OOA_Chris_Report

Chris Holman

2023-03-10

Libraries

```
library(dplyr)
library(ggplot2)
library(ggcorrplot)
library(factoextra)
library(randomForest)
library(gridExtra)
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Functions

```
# just converts 'null' to NA
nullToNA <- function(dataName) {
    return(replace(dataName, dataName == "null", NA))
}

# turns numbers as strings into numbers
factorize <- function(dataName) {
    return(as.numeric(as.character(dataName)))
}

tester = function(data, title = "data") {
    data = factorize(data)
    hist(data, main = c("Histogram of", title))
}
```

Initial Data Loading

```
load("C:/Users/chris/OneDrive/Desktop/Capstone/OOA_Proj/subset_adv_ltg_environment.RData")
```

Initial EDA

```
sub2 = adv_sub_ltg
for (i in 1:ncol(adv_sub_ltg)) {
    if (is.character(adv_sub_ltg[, i])) {
        sub2[, i] = nullToNA(adv_sub_ltg[, i])
    }
}
# columns with NA > 0
sum(colMeans(is.na(sub2)) > 0)
```

```
## [1] 143
```

```
# removing very high NA columns
sub3 <- sub2 %>%
    select(-c("Not_in_use", "TAS_ID", "UniqueID", "SubmitDate",
        "Source", "_rescued_data", "HomeAddressStreet1", "HomeAddressStreet2",
        "HomeAddressSystemID", "FirstName", "MiddleName", "LastName",
        "HomeCity"))
```

From the beginning we saw variables that did not seem useful so we removed them first. Our reasoning consisted of two ideas; NA rates that were too high or too general of information. Attributes like TAS_ID, UniqueID, and Address are unique to each individual so they wouldn't be much help for model building.

Checking Distributions

```
# some variables needed factorized
tester(sub3$cat_calc_social_score, "social score")
```

```
tester(sub3$amt_pop_per_capita_income, "percap income")
```

```
tester(sub3$val_donor_private_foundation, "private dno")
```

```
tester(sub3$val_donor_education_charities, "edu charities")
```

```
tester(sub3$ind_life_new_mover_12mos, "newmover12")
```

```
tester(sub3$ind_life_new_homeowner_12mos, "newhome12")
```

```
tester(sub3$cat_demo_dual_income, "dual income")
```

```
tester(sub3$ind_purchase_dm_multi_buyer, "multi buyer")
```

```
tester(sub3$n_purchase_mail_upscale_buyer, "upscale mail")
```

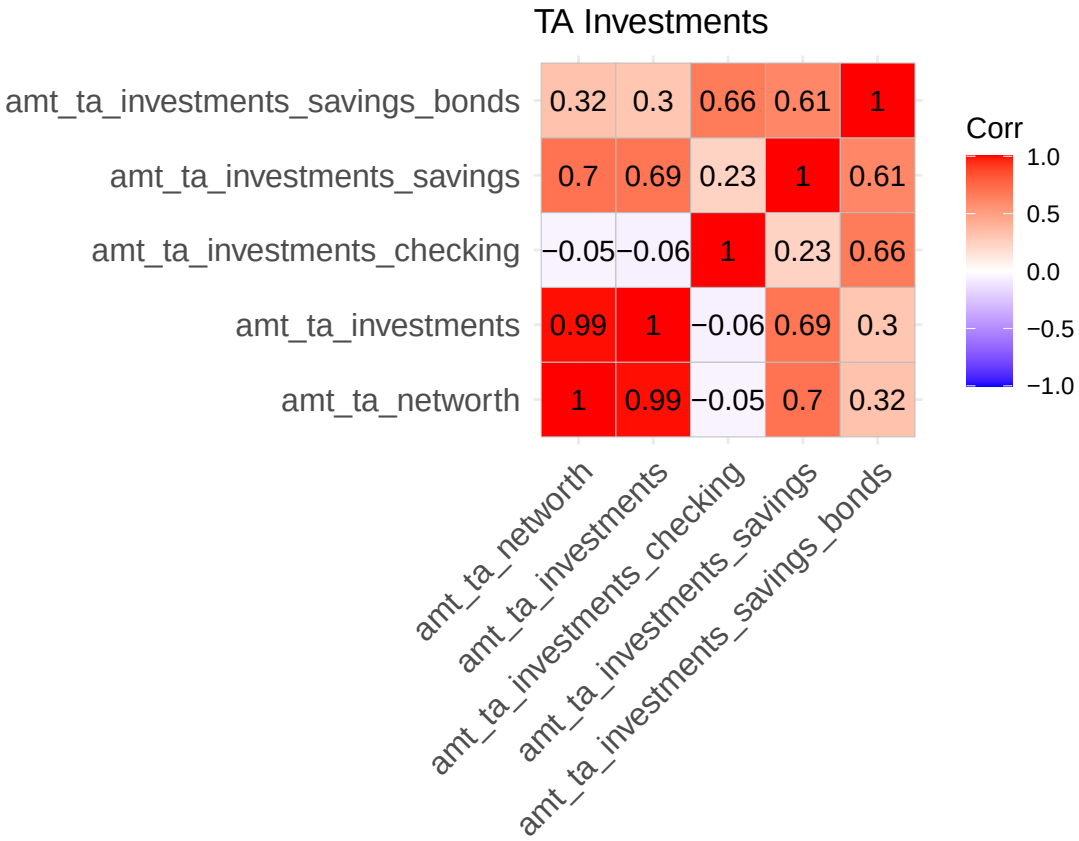
```
tester(sub3$cat_calc_political_persona, "politics")
```

