LHR: Amsterdam Bias Analysis

Justin Braun

2024-10-13

Load Libraries

```
library(dplyr)
```

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

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

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

```
library(tidyr)
library(ggplot2)
library(openxlsx)
library(stringr)
library(friendlyeval)
```

Load Data

```
# Load confusion matrices
cms_raw <- read.xlsx('../input/Results_LHR/Output/20240308_CMs_LHR_SlimmeCheck.xlsx')
# Load feature count data
feature_counts_raw <- read.xlsx('../input/Results_LHR/Output/20240308_Important_Features_Counts.xlsx')
```

Preprocesing

```
### Confusion Matrices ###
summary(cms_raw)
```

```
##    Dataset              Model              Feature          Feature_Value
##  Length:480         Length:480         Length:480         Length:480
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Metric              Value
##  Length:480         Min.   :   0.0
##  Class :character   1st Qu.:  32.0
##  Mode  :character   Median :  83.0
##                     Mean   : 159.9
##                     3rd Qu.: 202.2
##                     Max.   :1024.0
```

```
print(unique(cms_raw$Dataset))
```

```
## [1] "Pilot"         "Prepilot"      "TrainingTrain" "TrainingTest"
```

```
print(unique(cms_raw$Model))
```

```
## [1] "BR" "AR"
```

```
print(unique(cms_raw$Feature))
```

```
## [1] "geslacht"         "Leeftijd&lt;30"   "Leeftijd&lt;40"   "Leeftijd&lt;50"
## [5] "IsNederlands"     "IsWesters"        "IsFulltimeParent" "IsParttimeParent"
```

```
cms_raw <- cms_raw %>%
  #0s indicate small sample sizes but are unlikely to be correct
  mutate(Value = ifelse(Value == 0, NA, Value))

#combine train and test since the original split is not actually recreated
cms_train <- cms_raw %>%
  filter(Dataset == 'TrainingTrain') %>%
  rename(Value_Train = Value) %>%
  dplyr::select(-Dataset)

cms_test <- cms_raw %>%
  filter(Dataset == 'TrainingTest') %>%
  rename(Value_Test = Value) %>%
  dplyr::select(-Dataset)

cms_train_test <- cms_train %>%
  left_join(cms_test, by = c('Model', 'Feature', 'Feature_Value', 'Metric')) %>%
  mutate(Value = Value_Train + Value_Test,
         Dataset = 'TrainTest') %>%
  dplyr::select(-Value_Train, -Value_Test)

cms_wide <- cms_raw %>%
  #merge combined train/test data
```

```
filter(!(Dataset %in% c('TrainingTrain', 'TrainingTest'))) %>%
bind_rows(cms_train_test) %>%

#compute shares
group_by(Feature, Feature_Value, Dataset, Model) %>%
mutate(Share = (Value/sum(Value)) * 100,
       group_size = sum(Value)) %>%
ungroup() %>%
dplyr::select(-Value) %>%

# pivot to dataset x model x group level
pivot_wider(names_from = Metric, values_from = Share) %>%

# compute various performance/fairness metrics
mutate(TOTAL = TN+FP+TP+FN,
       ACT_N = FP + TN,
       ACT_P = FN + TP,
       PRED_P = FP + TP,
       PRED_N = FN + TN,
       FPR = (FP/ACT_N) * 100,
       PPV = (TP/PRED_P) * 100,
       FDR = (FP/PRED_P) * 100,
       TPR = (TP/ACT_P) * 100,
       STAT_PAR = PRED_P,
       ERROR = FP+FN) %>%

# translate features, values, and models
mutate(Feature_EN = case_when(Feature == 'geslacht' ~ 'gender',
                              Feature == 'Leeftijd&lt;30' ~ 'Age < 30',
                              Feature == 'Leeftijd&lt;40' ~ 'Age < 40',
                              Feature == 'Leeftijd&lt;50' ~ 'Age < 50',
                              Feature == 'IsNederlands' ~ 'Dutch',
                              Feature == 'IsWesters' ~ 'Western',
                              .default = Feature),
       Feature_Value_EN = case_when(Feature_Value == 'V' ~ 'F',
                                    Feature == 'Leeftijd&lt;30' & Feature_Value == 1 ~ 'below 30',
                                    Feature == 'Leeftijd&lt;30' & Feature_Value == 0 ~ 'above 30',
                                    Feature == 'Leeftijd&lt;40' & Feature_Value == 1 ~ 'below 40',
                                    Feature == 'Leeftijd&lt;40' & Feature_Value == 0 ~ 'above 40',
                                    Feature == 'Leeftijd&lt;50' & Feature_Value == 1 ~ 'below 50',
                                    Feature == 'Leeftijd&lt;50' & Feature_Value == 0 ~ 'above 50',
                                    Feature == 'IsNederlands' & Feature_Value == 1 ~ 'Dutch',
                                    Feature == 'IsNederlands' & Feature_Value == 0 ~ 'Not Dutch',
                                    Feature == 'IsWesters' & Feature_Value == 1 ~ 'Western',
                                    Feature == 'IsWesters' & Feature_Value == 0 ~ 'Not Western',
                                    Feature == 'IsFulltimeParent' & Feature_Value == 1 ~ 'Full-time pa
                                    Feature == 'IsFulltimeParent' & Feature_Value == 0 ~ 'Not full-tir
                                    Feature == 'IsParttimeParent' & Feature_Value == 1 ~ 'Part-time pa
                                    Feature == 'IsParttimeParent' & Feature_Value == 0 ~ 'Not part-tir
                                    .default = Feature_Value),
       is_privileged_group = case_when(Feature_Value == 'V' ~ 0,
                                       Feature_Value == 'M' ~ 1,
                                       Feature == 'Leeftijd&lt;30' & Feature_Value == 1 ~ 1,
```

```
                                        Feature == 'Leeftijd&lt;30' & Feature_Value == 0 ~ 0,
                                        Feature == 'Leeftijd&lt;40' & Feature_Value == 1 ~ 1,
                                        Feature == 'Leeftijd&lt;40' & Feature_Value == 0 ~ 0,
                                        Feature == 'Leeftijd&lt;50' & Feature_Value == 1 ~ 1,
                                        Feature == 'Leeftijd&lt;50' & Feature_Value == 0 ~ 0,
                                        Feature == 'IsNederlands' & Feature_Value == 1 ~ 1,
                                        Feature == 'IsNederlands' & Feature_Value == 0 ~ 0,
                                        Feature == 'IsWesters' & Feature_Value == 1 ~ 1,
                                        Feature == 'IsWesters' & Feature_Value == 0 ~ 0,
                                        Feature == 'IsFulltimeParent' & Feature_Value == 1 ~ 0,
                                        Feature == 'IsFulltimeParent' & Feature_Value == 0 ~ 1,
                                        Feature == 'IsParttimeParent' & Feature_Value == 1 ~ 0,
                                        Feature == 'IsParttimeParent' & Feature_Value == 0 ~ 1,
                                        .default = NA),
         stage = paste0(Dataset, '/', Model),
         Model_EN = case_when(Model == 'BR' ~ 'Before Reweighing',
                              Model == 'AR' ~ 'After Reweighing'))

write.csv(cms_wide, '../output/cms_wide.csv')

# convert to long
cms_long <- cms_wide %>%
  pivot_longer(cols = c("TN", "FP", "FN", "TP", "TOTAL", "ACT_N", "ACT_P", "PRED_P", "PRED_N", "FPR", "
               names_to = 'Metric', values_to = 'Value')

write.csv(cms_long, '../output/cms_long.csv')

# compute Amsterdam definitions
cms_amsti <- cms_wide %>%
  dplyr::select('Dataset', 'Model', 'Model_EN', 'Feature_EN', 'is_privileged_group', 'Feature_Value_EN'
  mutate(is_privileged_group = ifelse(is_privileged_group ==1, 'privileged', 'unprivileged')) %>%
  pivot_wider(names_from = 'is_privileged_group', values_from =  c('Feature_Value_EN', 'group_size', 'F
  mutate(FP_amsti_diff = 100*((FP_unprivileged-FP_privileged)/FP_privileged),
         FPR_amsti_diff = 100*((FPR_unprivileged-FPR_privileged)/FPR_privileged),
         FDR_amsti_diff = 100*((FDR_unprivileged-FPR_privileged)/FDR_privileged),
         PRED_P_amsti_diff = 100*((PRED_P_unprivileged-PRED_P_privileged)/PRED_P_privileged))
write.csv(cms_amsti, '../output/cms_amsti.csv')
```

