Assignment 2

Enric Reverter & Gerard Pons

14/10/2021

In this assignment, a binary model has been created to predict whether or not a candidate will work for a company, or in other words, if he or she will change jobs. The process of creating the model started from understanding and cleaning the data, which in this dataset was a challenging but very important task. Then it continued by progressively building the model, first by checking the best way to treat some variables (i.e as factor or numerical), assessing transformations, additions and interactions. After that, the residuals and influential observations were addressed, and the model was reevaluated. Finally, the predictive power of the model was assessed. The steps hereunder document this process in a more detailed way.

Required libraries

```
## Data manipulation
library(tidyverse)
library(dplyr)
options(dplyr.summarise.inform = FALSE)
library(mice)
library(Hmisc)
## Statistics
library(lsr)
library(missMDA)
library(VIM)
library(chemometrics)
library(arules)
library(skimr)
library(car)
library(FactoMineR)
library(factoextra)
library(effects)
## Plots
library(ggplot2)
library(ggExtra)
library(ggthemes)
library(processx)
library(plotly)
library(cowplot)
library(gridExtra)
library(RColorBrewer)
theme_set(theme_bw())
## Set data path
setwd("..")
```

```
data_path = file.path(getwd(), "data")
plot_path = file.path(getwd(), "plots")
```

Data Exploration

Sample from the original dataset:

```
data = read.csv(file.path(data_path, "aug_train.csv"))
set.seed(020198)
sample = sample(1:nrow(data), 5000)
df = data[sample,]
write.csv(df, file.path(data_path, "jobs.csv"), row.names = FALSE)
```

Or load the dataset in case it is already stored:

```
df = read.csv(file.path(data_path, "jobs.csv"))
```

Skim over it:

```
head(df)
summary(df)
str(df)
```

Convert data types to the proper format:

```
df = df %>%
  mutate(across(where(is.character), ~ na_if(., ""))) %>%
  mutate(across(where(is.character) | matches("target"), ~ as.factor(.)))
```

Detail of factors:

```
df %>%
  select(., where(is.factor)) %>%
  sapply(., table)
table(df$last_new_job)
```

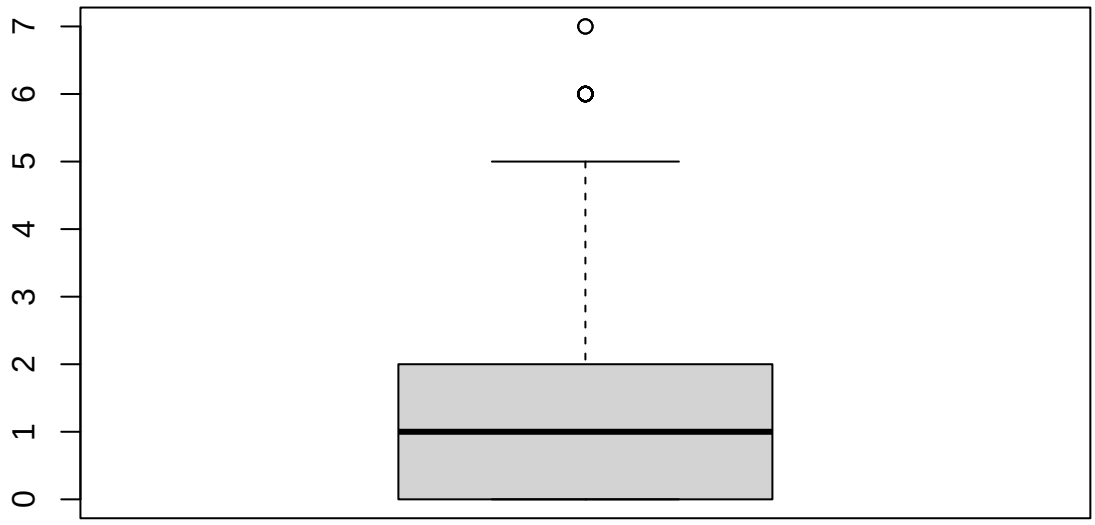
Missing Values

As it can be observed, the dataset contains a lot of missing values, in some cases even exceeding the 30% of values in a given attribute. These missing values might condition the imputation methods, which is first done using logic. Then, algorithms are used. Also, there is a set of 21 observations with more than 50% of the variables (that will be used) as NA, which have been decided to be deleted from the working set:

```
count_na = function(x) {sum(is.na(x))}
df = df %>%
  mutate(across(matches("company"), ~ as.character(.))) %>%
  mutate(across(matches("company"), ~ na_if(., "NA"))) %>%
  mutate(across(matches("company"), ~ as.factor(.))) %>%
  mutate(count_na = apply(., 1, count_na))
summary(df$count_na)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   1.088   2.000   7.000
```

```
boxplot(df$count_na)
```




```
table(df$count_na)
```

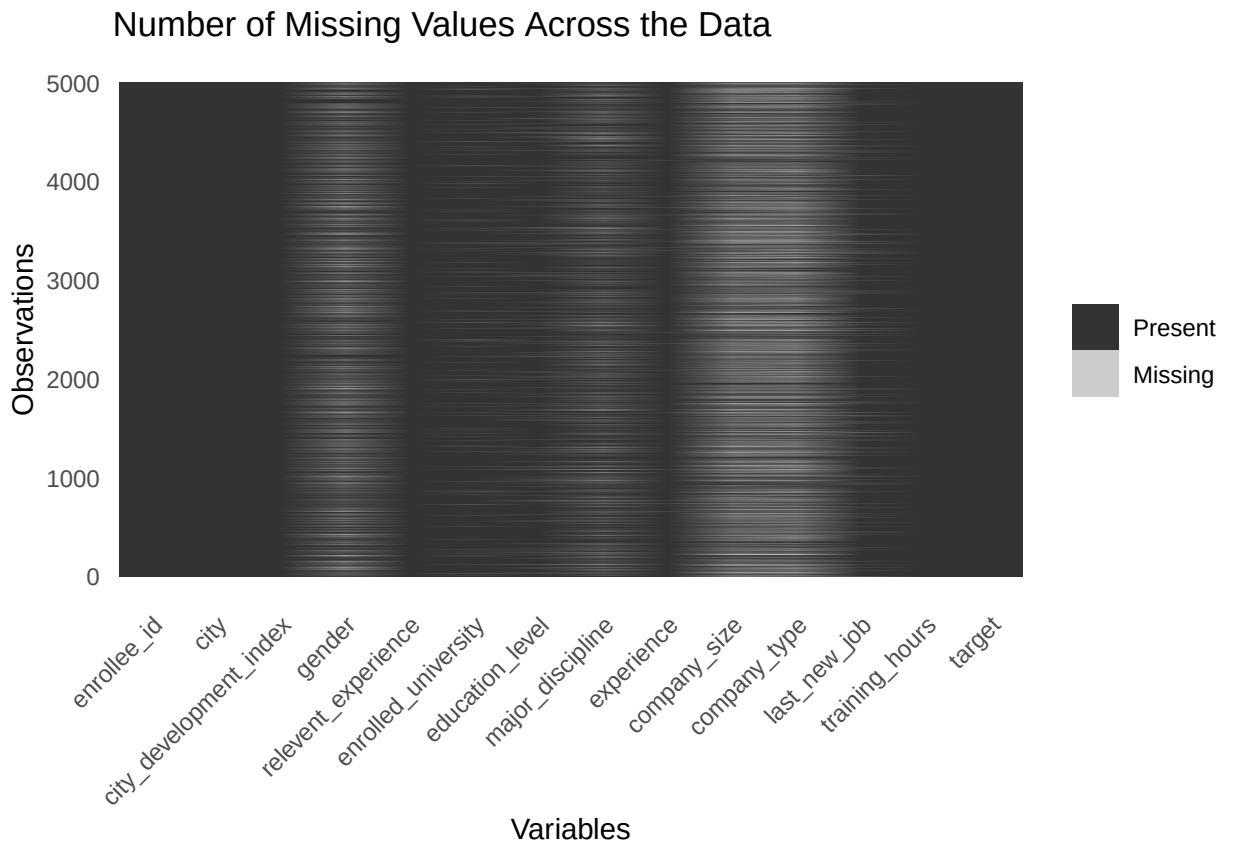
```
##
##    0    1    2    3    4    5    6    7
## 2344  932  991  495  171   46   19    2
```