```
tester(sub3$val_life_grandchildren, "grandchildren")
```

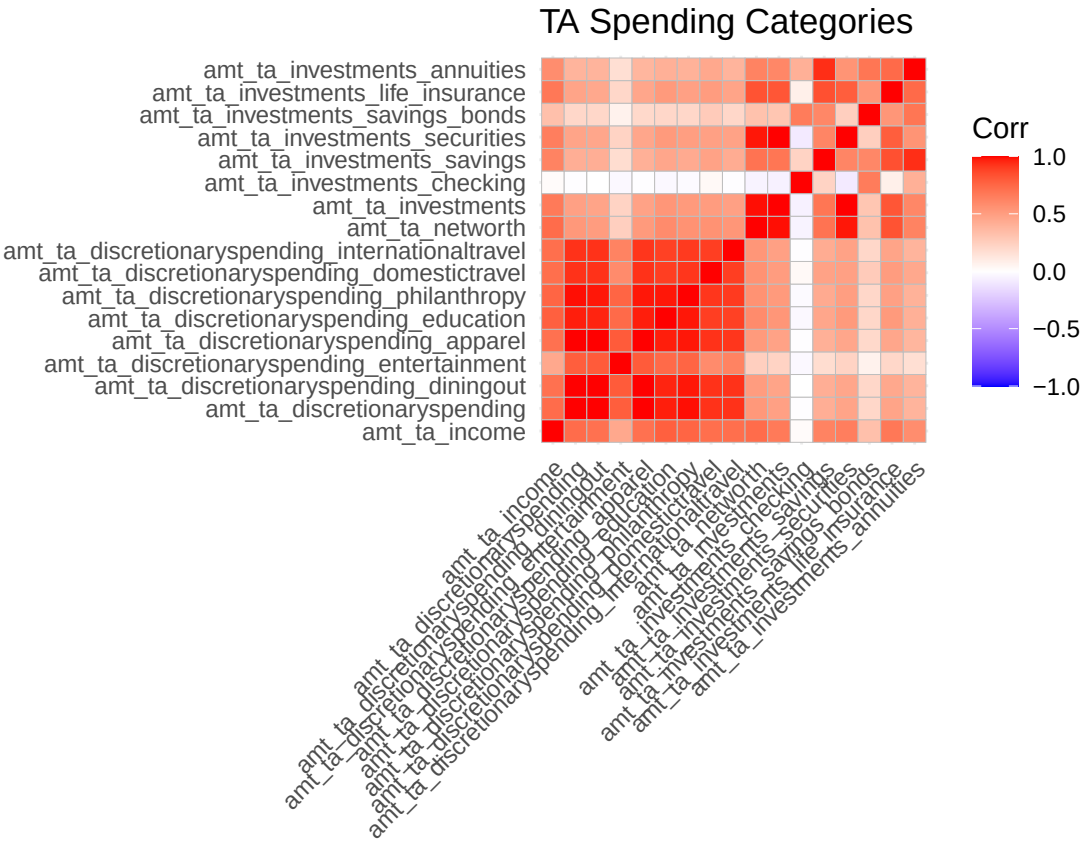
Checking the distributions for variables with consistent trends. Lots of variables have imbalanced class problem where most of the data is in one category. This won't be very helpful when building models, so they should potentially be removed. Examples of 'good' trends exist in variables like Social Score or Per Capita Income.

Checking Correlations and doing PCA on my section of intial data

