Creating groups

Created	@Jan 4, 2021 5:15 PM
NeedToAddress	□
Tags	

1. Load data and remove the pilot id's from list

```
setwd('G:\\My Drive\\Research\\UMAA\\Code\\ml_model')
ml_data <- fread("valid_predict_proba.csv")
all_data <- fread("all_data_for_final.csv")

setwd('G:\\My Drive\\Research\\UMAA\\Data\\Pilot')
pilot_exclusion <- fread("pilot_exclusion_ids.csv")
# 958 people in the pilot


# excluding the people in pilot
# for ml_data, order by descending order of probability and create a rnk variable
ml_data <- ml_data[!ID_DEMO %in% pilot_exclusion$ID_DEMO][order(-prob)][, .(ID_DEMO, prob)][, rnk:=1:.N]
all_data <- all_data[,.(ID_DEMO)][!ID_DEMO %in% pilot_exclusion$ID_DEMO]
```

2. randomly sample 5000 for each of random and control groups

```
# randomly sampling 5000 each non members for control and random group
set.seed(42)
# shuffle the data first
shuffle_all_data <- all_data[sample(.N, .N)]
random_5000 <- shuffle_all_data[sample(.N, 5000)]
control_5000 <- shuffle_all_data[sample(.N, 5000)]
ml_40000 <- ml_data[rnk<=40000]
```

3. randomly distribute the common people in random and control groups

```
# common folks in random and control, ml_group
common_random_control <- merge.data.table(random_5000, control_5000, all=FALSE)
# there are 56 people in common


# tossing them into either group
common_allocation <- sample(c(1,0), nrow(common_random_control), replace = TRUE)
common_random_control <- common_random_control[, allocation:=common_allocation]

# 21 people were put back in random group
common_alloc_random <- copy(common_random_control)[allocation==0,][, allocation:=NULL]
# the remaining 35 people were put back in control group
common_alloc_control <- copy(common_random_control)[allocation==1,][, allocation:=NULL]


# finalizing the random and control groups by adding these people back
```

```
random_group_1 <-random_5000[!ID_DEMO %in% common_alloc_control$ID_DEMO]
control_group_1 <- control_5000[!ID_DEMO %in% common_alloc_random$ID_DEMO]
```

4. randomly distribute the common people in ML and random groups

```
# Repeating the same thing if there are common people between ML and random group
common_random_ml <- merge.data.table(random_group_1, ml_40000, all=FALSE)
# 408 people are common.
common_allocation <- sample(c(1,0), nrow(common_random_ml), replace = TRUE)

common_random_ml <- common_random_ml[, allocation:=common_allocation]

# 215 people were put back in random group
common_alloc_random <- copy(common_random_ml)[allocation==0,][, allocation:=NULL]
# 205 remaining people were put back in ML group
common_alloc_ml <- copy(common_random_ml)[allocation==1,][, allocation:=NULL]


ml_group_1 <- copy(ml_40000)[!ID_DEMO %in% common_alloc_random$ID_DEMO]
random_group_2 <- copy(random_group_1)[!ID_DEMO %in% common_alloc_ml$ID_DEMO]
```

5. randomly distribute the common people in ML and control groups

```
# repeating the same thing for people common between ML and control groups
common_control_ml <- merge.data.table(control_group_1, ml_group_1, all=FALSE)
# 442 people are common.
common_allocation <- sample(c(1,0), nrow(common_control_ml), replace = TRUE)

common_control_ml <- common_control_ml[, allocation:=common_allocation]

# 229 people were put back in control group
common_alloc_control <- copy(common_control_ml)[allocation==0,][, allocation:=NULL]
# 213 remaining people were put back in ML group
common_alloc_ml <- copy(common_control_ml)[allocation==1,][, allocation:=NULL]


ml_group_2 <- copy(ml_group_1)[!ID_DEMO %in% common_alloc_control$ID_DEMO]
control_group_2 <- copy(control_group_1)[!ID_DEMO %in% common_alloc_ml$ID_DEMO]
```

6. Splitting the rest of ml group into two using a coin toos

```
# Dividing ML group into 2
ml_group_alloc <- sample(c(1,0), nrow(ml_group_2), replace = TRUE)
ml_group_3 <- copy(ml_group_2)[, allocation:=ml_group_alloc]
ml_group_main <- copy(ml_group_3)[allocation==0,][, allocation:=NULL]
# 19985 remaining people were put in main ML group
ml_group_control <- copy(ml_group_3)[allocation==1,][, allocation:=NULL]
# 19571 remaining people were put in main ML group
```

7. concatenating the main ml group, control group and random group and generating the list

```
# list of people to be excluded
ml_list <- copy(ml_group_main)[,.(ID_DEMO)][, group:='ml_group']
```

```
control_list <- copy(control_group_2)[, group:='control_group']
random_list <- copy(random_group_2)[, group:='random_group']


exclusion_list <- rbind(ml_list, control_list, random_list)
setwd('G:\\My Drive\\Research\\UMAA\\Code\\ml_model')
fwrite(exclusion_list, "exclusion_list.csv")
```