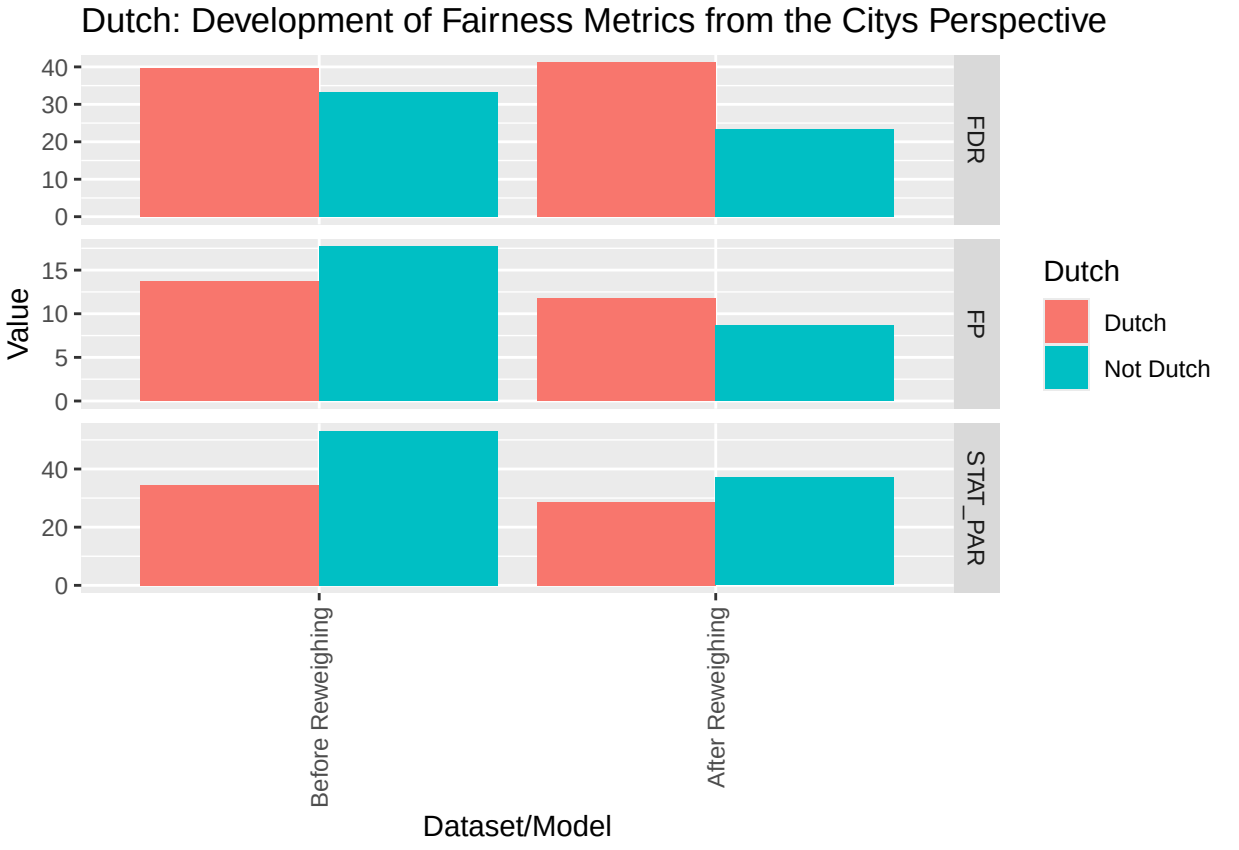
```
### Feature Importance ###
feature_counts <- feature_counts_raw %>%
  mutate(Count = ifelse(Count == 0, NA, Count),
         # translate
         Feature_EN = case_when(Feature == 'geslacht' ~ 'gender',
                                Feature == 'Leeftijd&lt;30' ~ 'Age < 30',
                                Feature == 'Leeftijd&lt;40' ~ 'Age < 40',
                                Feature == 'Leeftijd&lt;50' ~ 'Age < 50',
                                Feature == 'IsNederlands' ~ 'Dutch',
                                Feature == 'IsWesters' ~ 'Western',
                                .default = Feature),
         Feature_Value_EN = case_when(Feature == 'geslacht' & Value == 1 ~ 'F', #not sure about gender
                                      Feature == 'geslacht' & Value == 0 ~ 'M',
                                      Feature == 'Leeftijd&lt;30' & Value == 1 ~ 'below 30',
                                      Feature == 'Leeftijd&lt;30' & Value == 0 ~ 'above 30',
```