```
sub4 = sub3[, c(50:98, 105)]
char_cols = sapply(sub4, is.character)
char_cols[c("cat_ta_total_identified_assets", "cat_demo_political_affiliation")] = FALSE
sub4[, char_cols] = as.data.frame(apply(sub4[, char_cols], 2,
    as.numeric))
bigcor = cor(na.omit(sub4[, c(23:26, 28)]))
ggcorrplot(bigcor, hc.order = FALSE, type = "full", lab = TRUE,
    title = "TA Investments")
```



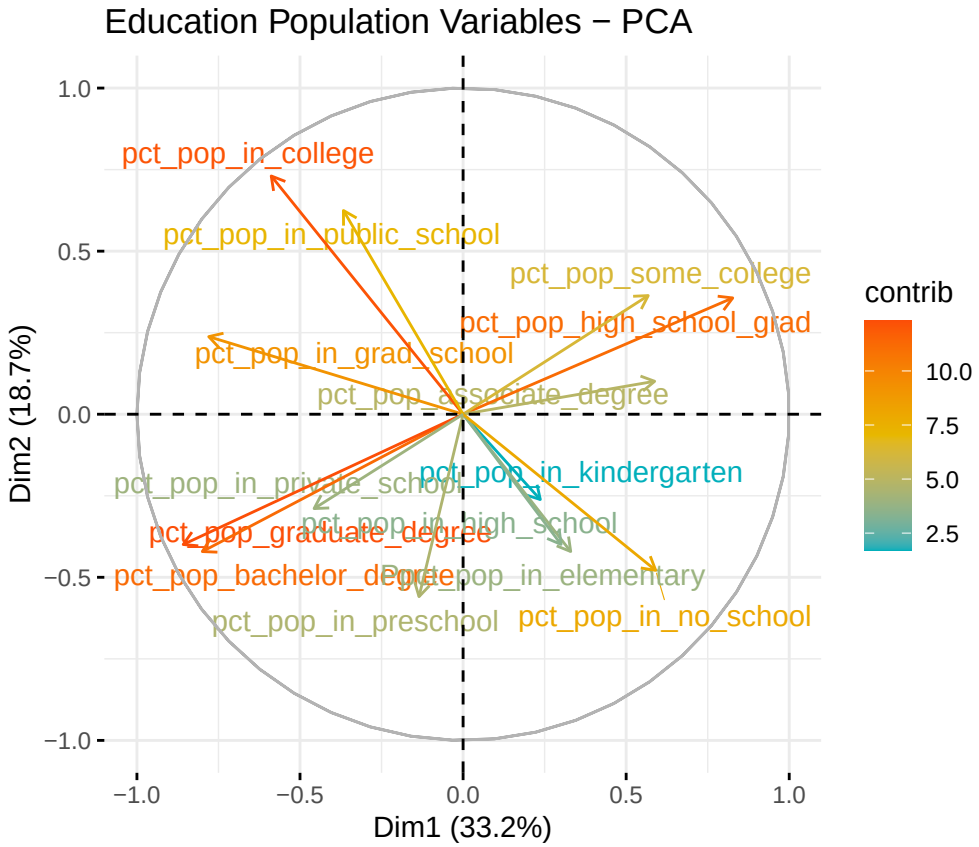

```
sub4_cor = na.omit(sub4[, 14:30])
corData = cor(sub4_cor)

ggcorrplot(corData, title = "TA Spending Categories", tl.cex = 9)
```



From correlations, we can see that we have a major multicollinearity problem. Lots of these variables from Target Analytics have very high correlations with each other. This means they encode similar information which can harm a model's ability to fit to the data accurately. We decided to remove the discretionary spending related variables except philanthropic spending. Also, we removed all investments related variables except investments in savings bonds because it had the least correlation with the other investments variables.