Visualizing the missing values prior to dropping them:

```
library(reshape2)
```

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

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

```
ggplot_missing <- function(data){
  df2 <- data %>% is.na %>% melt

  ggplot(df2, aes(Var2, Var1, fill=value)) +
    geom_raster() +
    theme_minimal() +
    theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
    scale_fill_grey(name="", labels=c("Present", "Missing")) +
    theme(axis.text.x  = element_text(angle=45, vjust=1, hjust=1)) +
    labs(title = "Number of Missing Values Across the Data",
         x = "Variables",
         y = "Observations")
}

ggplot_missing(select(df, -c("count_na")))
```



Deleting the observations with many NA's:

```
df = df %>%
  filter(., count_na < 5) %>%
  select(., -c("count_na"))
```

The rules used for logically imputation are stated as follow, always assuming that everyone in the dataset is currently working, as the target is looking or not for a job change:

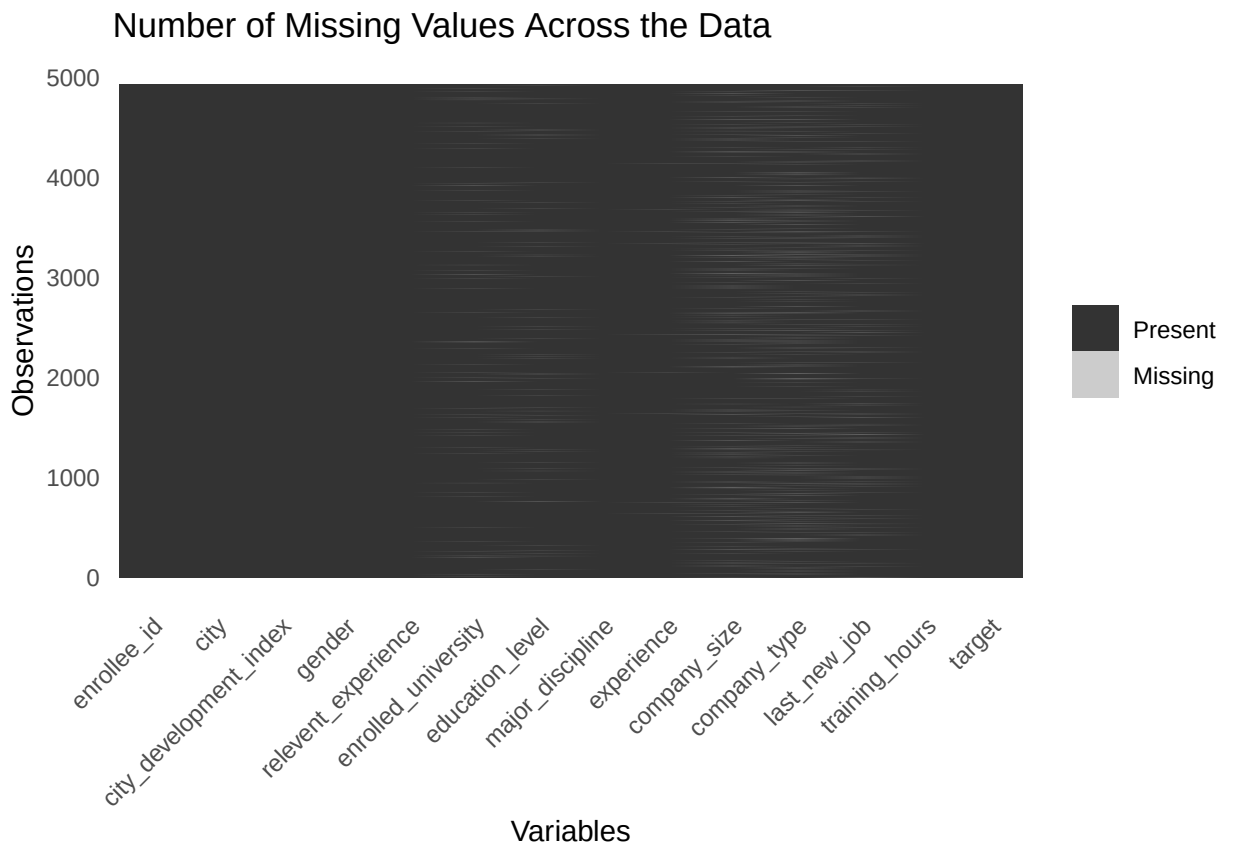
- If the education_level is null but they are enrolled in a university, the education is set to high school.