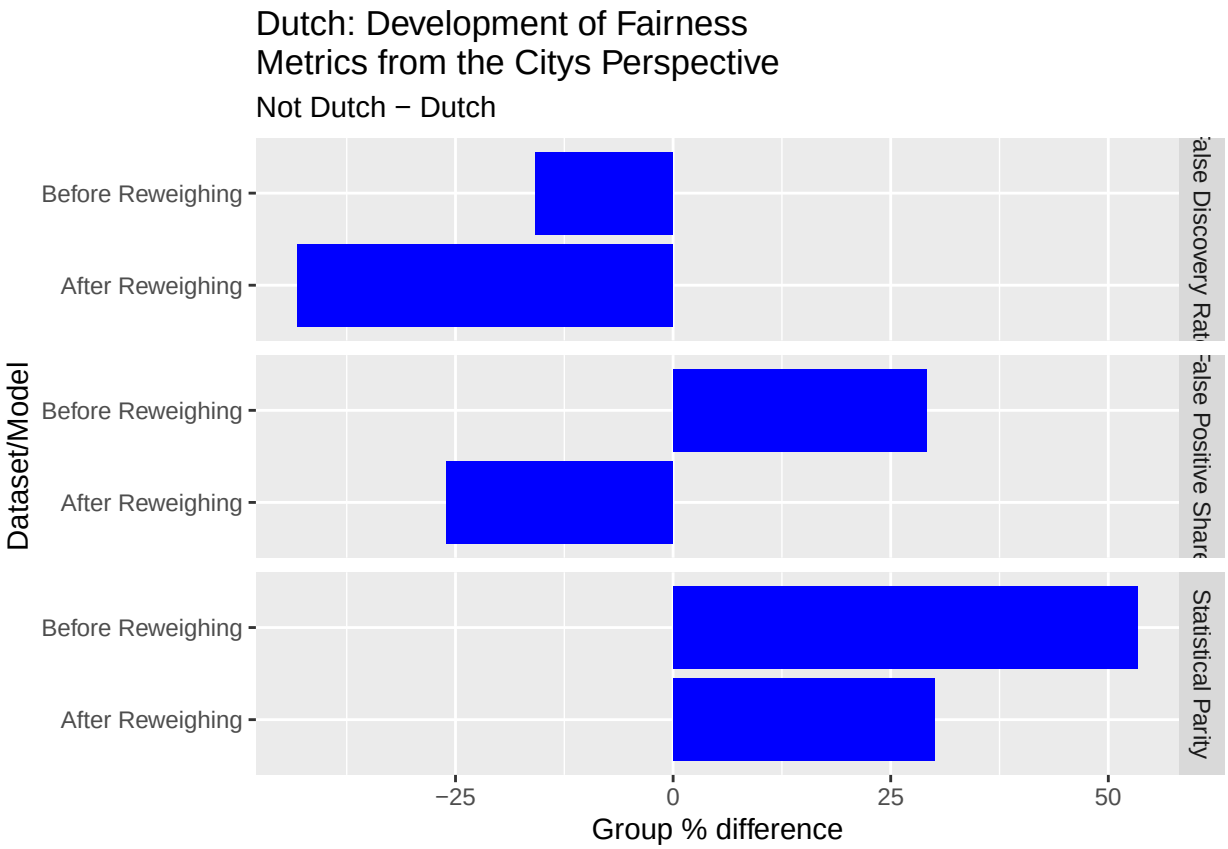
```
                                      Feature == 'Leeftijd&lt;40' & Value == 1 ~ 'below 40',
                                      Feature == 'Leeftijd&lt;40' & Value == 0 ~ 'above 40',
                                      Feature == 'Leeftijd&lt;50' & Value == 1 ~ 'below 50',
                                      Feature == 'Leeftijd&lt;50' & Value == 0 ~ 'above 50',
                                      Feature == 'IsNederlands' & Value == 1 ~ 'Dutch',
                                      Feature == 'IsNederlands' & Value == 0 ~ 'Not Dutch',
                                      Feature == 'IsWesters' & Value == 1 ~ 'Western',
                                      Feature == 'IsWesters' & Value == 0 ~ 'Not Western',
                                      Feature == 'IsFulltimeParent' & Value == 1 ~ 'Full-time parent',
                                      Feature == 'IsFulltimeParent' & Value == 0 ~ 'Not full-time paren
                                      Feature == 'IsParttimeParent' & Value == 1 ~ 'Part-time parent',
                                      Feature == 'IsParttimeParent' & Value == 0 ~ 'Not part-time paren
                                      .default = as.character(Value))) %>%
  group_by(Feature_EN, Feature_Value_EN, dataset) %>%
  mutate(share = (Count/sum(Count, na.rm = T)) * 100) %>% #note that I remove NAs which are presumably
  ungroup()
```

