028e-02  2.020e-02   1.499  0.14262
## veggies_day                  1.248e-03  6.069e-02   0.021  0.98371
## tortilla_calories            3.344e-04  2.425e-04   1.379  0.17640
## marital_status               1.176e-02  7.868e-02   0.149  0.88202
## ethnic_food                  9.978e-02  4.271e-02   2.336  0.02517 *
## turkey_calories              2.868e-04  3.177e-04   0.903  0.37267
```

```
## Gender                         1.135e-02  1.001e-01   0.113  0.91034
## nutritional_check              6.109e-02  3.941e-02   1.550  0.12982
## grade_level                    2.426e-02  4.411e-02   0.550  0.58574
## sports                         7.118e-03  1.107e-01   0.064  0.94910
## italian_food                   1.042e-01  7.143e-02   1.459  0.15332
## fruit_day                     -1.733e-01  7.678e-02  -2.257  0.03019 *
## on_off_campus                  1.258e-01  7.552e-02   1.666  0.10432
## ideal_diet_coded              -1.223e-02  2.038e-02  -0.600  0.55215
## comfort_food_reasons_coded.1 -2.799e-02  2.417e-02  -1.158  0.25440
## diet_current_coded            -8.523e-02  4.963e-02  -1.718  0.09447 .
## fav_food                       1.209e-02  5.510e-02   0.219  0.82757
## waffle_calories               -7.449e-05  1.898e-04  -0.393  0.69699
## self_perception_weight        -3.713e-02  4.408e-02  -0.842  0.40519
## healthy_feeling               -7.546e-02  2.358e-02  -3.200  0.00287 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2789 on 36 degrees of freedom
## Multiple R-squared:   0.61,  Adjusted R-squared:  0.3717
## F-statistic: 2.559 on 22 and 36 DF,  p-value: 0.005965
```