- If major_discipline is not null, the education level should be at least graduate.
- If company_type is known and company_size is missing, it is left for imputation and vice versa. If both are missing they are labeled as Unknown, as the number of missing values for company information exceeds 30%.
- If gender is missing, it is imputed with Unknown, as there are nearly 30% of missing values in gender.
- If major_discipline is null, it is imputed with Other if the education level is Graduate, Masters or PhD, and imputed to No Major otherwise.
- If experience , last_New_Job and company_ information are null, the experience is imputed to <1.

```
df = df %>%
  mutate(f.enrolled = case_when(enrolled_university == "no_enrollment" ~ "No",
                                !is.na(enrolled_university) ~ "Yes"))
df = df %>%
  # Convert factors to strings in order to impute them
  mutate(across(where(is.factor), ~ as.character(.))) %>%

  # Impute education level as mentioned above
  mutate(education_level = case_when(is.na(education_level) & f.enrolled == "Yes" ~ "High School",
                                     !is.na(major_discipline) & !(education_level %in% c("Graduate", "Ma
                                     TRUE ~ education_level)) %>%

  # Impute major_discipline as mentioned above
  mutate(major_discipline = case_when(is.na(major_discipline) & !(education_level %in% c("Graduate", "Ma
                                      is.na(major_discipline) & education_level %in% c("Graduate", "Mas
                                      TRUE ~ major_discipline)) %>%

  # Impute enrolled_university
  mutate(enrolled_university = case_when(is.na(enrolled_university) & education_level %in% c("Masters",
                                         TRUE ~ enrolled_university)) %>%

  # Impute experience
  mutate(experience = case_when(is.na(experience) & (is.na(last_new_job) & is.na(company_size) & is.na(
                                TRUE ~ experience)) %>%

  # Impute gender
  mutate(gender = case_when(is.na(gender) ~ "Other",
                            TRUE ~ gender)) %>%

  # Impute company
  mutate(company_size = case_when(is.na(company_size) & is.na(company_type) ~ "Unknown",
                                  TRUE ~ company_size)) %>%
  mutate(company_type = case_when(is.na(company_type) & company_size == "Unknown" ~ "Other",
                                  TRUE ~ company_type)) %>%

  # Convert back to factors
  mutate(across(where(is.character), ~ as.factor(.))) %>%

  # Drop unused columns
  select(., -c("f.enrolled"))
```

Visualizing the missing values posterior to the logical imputation:

```
ggplot_missing(df)
```



After the logical imputation, the NA values do not account for more than 2% in any of the categories, and it has been decided to impute them with factorial analysis for mixed data. It must be noted that a new flag attribute 'Imputed' has been created, in order to keep track of these imputed observations when modelling, as they could cause problems.

Indicator of rows which still have NA's:

```
colSums(is.na(df))
```

```
##            enrollee_id                   city city_development_index
##                      0                      0                      0
##                 gender    relevent_experience    enrolled_university
##                      0                      0                     56
##        education_level       major_discipline             experience
##                     51                      0                      9
##           company_size           company_type           last_new_job
##                    151                    197                    101
##         training_hours                 target
##                      0                      0
```

```
imputed_indicator = function(x) {if(count_na(x)>0) {return(TRUE)} else {return(FALSE)}}

df = df %>%
  mutate(imputed = apply(., 1, imputed_indicator))
```

FAMD Imputation

Impute with FAMD method:

```
res.famd = imputeFAMD(select(df, -c("target", "city", "enrollee_id", "imputed")))
```

As it can be seen, the class frequencies after imputation have been compared to the ones before it, and there is no notable change.

```
round(prop.table(table(df$education_level))*100,1)
```

```
##
##       Graduate    High School        Masters            Phd Primary School
##           62.6           11.3           22.6            2.0            1.6
```

```
round(prop.table(table(res.famd$completeObs$education_level))*100,1)
```

```
##
##       Graduate    High School        Masters            Phd Primary School
##           62.7           11.5           22.3            1.9            1.5
```

```
round(prop.table(table(df$last_new_job))*100,1)
```

```
##
##    >4     1     2     3     4 never
##  17.1  42.7  16.1   5.8   4.9  13.3
```

```
round(prop.table(table(res.famd$completeObs$last_new_job))*100,1)
```

```
##
##    last_new_job_>4     last_new_job_1     last_new_job_2     last_new_job_3
##               16.8               43.8               15.8                5.7
##     last_new_job_4 last_new_job_never
##                4.8               13.1
```

```
round(prop.table(table(df$enrolled_university))*100,1)
```

```
##
## Full time course    no_enrollment Part time course
##             19.9             73.4              6.7
```

```
round(prop.table(table(res.famd$completeObs$enrolled_university))*100,1)
```

```
##
## Full time course    no_enrollment Part time course
##             19.9             73.5              6.6
```

```
summary(df$training_hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   23.00   46.00   64.14   86.00  336.00
```

Store the complete dataset:

```
df = data.frame(res.famd$completeObs, select(df, c("target", "city", "enrollee_id", "imputed")))
```

Mutate strings after FAMD converted them into dummy variables:

```
df = df %>%
  mutate(across(where(is.factor), ~ as.character(.))) %>%
  mutate(gender = str_remove(gender, "gender_")) %>%
  mutate(major_discipline = str_remove(major_discipline, "major_discipline_")) %>%
  mutate(company_type = str_remove(company_type, "company_type_")) %>%
  mutate(experience = str_remove(experience, "experience_")) %>%
  mutate(last_new_job = str_remove(last_new_job, "last_new_job_")) %>%
  mutate(across(where(is.character), ~ as.factor(.)))
```

With the complete dataset, some new attributes have been crated: a new numerical variable has been created from the factor experience and a new factor has been created from the variable of city development index. In future steps, it will be decided which one is the most suitable for the modelling process. It must be noted that since company size had a lot of NA's, it has not been converted into numerical.

Convert experience into a numerical variable:

```
df = df %>%
  mutate(across(where(is.factor), ~ as.character(.))) %>%
  mutate(n.experience = case_when(experience == "<1" ~ "0",
                                  experience == ">20" ~ "25",
                                  TRUE ~ experience)) %>%
  mutate(n.experience = as.integer(n.experience)) %>%
  mutate(across(where(is.character), ~ as.factor(.)))

summary(df$n.experience)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     4.0     8.0    10.7    16.0    25.0
```

```
summary(df$experience)
```

```
##  <1 >20   1  10  11  12  13  14  15  16  17  18  19   2  20   3   4   5   6   7
## 131 831 135 240 176 113  92 153 157 146  87  65  90 291  37 344 373 394 316 270
##   8   9
## 229 263
```

Convert city development index into a categorical variable:

```
groups = 5

df$f.city_development_index = as.ordered(cut2(df$city_development_index, g=groups, m=nrow(df)/groups))
table(df$f.city_development_index)
```

```
##
## [0.448,0.691) [0.691,0.878) [0.878,0.920)         0.920 [0.921,0.949]
##          1004           970          1000          1294           665
```

Write the dataset:

```
write.csv(df, file.path(data_path, "jobs_complete.csv"), row.names = FALSE)
```