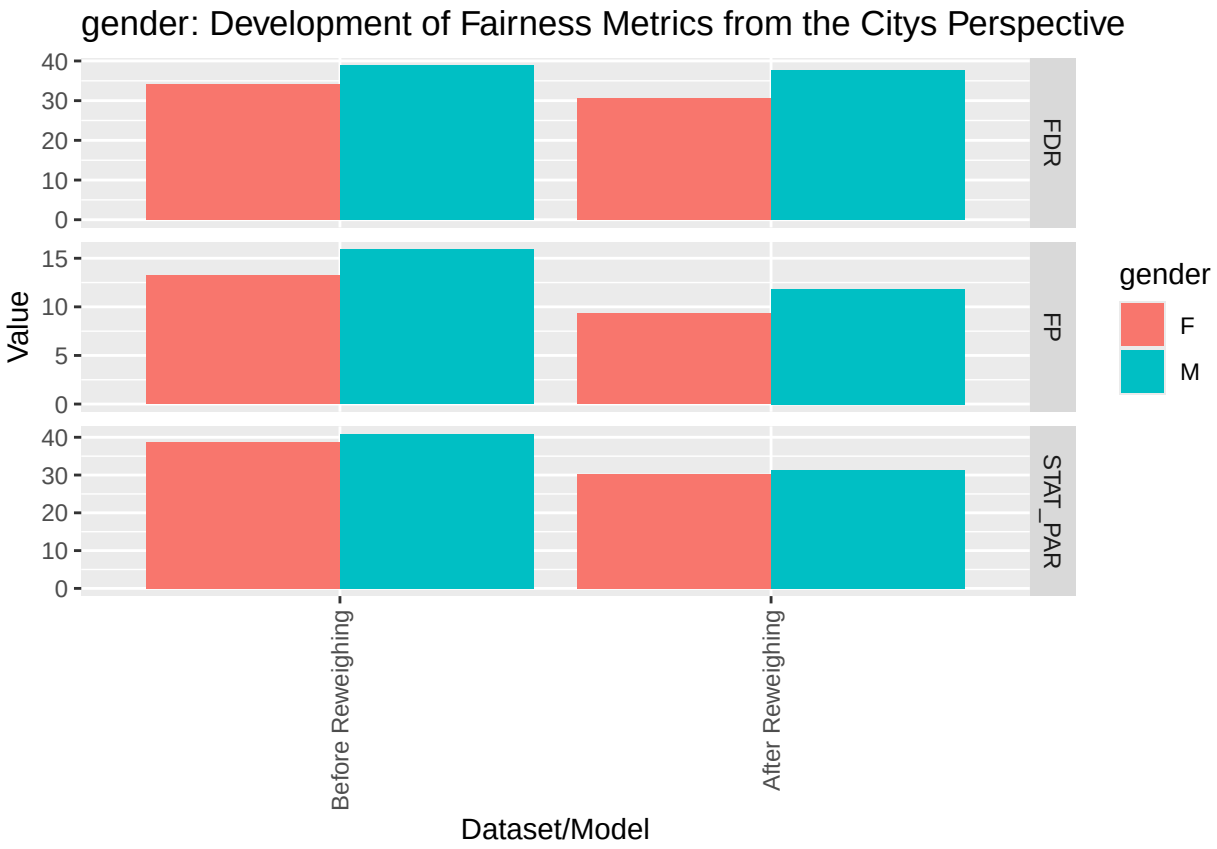
RQ 1 How did Reweighing affect various fairness metrics?

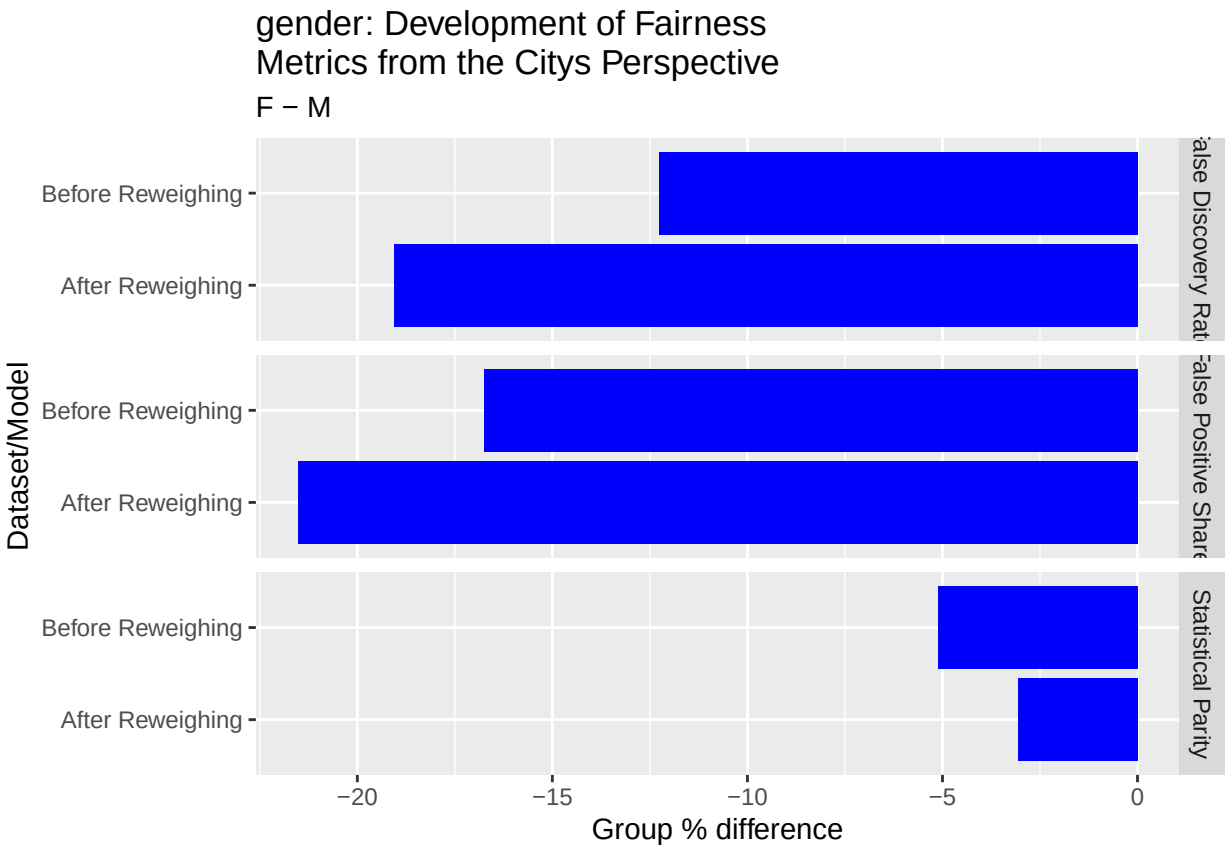
```
# subset to relevant definitions, datasets, and characteristics
cms_city_perspective <- cms_long %>%
  filter(stage %in% c('Prepilot/BR', 'Prepilot/AR'),
         Metric %in% c('STAT_PAR', 'FDR', 'FP'),
         Feature_EN %in% c('gender', 'Age < 30', 'Age < 50', 'Dutch', 'IsFulltimeParent', 'Western')) %
  mutate(order = case_when(stage == 'Prepilot/BR' ~ 1,
                           stage == 'Prepilot/AR' ~ 2,
                           .default = NA))

cms_char_diff <- data.frame()

#loope over characteristics
for(characteristic in c('Dutch', 'gender', 'IsFulltimeParent')){
  cms_char <- cms_city_perspective %>%
    filter(Feature_EN == characteristic)

  # compare metrics by group
  p1 <- ggplot(cms_char, aes(x = reorder(Model_EN, order), y = Value, fill = Feature_Value_EN))+
    geom_bar(stat = 'identity', position = position_dodge())+
    facet_grid(Metric ~., scales = "free_y")+
    labs(x = 'Dataset/Model',
         title = paste0(characteristic, ': Development of Fairness Metrics from the Citys Perspective')
         fill = characteristic)+
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

  print(p1)
  ggsave(paste0('../output/rq1_p1_', characteristic, '.png'), plot = p1, width = 10, height = 8)

  # compute Amsterdam-style comparisons
  feature_val_privileged <- unique(cms_char[cms_char$is_privileged_group == 1,]$Feature_Value_EN)[1]
  feature_val_unprivileged <- unique(cms_char[cms_char$is_privileged_group == 0,]$Feature_Value_EN)[1]
```

