EDA Serguro Driver Safety

Kensen Tan

October 11, 2017

```
data_path = "C:/Users/tanke/OneDrive/Kaggle/Driver Safety/Data"
train = read.csv(paste0(data_path, "/train.csv"), stringsAsFactors = F)
train = train[,-1]
#remove userid
train_cont = train[, -grep(paste(c("cat","bin"), collapse="|"), names(train))]
#train data with only numerical features
train_cat_bin = train[, c(1,grep(paste(c("cat","bin"), collapse="|"), names(train)))]
#train data with only binary/categorical features
```

Extract Significant Correlation Pairs

```
corr_threshold = 0.15
pvalue_threshold = 0.05
sig_corr = data.frame(Relationship = character(), Correlation = double(), P_Value = double(), stringsAs

for (i in 1:ncol(train_cont)){
  for (j in 1:ncol(train_cont)){

    ctest = cor.test(train_cont[,i], train_cont[, j])

    if(i != j && ctest$p.value < pvalue_threshold && ctest$estimate > corr_threshold){
        new_row = data.frame(relationship = paste0(names(train_cont)[i], " vs ", names(train_cont)[j]),
                             correlation = ctest$estimate, p_value = ctest$p.value, row.names = NULL)
        sig_corr = rbind(sig_corr, new_row)
    }

  }
}

print(sig_corr)
```

```
##              relationship correlation p_value
## 1  ps_ind_01 vs ps_ind_03   0.2234076       0
## 2  ps_ind_01 vs ps_reg_02   0.1838548       0
## 3  ps_ind_01 vs ps_car_12   0.1618422       0
## 4  ps_ind_03 vs ps_ind_01   0.2234076       0
## 5  ps_ind_03 vs ps_ind_15   0.1704486       0
## 6  ps_ind_15 vs ps_ind_03   0.1704486       0
## 7  ps_reg_01 vs ps_reg_02   0.4710271       0
## 8  ps_reg_01 vs ps_reg_03   0.6370345       0
## 9  ps_reg_02 vs ps_ind_01   0.1838548       0
## 10 ps_reg_02 vs ps_reg_01   0.4710271       0
## 11 ps_reg_02 vs ps_reg_03   0.5164572       0
## 12 ps_reg_02 vs ps_car_12   0.1714158       0
## 13 ps_reg_02 vs ps_car_13   0.1943160       0
## 14 ps_reg_03 vs ps_reg_01   0.6370345       0
```

```
## 15 ps_reg_03 vs ps_reg_02   0.5164572       0
## 16 ps_car_12 vs ps_ind_01   0.1618422       0
## 17 ps_car_12 vs ps_reg_02   0.1714158       0
## 18 ps_car_12 vs ps_car_13   0.6717203       0
## 19 ps_car_13 vs ps_reg_02   0.1943160       0
## 20 ps_car_13 vs ps_car_12   0.6717203       0
## 21 ps_car_13 vs ps_car_15   0.5295186       0
## 22 ps_car_15 vs ps_car_13   0.5295186       0
```