Outlier treatment

Univariate outliers can not be seen in the dataset for the two numerical variables. One could think that training_hours contains some outliers, as they are above the extreme threshold. However, they are not too extreme and all of them have a very plausible value, hence imputation would not be a good practice in this case.

```
extreme_out = quantile(df$training_hours)[[4]]+3*IQR(df$training_hours)

ggplot(data = df, aes(x="", y=training_hours)) +
  geom_boxplot(width=0.5) +
  geom_hline(yintercept = extreme_out, color="red") +
  scale_y_continuous(labels=scales::comma)
labs(title='Boxplot Training Hours',
     y="Training Hours") +
  # Do not show x axis
  theme(axis.text.x=element_blank(), axis.ticks.x = element_blank(), axis.line.x = element_blank(), axi

num_outliers = df %>%
  filter(., training_hours > extreme_out) %>%
  nrow()
num_outliers

outliers = df %>%
  filter(., training_hours > extreme_out)

# prop.table(table(df$gender))
# prop.table(table(outliers$gender))
# prop.table(table(df$relevent_experience))
# prop.table(table(outliers$relevent_experience))
# prop.table(table(df$enrolled_university))
# prop.table(table(outliers$enrolled_university))
# prop.table(table(df$education_level))
# prop.table(table(outliers$education_level))
# prop.table(table(df$major_discipline))
# prop.table(table(outliers$major_discipline))
# prop.table(table(df$last_new_job))
# prop.table(table(outliers$last_new_job))
```

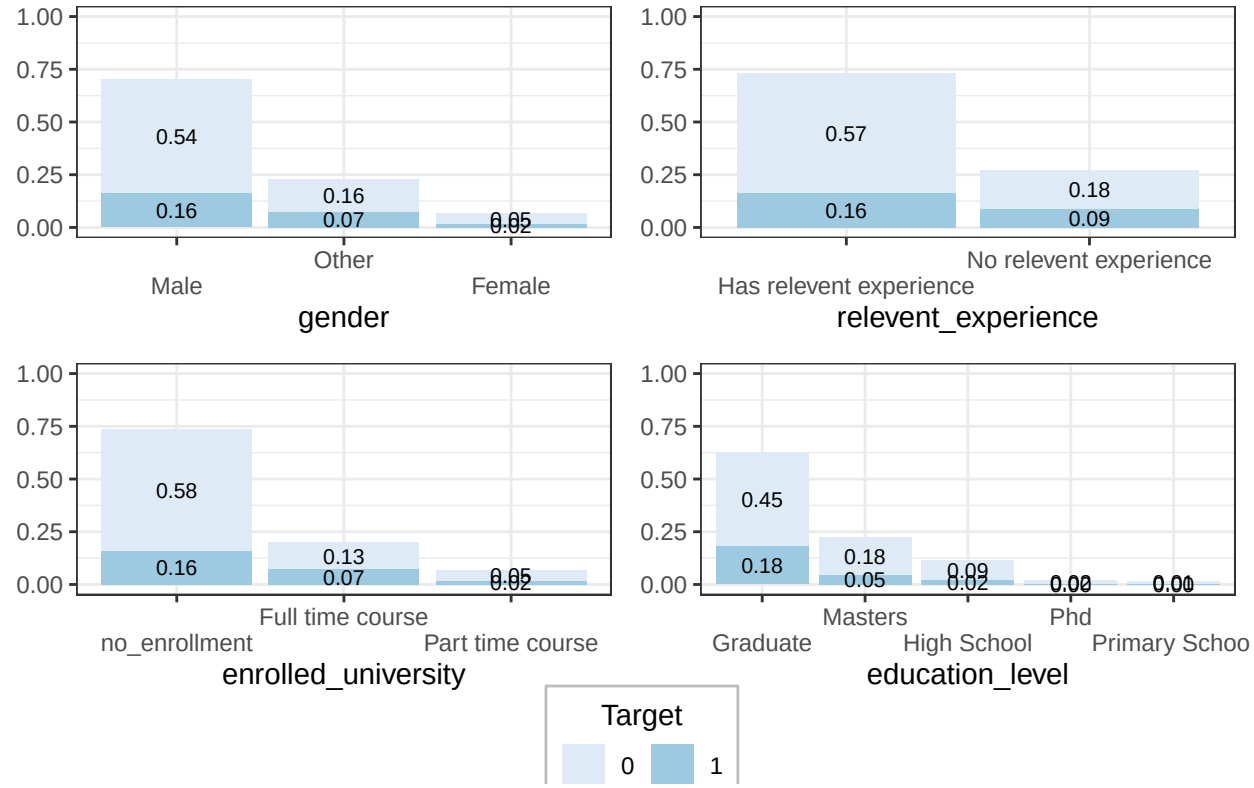
Factor Visualizations

Let's take a look at the categorical variables with which the modelling is done.