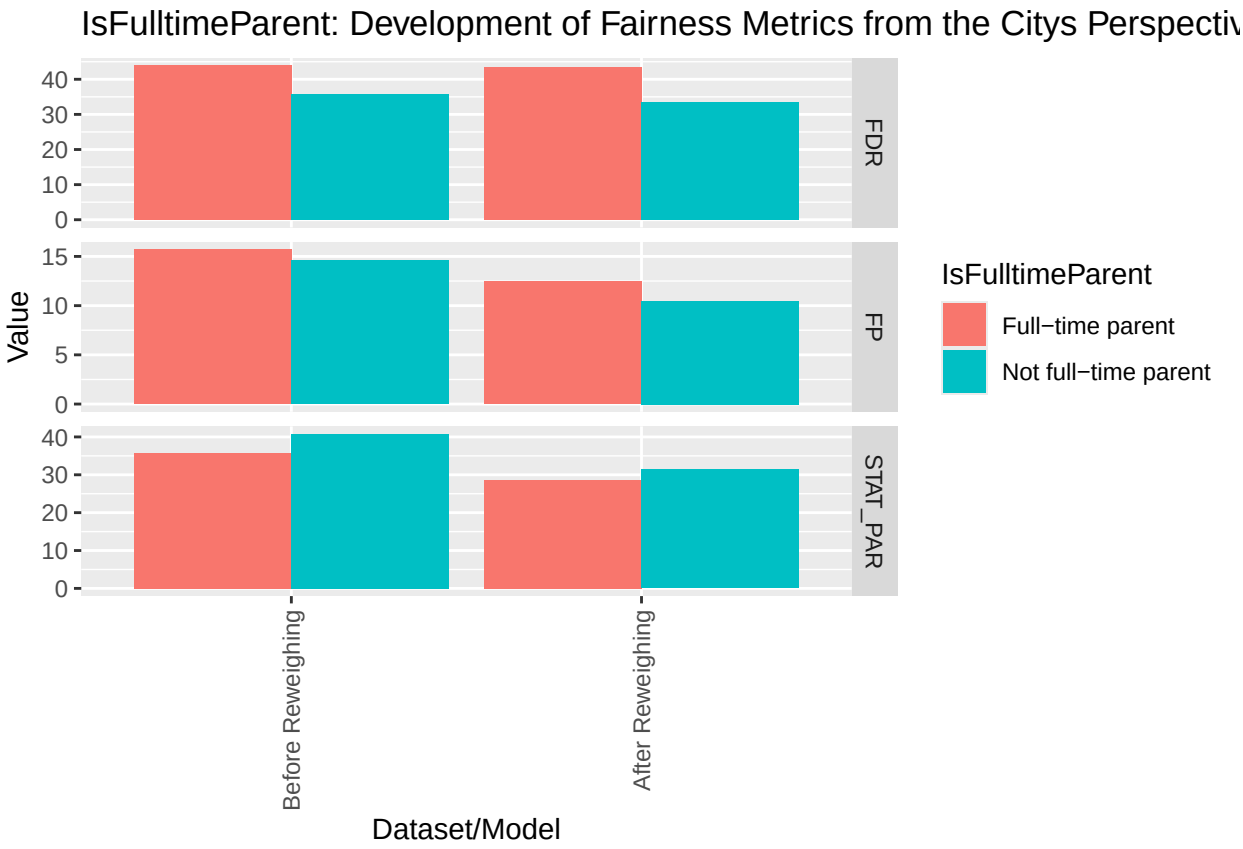
```r
  cms_char_diff <- cms_char %>%
    dplyr::select(-Feature_Value, -group_size, -is_privileged_group) %>%
    pivot_wider(names_from = 'Feature_Value_EN', values_from = 'Value') %>%
    #using Amsterdam's difference op here, though not sure the ref cat is always the same
    mutate(Diff = 100*(.data[[feature_val_unprivileged]] -  .data[[feature_val_privileged]])/.data[[fea
           Metric = case_when(Metric == 'FP' ~ 'False Positive Share',
                              Metric == 'FDR' ~ 'False Discovery Rate',
                              Metric == 'STAT_PAR' ~ 'Statistical Parity',
                              .default = NA))

  # plot difference as operationalized by the city reports
  p2 <- ggplot(cms_char_diff, aes(x = reorder(Model_EN, -order), y = Diff))+
    geom_bar(stat = 'identity', position = position_dodge(), fill = 'blue') +
    facet_grid(Metric ~., scales = "free_y")+
    labs(x = 'Dataset/Model',
         y = 'Group % difference',
         title = paste0(characteristic, ': Development of Fairness \nMetrics from the Citys Perspective
         subtitle = paste0(feature_val_unprivileged, ' - ', feature_val_privileged))+
    coord_flip()

  print(p2)
  ggsave(paste0('../output/rq1_p2_', characteristic, '.png'), plot = p2, width = 10, height = 8)

  write.csv(cms_char_diff, paste0('../output/fairness_definition_cmomp_', characteristic, '.csv'))
}
```