```
finance_pca = prcomp(na.omit(sub4[,23:30]), scale =TRUE)
#Scree Plot to determine number of PCs needed
#fviz_eig(finance_pca)

#fviz_pca_var(finance_pca,
#             col.var = "contrib", # Color by contributions to the PC
 #            gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
  #           repel = TRUE,     # Avoid text overlapping
   #          title='Finance Variables - PCA'
#)

#rotation matrix
#finance_pca$rotation

#checking seems least impactful (not by much, no major finding)

#on population dist
pop_pca = prcomp(na.omit(sub4[,36:49]), scale =TRUE)
#took more PC to explain substantial variance
#Scree Plot to determine number of PCs needed
#fviz_eig(pop_pca)
```

```
fviz_pca_var(pop_pca,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,     # Avoid text overlapping
             title='Education Population Variables - PCA'
)
```




```
#rotation matrix
#pop_pca$rotation
#kindergarten has lower contribution than the rest, maybe remove

par(mfrow=c(1,1))
```

PCA was used to understand how much dimension reduction was possible. For variables related to individuals finances, we saw that only 2 principle components was able to explain ~95% of the variation. As for the population education variables, it took almost all of the components to reach 95% of the variation. So most of these variable hold predictive power and should be used in the model. Specifically, variables such as Percent of Population in Kindergarten and Estimated Checking Account balance show low contribution to the principle components so they will be removed.

Loading in new dataset after removing vars from all of groups EDA

```
load("C:/Users/chris/OneDrive/Desktop/Capstone/OOA_Proj/projEnv.RData")
```

Renaming distributions with log transformations

```
logFix = c("amt_ta_income", "amt_ta_discretionaryspending_philanthropy",
    "pct_pop_in_college", "pct_pop_in_grad_school", "pct_pop_in_private_school",
    "amt_tran_total_dollars_purchase", "n_tran_credit_card_purchase",
    "amt_financial_assessed_home_value", "n_demo_length_of_residence",
    "amt_financial_estimated_available_equity", "amt_financial_estimated_monthly_mortgage")

for (col in logFix) {
    col_index <- which(colnames(subTrim) == col)  # Get the index of the column to modify
    colnames(subTrim)[col_index] <- paste(col, "log", sep = "_")  # Modify the column name
}
```

```
# cleaning data for random forest some variables log
# transformed weird so fixing -Inf values
subTrim = subTrim %>%
    select(c(-"n_tran_credit_card_purchase_log"))

subTrimRF = subTrim %>%
    select(-c("HomeState", "HomePostCode", "amt_financial_estimated_available_equity_log"))

subTrimRF$n_demo_length_of_residence_log[subTrimRF$n_demo_length_of_residence_log <
    0] <- 0

subTrimRF$amt_financial_assessed_home_value_log[subTrimRF$amt_financial_assessed_home_value_log <
    0] <- 0

subTrimRF$amt_financial_estimated_monthly_mortgage_log[subTrimRF$amt_financial_estimated_monthly_mortga
    0] <- 0


# drops from 250000 to 95762 :/
subTrimRF = na.omit(subTrimRF)
```

Running Random Forest

```
# persona response
rf <- randomForest(cat_score_p2p_persona_map ~ ., data = subTrimRF,
    importance = T, ntree = 25, maxnodes = 50)
z = importance(rf)
head(z[order(z[, 2], decreasing = T), ], n = 3)
```

```
##                               1 Go Getters 2 Caring Contributors
## val_score_p2p_diy_score           4.073514              4.990280
## val_score_philanthropic_score     1.104172              4.621531
## val_score_telemarketing_score     3.518039              4.557508
```