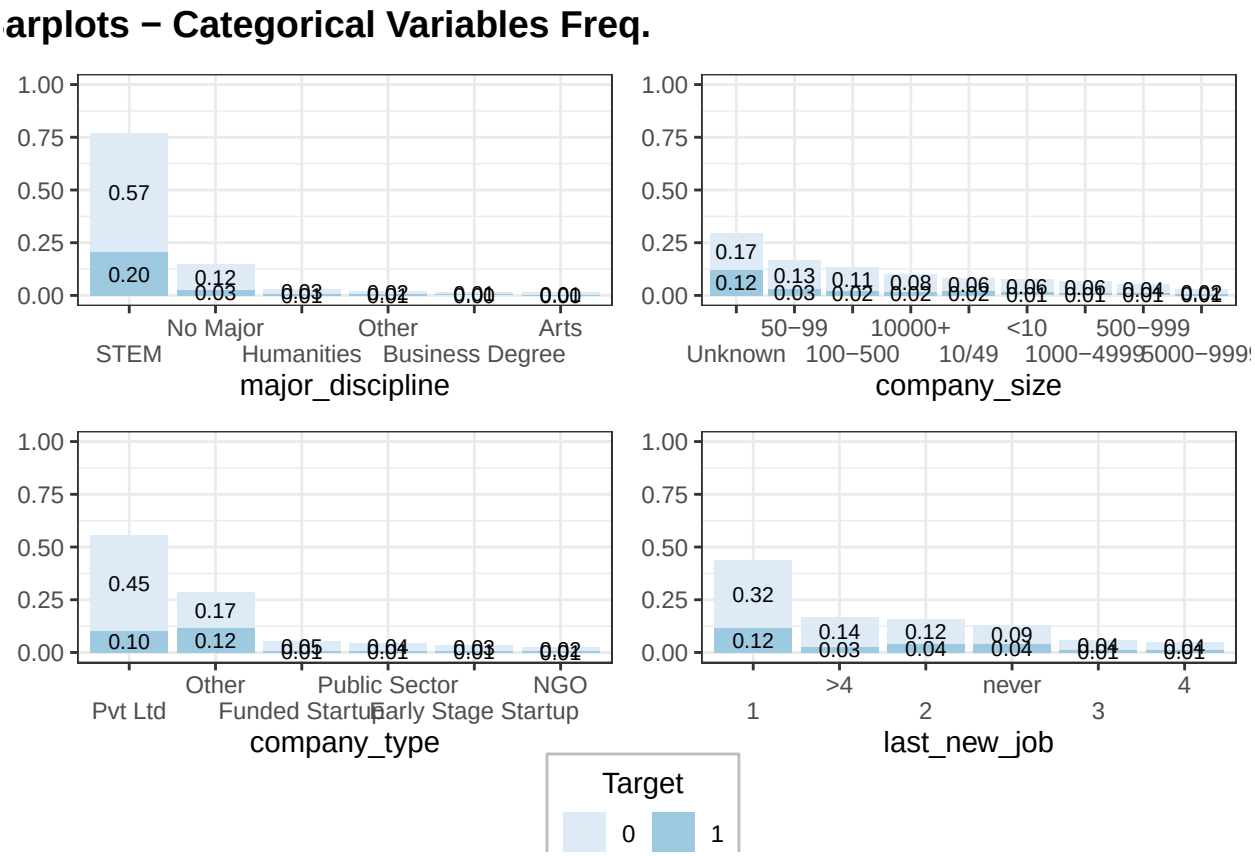
```
plist = list()
cat_vars = c("gender", "relevent_experience", "enrolled_university", "education_level", "major_discipli
```

```
for (i in 1:(length(cat_vars))) {
  plist[[i]] = df %>%
    group_by(!!as.name(cat_vars[i]), target) %>%
    summarise(n = n()/nrow(df)) %>%
    ggplot(data=., aes(x=reorder(!!as.name(cat_vars[i]), -n), y=n, fill=target)) +
    geom_bar(position="stack", stat="identity") +
    scale_fill_brewer(palette = "Blues") +
    scale_y_continuous(limits=c(0, 1)) +
    geom_text(aes(label=sprintf("%0.2f", round(n, digits = 2))), position = position_stack(vjust = 0.5)
    labs(x=cat_vars[i],
           fill="Target") +
    theme(legend.position="none") +
    theme(axis.title.y=element_blank()) +
    scale_x_discrete(labels = function(labels) {
      fixedLabels = c()
      for (l in 1:length(labels)) {
        fixedLabels[l] = paste0(ifelse(l %% 2 == 0, '', '\n'), labels[l])
      }
      return(fixedLabels)
    })
}

p = df %>%
    group_by(!!as.name(cat_vars[i]), target) %>%
    summarise(n = n()) %>%
    ggplot(data=., aes(x=reorder(!!as.name(cat_vars[i]), -n), y=n, fill=target)) +
    geom_bar(position="stack", stat="identity") +
    scale_fill_brewer(palette = "Blues") +
    labs(x=cat_vars[i], fill="Target") +
    guides(fill = guide_legend(nrow = 1))

legend = get_legend(p + theme(legend.box.margin = margin(0, 0, 0, 12),
                              legend.box = "horizontal",
                              legend.title.align=0.5,
                              legend.background = element_rect(linetype="solid",
                                                               color="grey")))

title = ggdraw() + draw_label("Barplots - Categorical Variables Freq.", fontface='bold')
empty = ggdraw()
p = plot_grid(title, empty, plotlist = plist[1:4], ncol = 2, rel_heights = c(0.2,1,1))
q = plot_grid(title, empty, plotlist = plist[5:8], ncol = 2, rel_heights = c(0.2,1,1))

pp = plot_grid(p, legend, ncol = 1, rel_heights = c(1, 0.1))
qq = plot_grid(q, legend, ncol = 1, rel_heights = c(1, 0.1))

pp
```

arplots − Categorical Variables Freq.




```
qq
```




```
# ggsave(file=file.path(plot_path,"barplot_freq1.png"), plot=pp)
# ggsave(file=file.path(plot_path,"barplot_freq2.png"), plot=qq)
```