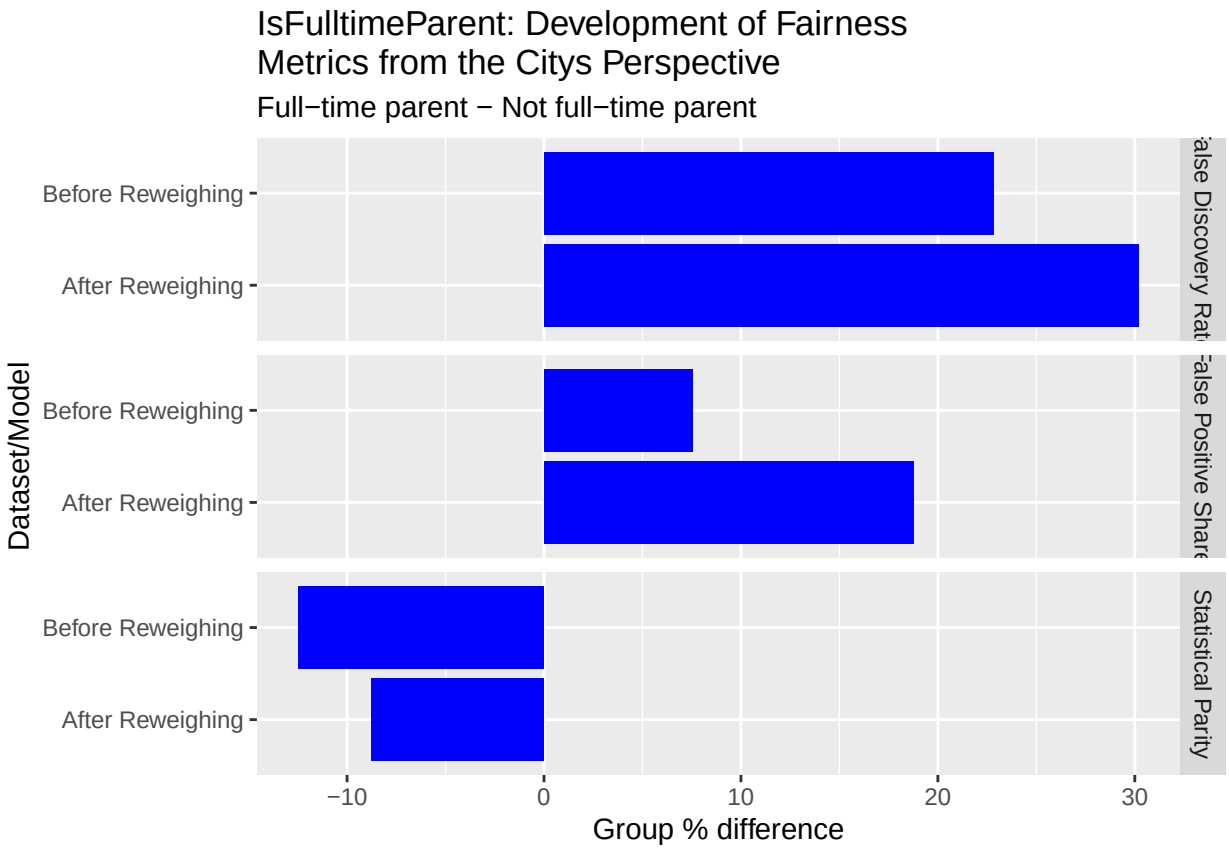



Dutch: Development of Fairness
Metrics from the Citys Perspective
Not Dutch − Dutch
Dataset/Model
Before Reweighing
After Reweighing
Before Reweighing
After Reweighing
Before Reweighing
After Reweighing
alse Discovery Rat
alse Positive Share
Statistical Parity
−25
0
25
50
Group % difference


gender: Development of Fairness Metrics from the Citys Perspective
Value
FDR
FP
STAT_PAR
40
30
20
10
0
15
10
5
0
Before Reweighing
After Reweighing
Dataset/Model
gender
F
M


gender: Development of Fairness Metrics from the Citys Perspective
F − M
Dataset/Model
Before Reweighing
After Reweighing
Before Reweighing
After Reweighing
Before Reweighing
After Reweighing
alse Discovery Rat
alse Positive Share
Statistical Parity
−20
−15
−10
−5
0
Group % difference


IsFulltimeParent: Development of Fairness Metrics from the Citys Perspectiv
Value
FDR
FP
STAT_PAR
40
30
20
10
0
15
10
5
0
Before Reweighing
After Reweighing
Dataset/Model
IsFulltimeParent
Full−time parent
Not full−time parent



RQ2 Error rate in Pilot was much higher than in tests

```
# subset to relevant characteristics and metrics
cms_error <- cms_long %>%
  filter(stage %in% c( 'Prepilot/AR', 'Pilot/AR'),
         Metric %in% c('FP', 'FN', 'ERROR'),
         Feature_EN == 'Dutch') %>%
  mutate(order = case_when(stage == 'Prepilot/AR' ~ 1,
                           stage == 'Pilot/AR' ~ 1,
                           .default = NA))

# produce plots for error rate for Dutch vs non-Dutch applicants
p4 <- ggplot(cms_error, aes(x = reorder(stage, order), y = Value, fill = Metric))+
  geom_bar(stat = 'identity', position = position_dodge())+
  facet_wrap(.~Feature_Value_EN)+
 labs(x = 'Dataset/Model', y = '% Error',
       title = paste0(characteristic, ': Development of Error rates across model development'),
       fill = 'Error Metric')+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
print(p4)
```