Extract Interaction Pairs

```
target_correlation = round(cor(train_cont)[1,],3)
target_correlation_sorted = sort(target_correlation, decreasing = T)
print(target_correlation_sorted)
```

```
##     target  ps_car_13  ps_car_12  ps_reg_02  ps_reg_03  ps_car_15
##      1.000      0.054      0.039      0.035      0.031      0.028
##  ps_reg_01  ps_ind_01  ps_ind_03  ps_ind_14 ps_calc_01 ps_calc_03
##      0.023      0.019      0.008      0.007      0.002      0.002
## ps_calc_02 ps_calc_05 ps_calc_09 ps_calc_10 ps_calc_14 ps_calc_04
##      0.001      0.001      0.001      0.001      0.001      0.000
## ps_calc_06 ps_calc_07 ps_calc_11 ps_calc_13  ps_car_11 ps_calc_08
##      0.000      0.000      0.000      0.000     -0.001     -0.001
## ps_calc_12  ps_car_14  ps_ind_15
##     -0.001     -0.004     -0.022
```

```
interact_threshold = mean(target_correlation_sorted[2:6]) #0.0374
pvalue_threshold = 0.05
interact_corr = data.frame(Interact_Term = character(), Correlation = double(), P_Value = double(), str


for (i in 2:ncol(train_cont)){
  for(j in 2:ncol(train_cont)){

    ctest = cor.test(train_cont[,1], train_cont[,j]*train_cont[,i]) #correlation test between Y and int
    individualterm_threshold =  max(target_correlation[i], target_correlation[j])
    #threshold to ensure the interaction term has better correlation than only one componenent term

    if(ctest$estimate > max(individualterm_threshold, interact_threshold) && ctest$p.value < pvalue_thr
        new_row = data.frame(Interact_Term = paste0(names(train_cont)[i], " X ", names(train_cont)[j]),
                             correlation = ctest$estimate, p_value = ctest$p.value, row.names = NULL)
        interact_corr = rbind(interact_corr, new_row)

    }

  }
}


print(interact_corr)
```

```
##           Interact_Term correlation       p_value
## 1 ps_reg_02 X ps_car_15  0.03869700 5.465860e-196
## 2 ps_car_12 X ps_car_15  0.04405461 1.907596e-253
```

```
## 3 ps_car_15 X ps_reg_02  0.03869700 5.465860e-196
## 4 ps_car_15 X ps_car_12  0.04405461 1.907596e-253
```

Information Value

```
library(InformationValue)
```

```
## Warning: package 'InformationValue' was built under R version 3.3.3
```

```
IV_table = data.frame(Name = character(), IV = numeric())

for (i in 2:ncol(train_cat_bin)){
  IV_ = IV(X=factor(train_cat_bin[,i]), Y=train_cat_bin$target)
  new_row = data.frame(Name = colnames(train_cat_bin)[i], IV_)
  IV_table = rbind(IV_table, new_row)
}

print(IV_table[order(IV_table$IV_, decreasing = T),])
```

```
##              Name          IV_
## 25  ps_car_11_cat 6.847769e-02
## 3   ps_ind_05_cat 4.110201e-02
## 15  ps_car_01_cat 4.028846e-02
## 18  ps_car_04_cat 3.495814e-02
## 4   ps_ind_06_bin 3.459194e-02
## 20  ps_car_06_cat 3.381718e-02
## 13  ps_ind_17_bin 3.288829e-02
## 5   ps_ind_07_bin 3.081699e-02
## 21  ps_car_07_cat 2.833690e-02
## 17  ps_car_03_cat 2.761610e-02
## 16  ps_car_02_cat 2.594897e-02
## 12  ps_ind_16_bin 2.113447e-02
## 23  ps_car_09_cat 1.822207e-02
## 19  ps_car_05_cat 1.502336e-02
## 22  ps_car_08_cat 1.087619e-02
## 2   ps_ind_04_cat 7.012662e-03
## 6   ps_ind_08_bin 4.658925e-03
## 1   ps_ind_02_cat 3.835845e-03
## 7   ps_ind_09_bin 1.999937e-03
## 10  ps_ind_12_bin 1.468236e-03
## 14  ps_ind_18_bin 5.785738e-04
## 11  ps_ind_13_bin 1.454926e-04
## 9   ps_ind_11_bin 1.049083e-04
## 30 ps_calc_19_bin 8.682940e-05
## 8   ps_ind_10_bin 7.717706e-05
## 31 ps_calc_20_bin 3.290504e-05
## 24  ps_car_10_cat 3.173514e-05
## 27 ps_calc_16_bin 1.109028e-05
## 29 ps_calc_18_bin 8.678863e-06
## 26 ps_calc_15_bin 6.867809e-06
## 28 ps_calc_17_bin 8.240801e-07
```

No categorical/binary features have high IV values. (all of them are conidered "not predictive" by the IV table). More info about IV and WoE: http://documentation.statsoft.com/STATISTICAHelp.aspx?path=

WeightofEvidence/WeightofEvidenceWoEIntroductoryOverview.

PCA