Modelization

Before starting with the model, it is interesting to describe the response variable. It can be seen that it is significantly associated with all the numerical and categorical variables, except for training_hours, which sits really close to the threshold. It is also worth noting that the variables which have been kept in purpose both in numerical and categorical form are the ones that have a more significant association, meaning that the future assessment of how to treat them will be of particular interest. Overall, what can be said is that in general, people who want to change jobs tend to be from less developed cities with no data regarding the company, have less experience, and a higher education.

```
cat = FactoMineR::catdes(df[,-c(13:15)], 12)
cat$test.chi2
```

```
##                                p.value df
## f.city_development_index 3.975953e-152  4
## company_size              5.157409e-61  8
## company_type              7.792823e-60  5
## experience                2.796508e-30 21
## enrolled_university       5.728866e-19  2
## education_level           3.535428e-14  4
```

```
## relevent_experience       2.341751e-13  1
## last_new_job              2.062908e-08  5
## gender                    8.397510e-08  2
## major_discipline          3.170616e-05  5
```

```
cat$quanti.var
```

```
##                               Eta2       P-value
## city_development_index 0.1274969888 2.903870e-148
## n.experience           0.0305698714  3.685682e-35
## training_hours         0.0008362456  4.225817e-02
```

```
cat$category
```

```
## $`0`
##                                                Cla/Mod   Mod/Cla    Global
## company_type=Pvt Ltd                          81.50585 60.237709 55.463207
## f.city_development_index=[0.878,0.920)        88.10000 23.797947 20.271640
## f.city_development_index=[0.921,0.949]        89.92481 16.153431 13.480641
## enrolled_university=no_enrollment             78.21891 76.634252 73.525238
## experience=>20                                84.95788 19.070773 16.845733
## relevent_experience=Has relevent experience 77.79012 75.688817 73.018447
## last_new_job=>4                               83.47407 18.692599 16.805190
## company_size=100-500                          83.84146 14.856834 13.298196
## company_size=1000-4999                        86.48649  7.779579  6.750456
## gender=Male                                   77.15108 72.420313 70.443949
## company_size=50-99                            81.85185 17.909238 16.420028
## company_type=Funded Startup                   86.79245  6.212858  5.371985
## major_discipline=No Major                     81.85596 15.964344 14.636124
## education_level=High School                   82.65487 12.614803 11.453477
## education_level=Masters                       79.76407 23.743922 22.339347
## f.city_development_index=0.920                79.05719 27.633712 26.231502
## company_size=10000+                           82.03593 11.102107 10.156092
## experience=16                                 87.67123  3.457590  2.959659
## company_size=500-999                          82.75862  5.834684  5.290898
## company_size=<10                              81.38298  8.265802  7.622137
## education_level=Phd                           87.50000  2.269044  1.946077
## experience=10                                 82.91667  5.375473  4.865194
## education_level=Primary School                88.15789  1.809833  1.540645
## f.city_development_index=[0.691,0.878)        78.35052 20.529444 19.663491
## experience=17                                 86.20690  2.025932  1.763633
## company_type=Public Sector                    81.14035  4.997299  4.621934
## experience=14                                 82.35294  3.403566  3.101561
## experience=15                                 82.16561  3.484603  3.182647
## major_discipline=Humanities                   82.05128  3.457590  3.162376
## experience=6                                  70.25316  5.996759  6.405838
## experience=5                                  69.79695  7.428417  7.987026
## last_new_job=never                            70.89783 12.371691 13.095479
## last_new_job=1                                73.05556 42.625608 43.786742
## experience=2                                  67.01031  5.267423  5.899047
## experience=3                                  65.40698  6.077796  6.973444
## experience=1                                  57.03704  2.079957  2.736671
## experience=4                                  64.61126  6.509995  7.561322
```

```
## experience=<1                                   56.48855  1.998920  2.655585
## major_discipline=STEM                           73.49619 75.580767 77.174133
## gender=Other                                    68.71705 21.123717 23.069126
## relevent_experience=No relevent experience      67.61833 24.311183 26.981553
## education_level=Graduate                        71.26697 59.562399 62.720454
## enrolled_university=Full time course            64.01631 16.963803 19.886479
## company_type=Other                              58.69721 22.150189 28.319481
## company_size=Unknown                            58.85989 23.149649 29.515508
## f.city_development_index=[0.448,0.691)          43.82470 11.885467 20.352727
##                                                       p.value     v.test
## company_type=Pvt Ltd                           1.696974e-31  11.675683
## f.city_development_index=[0.878,0.920)         8.949731e-30  11.333555
## f.city_development_index=[0.921,0.949]         6.337961e-25  10.310184
## enrolled_university=no_enrollment              4.181199e-17   8.407703
## experience=>20                                 4.175352e-14   7.555412
## relevent_experience=Has relevent experience    6.044670e-13   7.199471
## last_new_job=>4                                1.977683e-10   6.363064
## company_size=100-500                           6.384243e-09   5.806368
## company_size=1000-4999                         1.261415e-07   5.284365
## gender=Male                                    1.836673e-07   5.215151
## company_size=50-99                             5.006137e-07   5.026077
## company_type=Funded Startup                    1.482631e-06   4.813563
## major_discipline=No Major                      2.593774e-06   4.700616
## education_level=High School                    4.551971e-06   4.584444
## education_level=Masters                        3.056277e-05   4.169232
## f.city_development_index=0.920                 8.626808e-05   3.926285
## company_size=10000+                            8.706940e-05   3.924059
## experience=16                                  1.421461e-04   3.804404
## company_size=500-999                           2.274236e-03   3.051866
## company_size=<10                               2.470553e-03   3.026925
## education_level=Phd                            2.632041e-03   3.007733
## experience=10                                  2.869768e-03   2.981354
## education_level=Primary School                 4.879655e-03   2.814873
## f.city_development_index=[0.691,0.878)         7.387903e-03   2.678834
## experience=17                                  1.128273e-02   2.533818
## company_type=Public Sector                     2.631339e-02   2.221555
## experience=14                                  2.976693e-02   2.173178
## experience=15                                  3.204593e-02   2.143837
## major_discipline=Humanities                    3.566666e-02   2.100707
## experience=6                                   4.517093e-02  -2.003059
## experience=5                                   1.369071e-02  -2.465278
## last_new_job=never                             9.927028e-03  -2.578361
## last_new_job=1                                 4.441380e-03  -2.844983
## experience=2                                   1.503261e-03  -3.174053
## experience=3                                   3.430779e-05  -4.142809
## experience=1                                   3.792657e-06  -4.622433
## experience=4                                   2.943942e-06  -4.674693
## experience=<1                                  2.819389e-06  -4.683557
## major_discipline=STEM                          2.561324e-06  -4.703186
## gender=Other                                   3.236147e-08  -5.528156
## relevent_experience=No relevent experience     6.044670e-13  -7.199471
## education_level=Graduate                       7.189505e-16  -8.067257
## enrolled_university=Full time course           4.870655e-18  -8.656361
## company_type=Other                             5.171400e-59 -16.198450
```