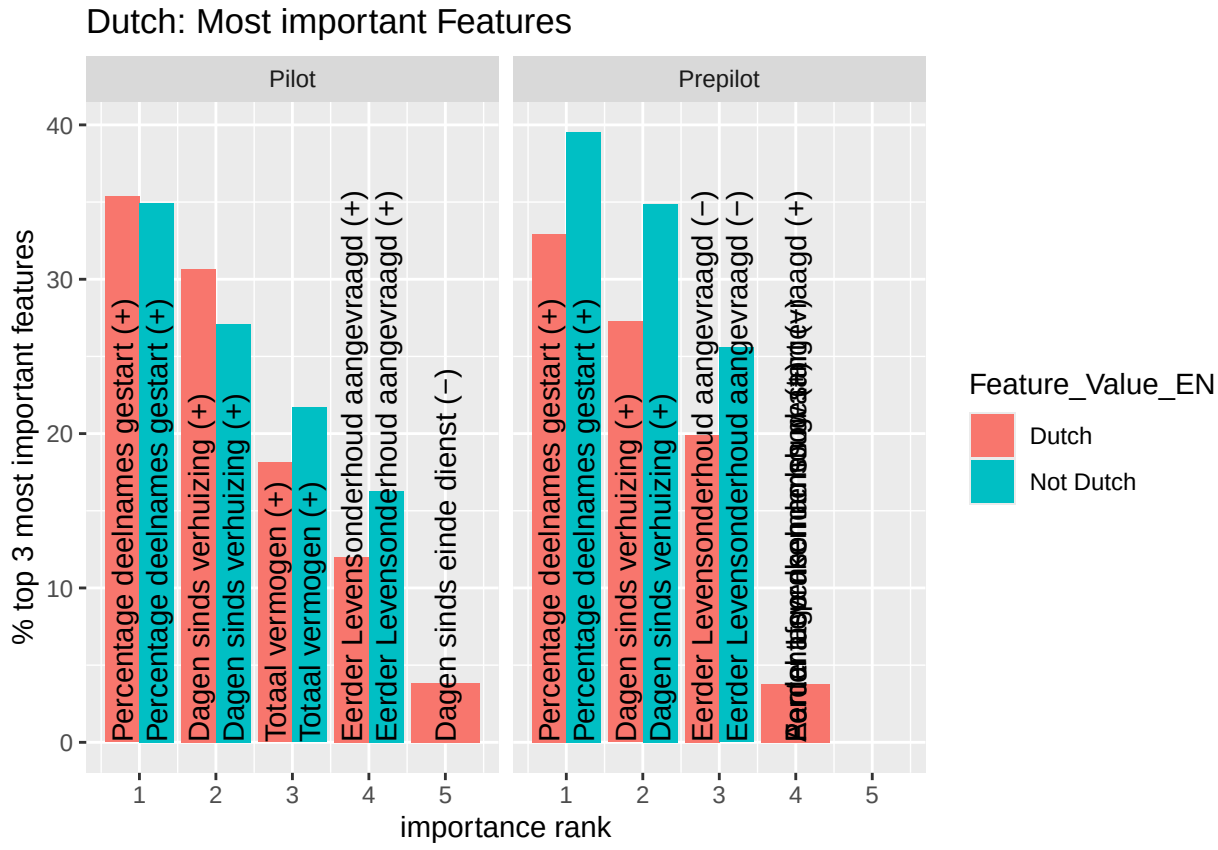

```
ggsave(paste0('../output/rq3_p4_error_', characteristic, '.png'), plot = p4, width = 10, height = 8)

cms_error <- cms_error %>%
  mutate(Metric = case_when(Metric == 'FP' ~ 'False Positives',
                            Metric == 'FN' ~ 'False Negatives',
                            Metric == 'ERROR' ~ 'Overall Error Rate'))
write.csv(cms_char, paste0('../output/', characteristic ,'_error_rate.csv'))
```