```
##                               3 Casual Contributors 4 Do Gooders
## val_score_p2p_diy_score                    3.554379   -0.7372299
## val_score_philanthropic_score             -1.436200    2.6740680
## val_score_telemarketing_score             -2.052927    2.5207076
##                               5 Generous Joes 6 Over Achievers
## val_score_p2p_diy_score            1.5432142        0.7774499
## val_score_philanthropic_score     -0.3173565        2.8945071
## val_score_telemarketing_score      8.1993047       -0.5850514
##                               7 Cause Enthusiasts 8 Thrill Seekers Average Joes
## val_score_p2p_diy_score                 -1.706693         2.903424     3.628204
## val_score_philanthropic_score            3.586202         4.544039     2.353092
## val_score_telemarketing_score            5.602689         7.934448     5.155071
##                               MeanDecreaseAccuracy MeanDecreaseGini
## val_score_p2p_diy_score                   5.375455        1655.9365
## val_score_philanthropic_score             7.792200         732.0599
## val_score_telemarketing_score             9.215079        2241.0205
```

```
subTrimRF2 = subTrimRF %>%
    select(-c("cat_score_p2p_persona_map"))

# LTG response
rf2 <- randomForest(amt_lifetime_giving_log ~ ., data = subTrimRF2,
    importance = T, ntree = 25, maxnodes = 50)

z2 = importance(rf2)
head(z2[order(z2[, 2], decreasing = T), ], n = 5)
```

```
##                                %IncMSE IncNodePurity
## val_score_telemarketing_score 12.026332     31267.922
## val_score_sustainer_score      7.590231     20673.938
## val_demo_age                  20.297923     14846.984
## val_score_p2p_diy_score        7.118600     12037.643
## val_score_philanthropic_score  8.909524      9374.465
```

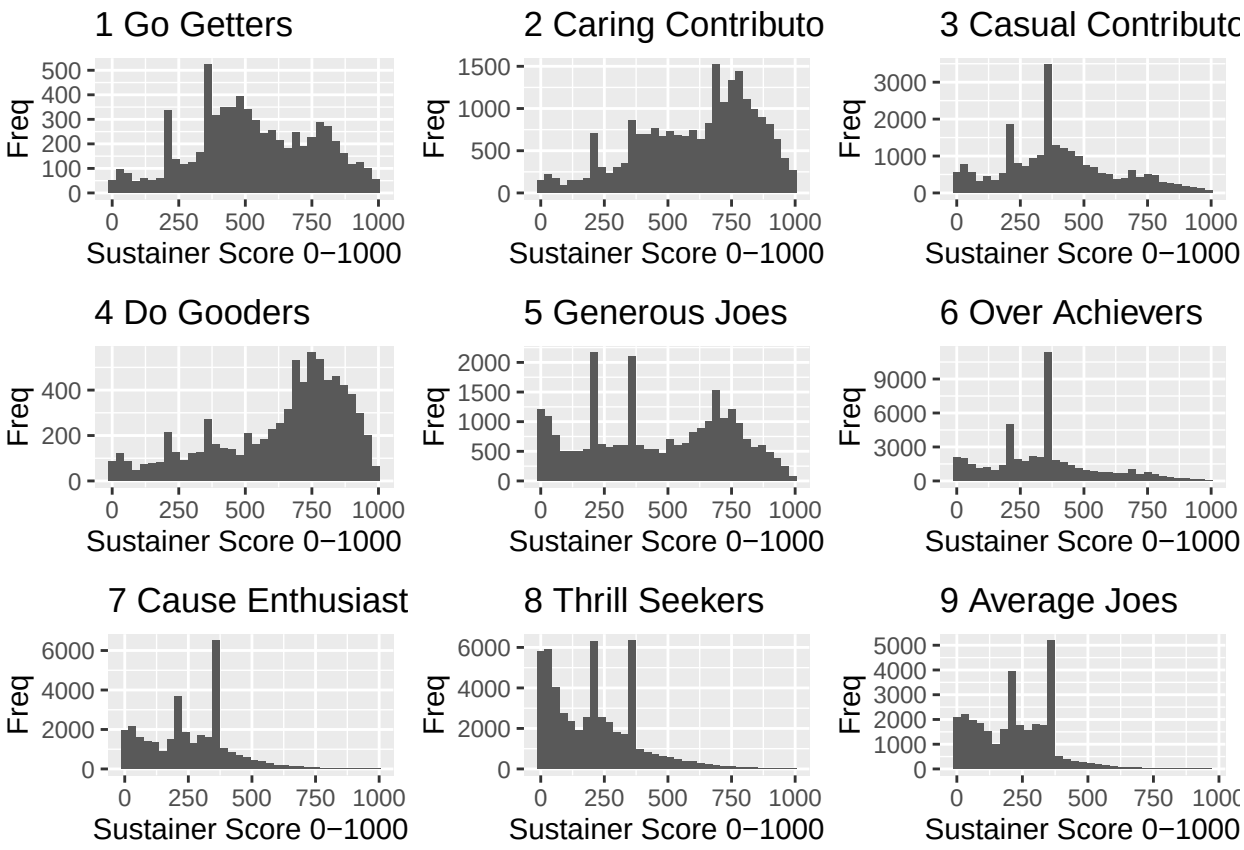
Ran the random forest with two response variables. The first model predicted the individual's lifetime giving total and the second predicted their p2p persona generated by Blackbaud. Both showed the Philantropic score, Telemarketing score, and Sustainer score were useful for predicting both responses.

Checking Trends with personas among important vars from RF

Sustainer Score