```
## company_size=Unknown                            1.783083e-61 -16.543515
## f.city_development_index=[0.448,0.691)      1.426886e-129 -24.218315
##
## $`1`
##                                                 Cla/Mod    Mod/Cla    Global
## f.city_development_index=[0.448,0.691)      56.17530 45.8164094 20.352727
## company_size=Unknown                        41.14011 48.6596263 29.515508
## company_type=Other                          41.30279 46.8724614 28.319481
## enrolled_university=Full time course        35.98369 28.6758733 19.886479
## education_level=Graduate                    28.73303 72.2177092 62.720454
## relevent_experience=No relevent experience  32.38167 35.0121852 26.981553
## gender=Other                                31.28295 28.9195776 23.069126
## major_discipline=STEM                       26.50381 81.9658814 77.174133
## experience=<1                               43.51145  4.6303818  2.655585
## experience=4                                35.38874 10.7229894  7.561322
## experience=1                                42.96296  4.7116166  2.736671
## experience=3                                34.59302  9.6669374  6.973444
## experience=2                                32.98969  7.7985378  5.899047
## last_new_job=1                              26.94444 47.2786353 43.786742
## last_new_job=never                          29.10217 15.2721365 13.095479
## experience=5                                30.20305  9.6669374  7.987026
## experience=6                                29.74684  7.6360682  6.405838
## major_discipline=Humanities                 17.94872  2.2745735  3.162376
## experience=15                               17.83439  2.2745735  3.182647
## experience=14                               17.64706  2.1933387  3.101561
## company_type=Public Sector                  18.85965  3.4930950  4.621934
## experience=17                               13.79310  0.9748172  1.763633
## f.city_development_index=[0.691,0.878)      21.64948 17.0593014 19.663491
## education_level=Primary School              11.84211  0.7311129  1.540645
## experience=10                               17.08333  3.3306255  4.865194
## education_level=Phd                         12.50000  0.9748172  1.946077
## company_size=<10                            18.61702  5.6864338  7.622137
## company_size=500-999                        17.24138  3.6555646  5.290898
## experience=16                               12.32877  1.4622258  2.959659
## company_size=10000+                         17.96407  7.3111292 10.156092
## f.city_development_index=0.920              20.94281 22.0146223 26.231502
## education_level=Masters                     20.23593 18.1153534 22.339347
## education_level=High School                 17.34513  7.9610073 11.453477
## major_discipline=No Major                   18.14404 10.6417547 14.636124
## company_type=Funded Startup                 13.20755  2.8432169  5.371985
## company_size=50-99                          18.14815 11.9415110 16.420028
## gender=Male                                 22.84892 64.5004062 70.443949
## company_size=1000-4999                      13.51351  3.6555646  6.750456
## company_size=100-500                        16.15854  8.6108855 13.298196
## last_new_job=>4                             16.52593 11.1291633 16.805190
## relevent_experience=Has relevent experience 22.20988 64.9878148 73.018447
## experience=>20                              15.04212 10.1543461 16.845733
## enrolled_university=no_enrollment           21.78109 64.1754671 73.525238
## f.city_development_index=[0.921,0.949]      10.07519  5.4427295 13.480641
## f.city_development_index=[0.878,0.920)      11.90000  9.6669374 20.271640
## company_type=Pvt Ltd                        18.49415 41.1047929 55.463207
##                                                   p.value     v.test
## f.city_development_index=[0.448,0.691)      1.426886e-129  24.218315
## company_size=Unknown                         1.783083e-61  16.543515
```

```
## company_type=Other                                5.171400e-59  16.198450
## enrolled_university=Full time course              4.870655e-18   8.656361
## education_level=Graduate                          7.189505e-16   8.067257
## relevent_experience=No relevent experience        6.044670e-13   7.199471
## gender=Other                                      3.236147e-08   5.528156
## major_discipline=STEM                             2.561324e-06   4.703186
## experience=<1                                     2.819389e-06   4.683557
## experience=4                                      2.943942e-06   4.674693
## experience=1                                      3.792657e-06   4.622433
## experience=3                                      3.430779e-05   4.142809
## experience=2                                      1.503261e-03   3.174053
## last_new_job=1                                    4.441380e-03   2.844983
## last_new_job=never                                9.927028e-03   2.578361
## experience=5                                      1.369071e-02   2.465278
## experience=6                                      4.517093e-02   2.003059
## major_discipline=Humanities                       3.566666e-02  -2.100707
## experience=15                                     3.204593e-02  -2.143837
## experience=14                                     2.976693e-02  -2.173178
## company_type=Public Sector                        2.631339e-02  -2.221555
## experience=17                                     1.128273e-02  -2.533818
## f.city_development_index=[0.691,0.878)            7.387903e-03  -2.678834
## education_level=Primary School                    4.879655e-03  -2.814873
## experience=10                                     2.869768e-03  -2.981354
## education_level=Phd                               2.632041e-03  -3.007733
## company_size=<10                                  2.470553e-03  -3.026925
## company_size=500-999                              2.274236e-03  -3.051866
## experience=16                                     1.421461e-04  -3.804404
## company_size=10000+                               8.706940e-05  -3.924059
## f.city_development_index=0.920                    8.626808e-05  -3.926285
## education_level=Masters                           3.056277e-05  -4.169232
## education_level=High School                       4.551971e-06  -4.584444
## major_discipline=No Major                         2.593774e-06  -4.700616
## company_type=Funded Startup                       1.482631e-06  -4.813563
## company_size=50-99                                5.006137e-07  -5.026077
## gender=Male                                       1.836673e-07  -5.215151
## company_size=1000-4999                            1.261415e-07  -5.284365
## company_size=100-500                              6.384243e-09  -5.806368
## last_new_job=>4                                   1.977683e-10  -6.363064
## relevent_experience=Has relevent experience       6.044670e-13  -7.199471
## experience=>20                                    4.175352e-14  -7.555412
## enrolled_university=no_enrollment                 4.181199e-17  -8.407703
## f.city_development_index=[0.921,0.949]            6.337961e-25 -10.310184
## f.city_development_index=[0.878,0.920)            8.949731e-30 -11.333555
## company_type=Pvt Ltd                              1.696974e-31 -11.675683
```

```
cat$quanti
```

```
## $`0`
##                           v.test Mean in category Overall mean sd in category
## city_development_index 25.076187        0.8531799     0.827733       0.104995
## n.experience           12.278868       11.4989195    10.701601       7.996310
## training_hours          2.030853       65.1220962    64.140888      59.865640
##                        Overall sd       p.value
## city_development_index  0.1235873 9.047238e-139
```

```
## n.experience            7.9081330  1.176323e-34
## training_hours         58.8414622  4.226992e-02
##
## $`1`
##                            v.test Mean in category Overall mean sd in category
## training_hours          -2.030853       61.1900894    64.140888     55.5436284
## n.experience           -12.278868        8.3038180    10.701601      7.1175788
## city_development_index -25.076187        0.7512063     0.827733      0.1423052
##                         Overall sd       p.value
## training_hours         58.8414622  4.226992e-02
## n.experience            7.9081330  1.176323e-34
## city_development_index  0.1235873 9.047238e-139
```