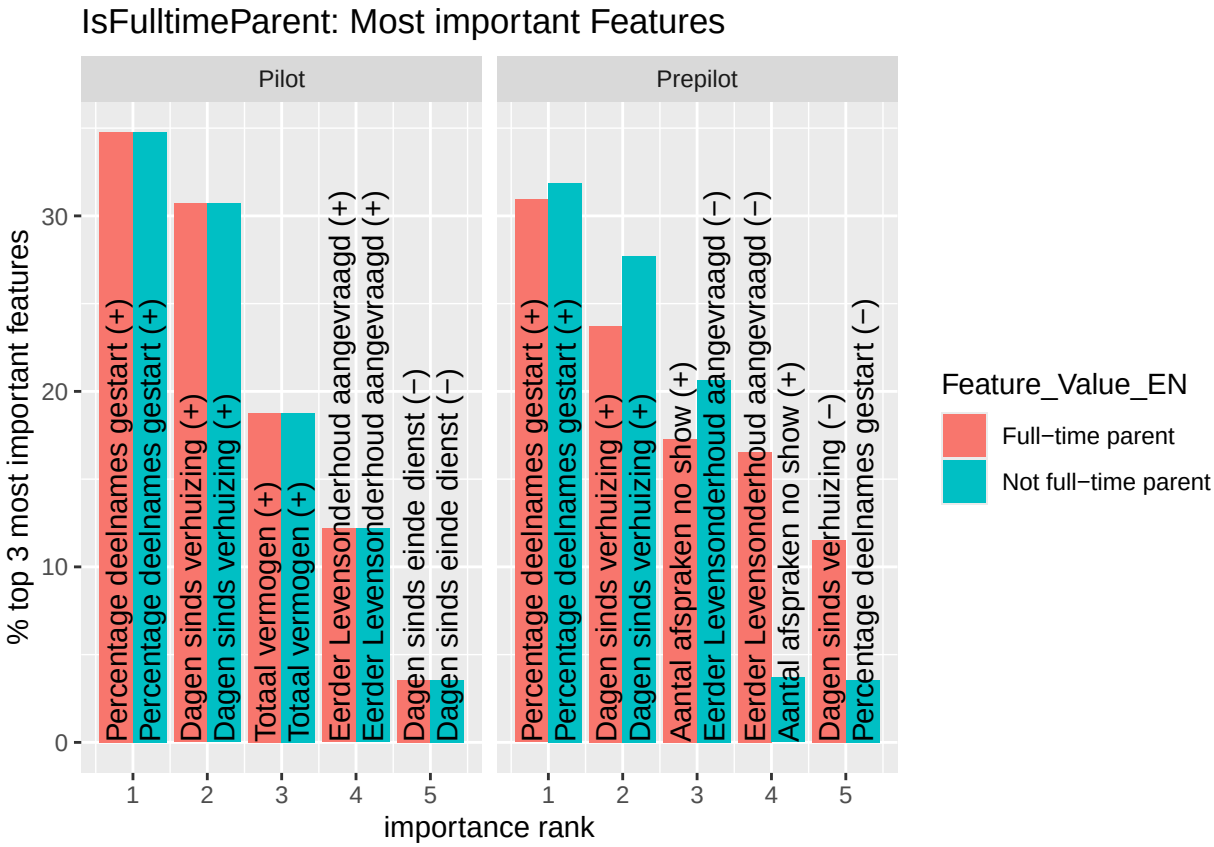
RQ 3 Feature importance

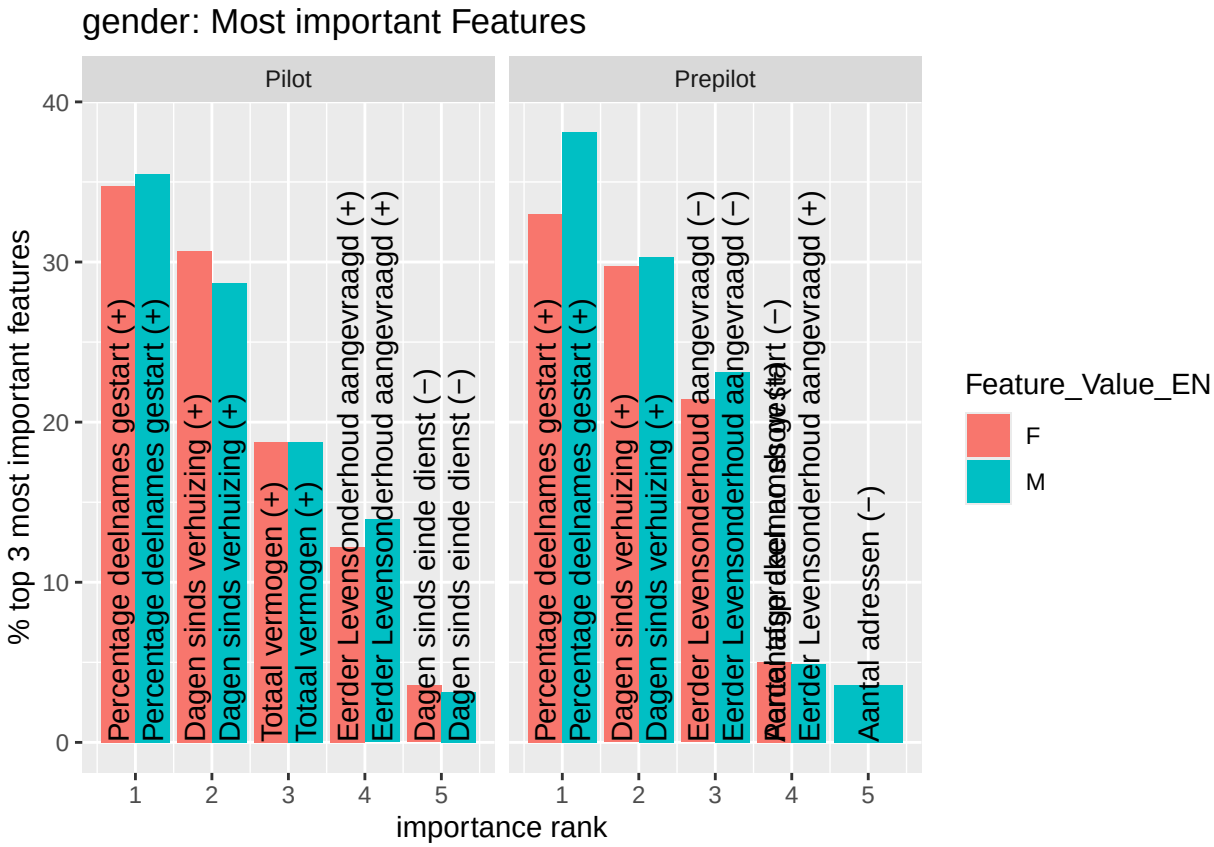
Along with the classification, the model provided caseworkers with the three most important features used by the model to come to its determination. Loek provided us access to the most important feature by demographic group. This allows us 1) to see if caseworkers could deduce beneficiary charactsitics from the highlighted features, potentially activating their biases, and 2) whether the model used different features for different demographic groups in coming to its determination. The latter could be concerning under due process considerations.

```
feature_counts_restricted <- feature_counts %>%
  filter(Feature_EN %in% c("gender", "IsFulltimeParent", "Dutch")) %>%
  group_by(Feature_EN, Feature_Value_EN, dataset) %>%
  arrange(desc(share)) %>%
  mutate(rank = dense_rank(desc(share))) %>%
  slice_max(n = 5, order_by = share) %>%
  ungroup() %>%
```

```r
  filter(!is.na(share))

# plot top 5 most important feature by applicant characteristic and development stage
for(characteristic in unique(feature_counts_restricted$Feature_EN)){
  feature_counts_char <- feature_counts_restricted %>%
    filter(Feature_EN == characteristic)

  p3 <- ggplot(feature_counts_char, aes(x = rank, y = share, fill = Feature_Value_EN, label = Important
    geom_bar(stat = 'identity', position = position_dodge())+
    geom_text(aes(y = 0), hjust = 0, angle = 90, position = position_dodge(width = .9))+
    facet_grid(.~dataset)+
    labs(x = 'importance rank',
         y = '% top 3 most important features',
         title = paste0(characteristic, ': Most important Features'))

  print(p3)
  ggsave(paste0('../output/rq3_feature_importance_', characteristic, '.png'), plot = p3, width = 10, he

}
```




IsFulltimeParent: Most important Features
Pilot
Prepilot
% top 3 most important features
0
10
20
30
1
2
3
4
5
importance rank
Percentage deelnames gestart (+)
Percentage deelnames gestart (+)
Dagen sinds verhuizing (+)
Dagen sinds verhuizing (+)
Totaal vermogen (+)
Totaal vermogen (+)
Eerder Levensonderhoud aangevraagd (+)
Eerder Levensonderhoud aangevraagd (+)
Dagen sinds einde dienst (−)
Dagen sinds einde dienst (−)
Percentage deelnames gestart (+)
Percentage deelnames gestart (+)
Dagen sinds verhuizing (+)
Dagen sinds verhuizing (+)
Aantal afspraken no show (+)
Eerder Levensonderhoud aangevraagd (−)
Eerder Levensonderhoud aangevraagd (−)
Aantal afspraken no show (+)
Dagen sinds verhuizing (−)
Percentage deelnames gestart (−)
Feature_Value_EN
Full−time parent
Not full−time parent


gender: Most important Features
Pilot
Prepilot
% top 3 most important features
importance rank
Feature_Value_EN
F
M
Percentage deelnames gestart (+)
Dagen sinds verhuizing (+)
Totaal vermogen (+)
Eerder Levensonderhoud aangevraagd (+)
Dagen sinds einde dienst (−)
Aantal adressen (−)