```
# fixing average joes to match formatting of other personas
levels(subTrim$cat_score_p2p_persona_map)[levels(subTrim$cat_score_p2p_persona_map) ==
    "Average Joes"] <- "9 Average Joes"
# vector to loop over all personas
personas = c("1 Go Getters", "2 Caring Contributors", "3 Casual Contributors",
    "4 Do Gooders", "5 Generous Joes", "6 Over Achievers", "7 Cause Enthusiasts",
    "8 Thrill Seekers", "9 Average Joes")
```

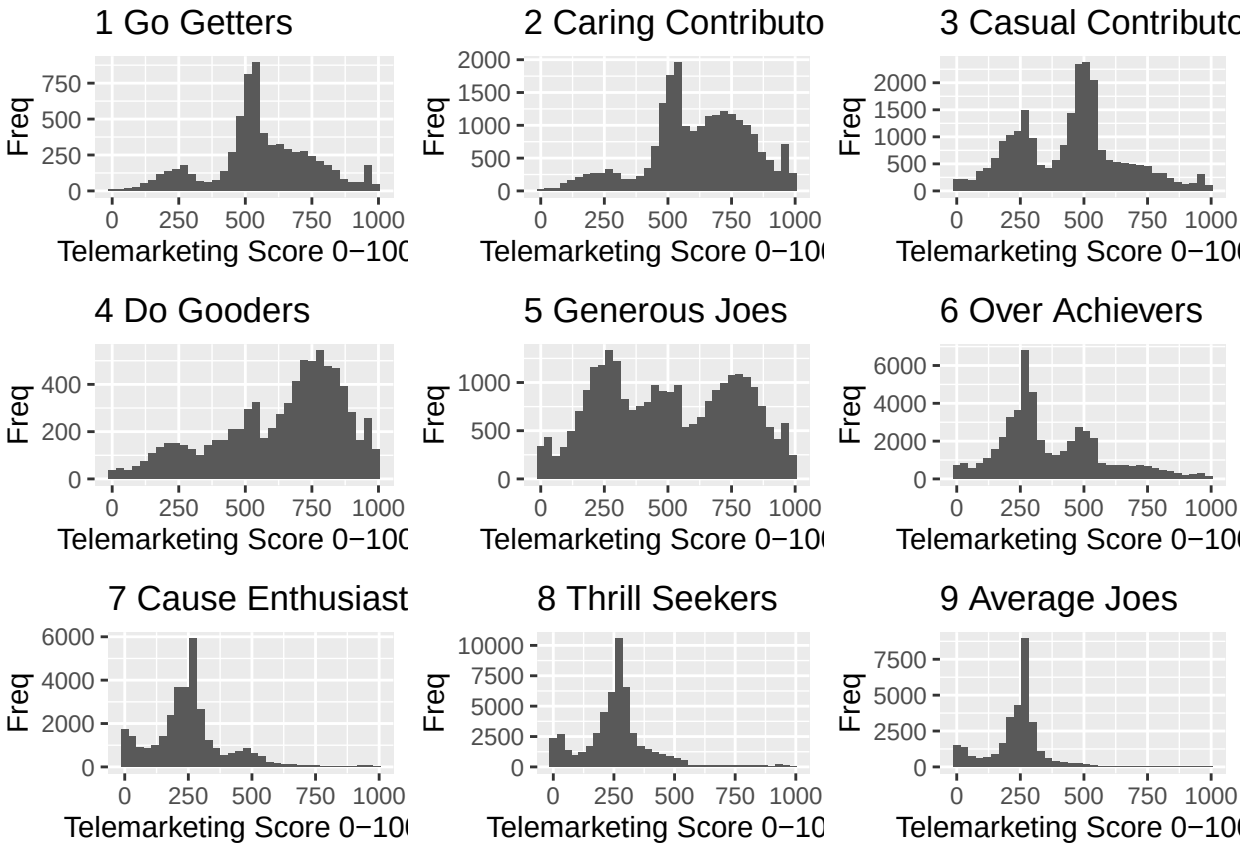
```
susPlots <- list()
for (i in 1:10) {
    susPlots[[i]] = ggplot(subTrim[subTrim$cat_score_p2p_persona_map ==
        personas[i], ], aes(x = val_score_sustainer_score)) +
        geom_histogram(binwidth = 30) + xlab("Sustainer Score 0-1000") +
        ggtitle(paste(substring(personas[i], 1, nchar(personas[i])))) +
        ylab("Freq")
}
grid.arrange(susPlots[[1]], susPlots[[2]], susPlots[[3]], susPlots[[4]],
    susPlots[[5]], susPlots[[6]], susPlots[[7]], susPlots[[8]],
    susPlots[[9]], ncol = 3)
```



Sustainer score was important according to our random forest models. Its distribution across the personas generally follows a constant decrease except for personas 2 and 4. Especially group 4, the Do Gooders, stand out as a group with a high frequency of individuals with high propensity to be a sustained donor.

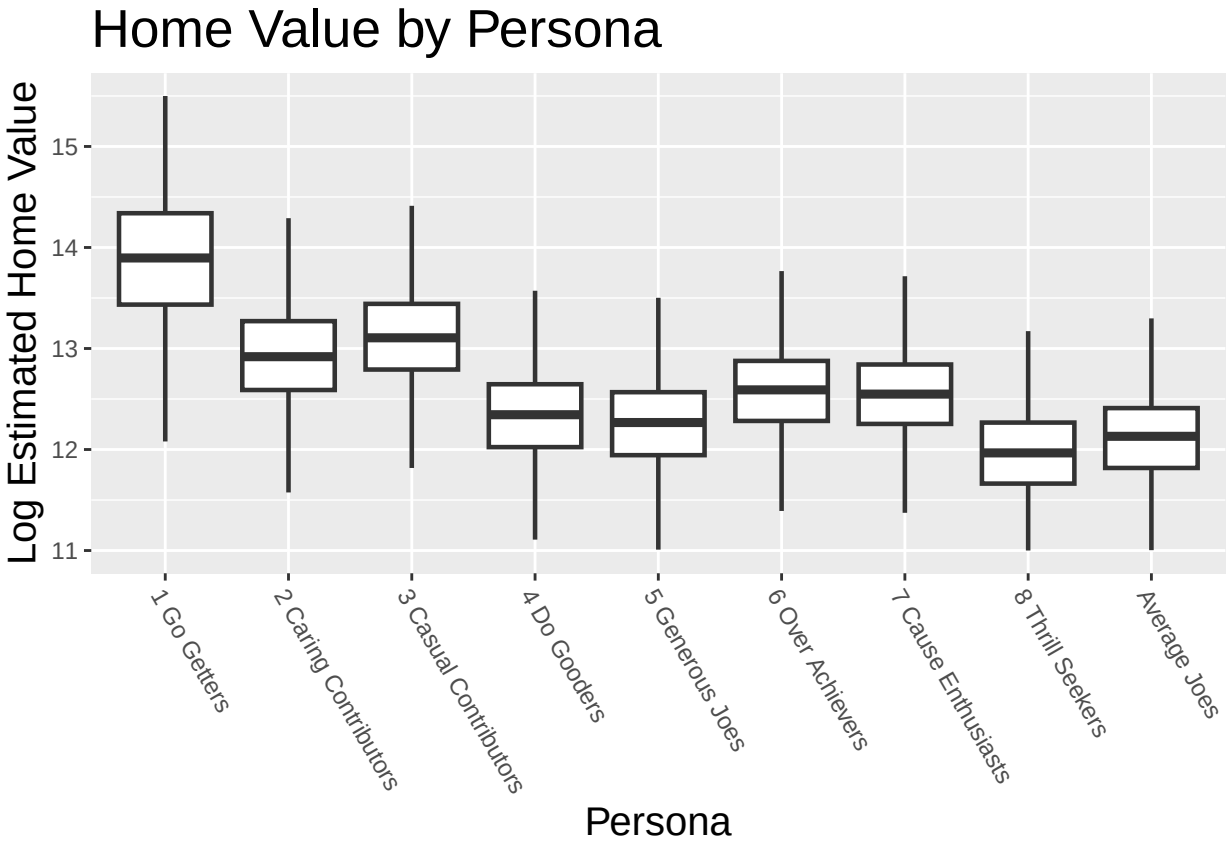
```
telePlots <- list()
for (i in 1:10) {
    telePlots[[i]] = ggplot(subTrim[subTrim$cat_score_p2p_persona_map ==
        personas[i], ], aes(x = val_score_telemarketing_score)) +
        geom_histogram(binwidth = 30) + xlab("Telemarketing Score 0-1000") +
        ggtitle(paste(substring(personas[i], 1, nchar(personas[i])))) +
        ylab("Freq")

}
grid.arrange(telePlots[[1]], telePlots[[2]], telePlots[[3]],
```

```
    telePlots[[4]], telePlots[[5]], telePlots[[6]], telePlots[[7]],
    telePlots[[8]], telePlots[[9]], ncol = 3)
```



Here we show another important variable, Telemarketing score, across all the p2p personas. The same constant decrease trend is still shown as we move down the personas. However, here groups 2 and 4 stand out quite a bit. These two groups have much higher frequencies of individuals with a high propensity to make a telemarketing gift.

```
# subTrimRF$cat_score_p2p_persona_map

ggplot(subTrimRF, aes(y = amt_financial_assessed_home_value_log,
    x = cat_score_p2p_persona_map)) + geom_boxplot(outlier.shape = NA,
    lwd = 0.8) + theme(axis.text.x = element_text(angle = 300,
    vjust = 1, hjust = 0), plot.title = element_text(size = 20),
    axis.title = (element_text(size = 15))) + xlab("Persona") +
    ylab("Log Estimated Home Value") + ggtitle("Home Value by Persona") +
    ylim(11, 15.5)
```



This plot sums up the overall trend of the financial variables that were found important by the random forest. The top three personas (groups 1-3) always were more well off than the rest of the bunch. However, within those three persona 2 typically ranked the lowest. This does not match the usual constantly decreasing trend seen in most financial variables.

Takeaways

After modeling with the random forest the main takeaways are the discoveries around personas 2 and 4. Persona 4, the Do Gooders, show high propensity to be sustained donors and to make telemarketing gifts. In the future, they could be good candidates for a phone campaign or a even a long term annual donation plan. Similarly, it seems that persona two, the Caring Contributors had propensity to given even though their financial standing was lower than expected. This could be interpreted as the persona describes people who give more than typical for their financial bracket. Further analysis could be done on these two groups to try to figure out attributes that link strongly with propensity to give.