Before starting with the modelling, the data should be split into working and test datasets, so that the created model can be compared and assessed with data that it has not seen, hence limiting overfitting. The chosen splitting size was 75-25.

```
library(caret)
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(020198)
trainIndex = createDataPartition(df$target, p = 0.75, list = FALSE, times = 1)
train = df[trainIndex,]
test = df[-trainIndex,]
```

Inspect the null model:

```
df = select(train, -c("city", "enrollee_id"))
m0 = glm(target ~ 1, data=df, family=binomial)
summary(m0)
```

```
##
## Call:
## glm(formula = target ~ 1, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7579  -0.7579  -0.7579  -0.7579   1.6659
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.10041    0.03798  -28.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4159.7  on 3700  degrees of freedom
## Residual deviance: 4159.7  on 3700  degrees of freedom
## AIC: 4161.7
##
## Number of Fisher Scoring iterations: 4
```

After computing the null model, it was assessed how to treat the attribute experience: as a factor or as a numerical variable.

Regarding being numerical, polynomial transformations were applied to it. It was seen that the p-value for a third degree polynomial suggests that this transformation is not needed (this conclusion can only be drawn because the variables constructed by Poly function are orthogonal), hence only a second order polynomial was kept. Using deviance tests, the comparison with the normal variable and the transformed one yield significantly different models, and with a better performance for the transformed one.

```
mnexp = glm(target ~ n.experience, data=df, family=binomial)
summary(mnexp)
```

```
##
## Call:
## glm(formula = target ~ n.experience, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9708  -0.8342  -0.6733  -0.4932   2.0818
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.507680   0.062302  -8.149 3.68e-16 ***
## n.experience -0.061508   0.005613 -10.959  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4159.7  on 3700  degrees of freedom
## Residual deviance: 4023.8  on 3699  degrees of freedom
## AIC: 4027.8
##
## Number of Fisher Scoring iterations: 4
```

```
mnexppoly3 = glm(target ~ poly(n.experience,3), data=df, family=binomial)
summary(mnexppoly3)
```

```
##
## Call:
## glm(formula = target ~ poly(n.experience, 3), family = binomial,
```

```
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0750  -0.8145  -0.6234  -0.5504   1.9810
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.15575    0.03990 -28.964  < 2e-16 ***
## poly(n.experience, 3)1 -27.44227    2.61067 -10.512  < 2e-16 ***
## poly(n.experience, 3)2   8.86474    2.48042   3.574 0.000352 ***
## poly(n.experience, 3)3  -0.05374    2.40147  -0.022 0.982147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4159.7  on 3700  degrees of freedom
## Residual deviance: 4010.6  on 3697  degrees of freedom
## AIC: 4018.6
##
## Number of Fisher Scoring iterations: 4
```

```
mnexppoly2 = glm(target ~ poly(n.experience,2), data=df, family=binomial)
summary(mnexppoly2)
```

```
##
## Call:
## glm(formula = target ~ poly(n.experience, 2), family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0742  -0.8148  -0.6230  -0.5505   1.9808
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.15578    0.03989 -28.975  < 2e-16 ***
## poly(n.experience, 2)1 -27.44299    2.61020 -10.514  < 2e-16 ***
## poly(n.experience, 2)2   8.87522    2.43592   3.643 0.000269 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4159.7  on 3700  degrees of freedom
## Residual deviance: 4010.6  on 3698  degrees of freedom
## AIC: 4016.6
##
## Number of Fisher Scoring iterations: 4
```

```
anova(mnexp,mnexppoly2,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ n.experience
## Model 2: target ~ poly(n.experience, 2)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3699     4023.8
## 2      3698     4010.6  1   13.188 0.0002817 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regarding experience as a factor, since it has more than 20 categories some collapses have been found to improve the model results: -Collapsing by quantiles -Collapsing the model logically in Entry Level, Junior Level, Mid Level, Senior Level and Chief Level, using some well defined year ranges for the Data Science field.

```
entry_level = c('<1','1','2')
junior_level = c('3','4')
mid_level = c('5','6')
senior_level = c('7','8','9','10')
chief_level = c('11','12','13','14','15','16','17','18','19','20','>20')

df = df %>%
  mutate(across(where(is.factor), ~ as.character(.))) %>%
  mutate(collapsed_exp = case_when(experience %in% entry_level ~ "Entry Level",
                                   experience %in% junior_level ~ "Junior Level",
                                   experience %in% mid_level ~ "Mid Level",
                                   experience %in% senior_level ~ "Senior Level",
                                   experience %in% chief_level ~ "Chief Level",
                                   TRUE ~ experience)) %>%
  mutate(across(where(is.character), ~ as.factor(.)))


groups = 5

df$collapsed_exp2 = as.ordered(cut2(df$n.experience, g=groups, m=nrow(df)/groups))
# table(df$collapsed_exp)
# table(df$collapsed_exp2)
```

Comparing both collapsed models, it can be seen that the one collapsed by quantiles is better.

```
mcexp = glm(target ~ experience, data=df, family=binomial)
summary(mcexp)
```

```
##
## Call:
## glm(formula = target ~ experience, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1040  -0.8359  -0.6011  -0.4921   2.0839
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)   -0.18924    0.19513  -0.970 0.332127
## experience>20 -1.61790    0.22718  -7.122 1.07e-12 ***
## experience1    0.01404    0.27786   0.051 0.959707
## experience10  -1.44567    0.27681  -5.223 1.76e-07 ***
## experience11  -1.05395    0.28465  -3.703 0.000213 ***
## experience12  -0.81840    0.31213  -2.622 0.008742 **
## experience13  -0.81221    0.33770  -2.405 0.016166 *
## experience14  -1.28802    0.31075  -4.145 3.40e-05 ***
## experience15  -1.43015    0.31301  -4.569 4.90e-06 ***
## experience16  -1.86093    0.35341  -5.266 1.40e-07 ***
## experience17  -1.66314    0.42741  -3.891 9.98e-05 ***
## experience18  -1.19705    0.42067  -2.846 0.004433 **
## experience19  -1.58383    0.40994  -3.864 0.000112 ***
## experience2   -0.54529    0.24273  -2.247 0.024670 *
## experience20  -1.64334    0.57278  -2.869 0.004117 **
## experience3   -0.48589    0.23673  -2.053 0.040120 *
## experience4   -0.42466    0.22968  -1.849 0.064473 .
## experience5   -0.59047    0.23150  -2.551 0.010752 *
## experience6   -0.68260    0.24213  -2.819 0.004816 **
## experience7   -0.76772    0.25336  -3.030 0.002444 **
## experience8   -1.19705    0.27164  -4.407 1.05e-05 ***
## experience9   -0.99275    0.25784  -3.850 0.000118 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4159.7  on 3700  degrees of freedom
## Residual deviance: 3990.4  on 3679  degrees of freedom
## AIC: 4034.4
##
## Number of Fisher Scoring iterations: 4
```

```
mcexpcol = glm(target ~ collapsed_exp, data=df, family=binomial)
mcexpcol2 = glm(target ~ collapsed_exp2, data=df, family=binomial)
```

After getting the best numerical and categorical transformations for the variable, the models created with them were compared. As they are not nested models, the deviance test anova() can not be applied, and it was decided to use AIC instead. It can be clearly seen that the numerical treatment of the variable outperforms the categorical, hence is the one that will be used in the following models.

```
AIC(mcexp,mcexpcol,mnexppoly2,mnexp, mcexpcol2)
```

```
##            df      AIC
## mcexp      22 4034.415
## mcexpcol    5 4036.222
## mnexppoly2  3 4016.566
## mnexp       2 4027.754
## mcexpcol2   5 4031.404
```

The same analysis can be done for the city development index (which will not be extensively reported). Even after performing the transformation suggested by the MarginalModelPlots, the discretized version of the city development index is much better.

```r
mncdi = glm(target ~ city_development_index, data=df, family=binomial); summary(mncdi) # Numerical
```

```
##
## Call:
## glm(formula = target ~ city_development_index, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6332  -0.6089  -0.5470  -0.5011   2.0675
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)              4.1256     0.2474   16.68   <2e-16 ***
## city_development_index  -6.4672     0.3086  -20.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4159.7  on 3700  degrees of freedom
## Residual deviance: 3692.0  on 3699  degrees of freedom
## AIC: 3696
##
## Number of Fisher Scoring iterations: 4
```

```r
mfcdi = glm(target ~ f.city_development_index, data=df, family=binomial); summary(mfcdi) # Categorical
```

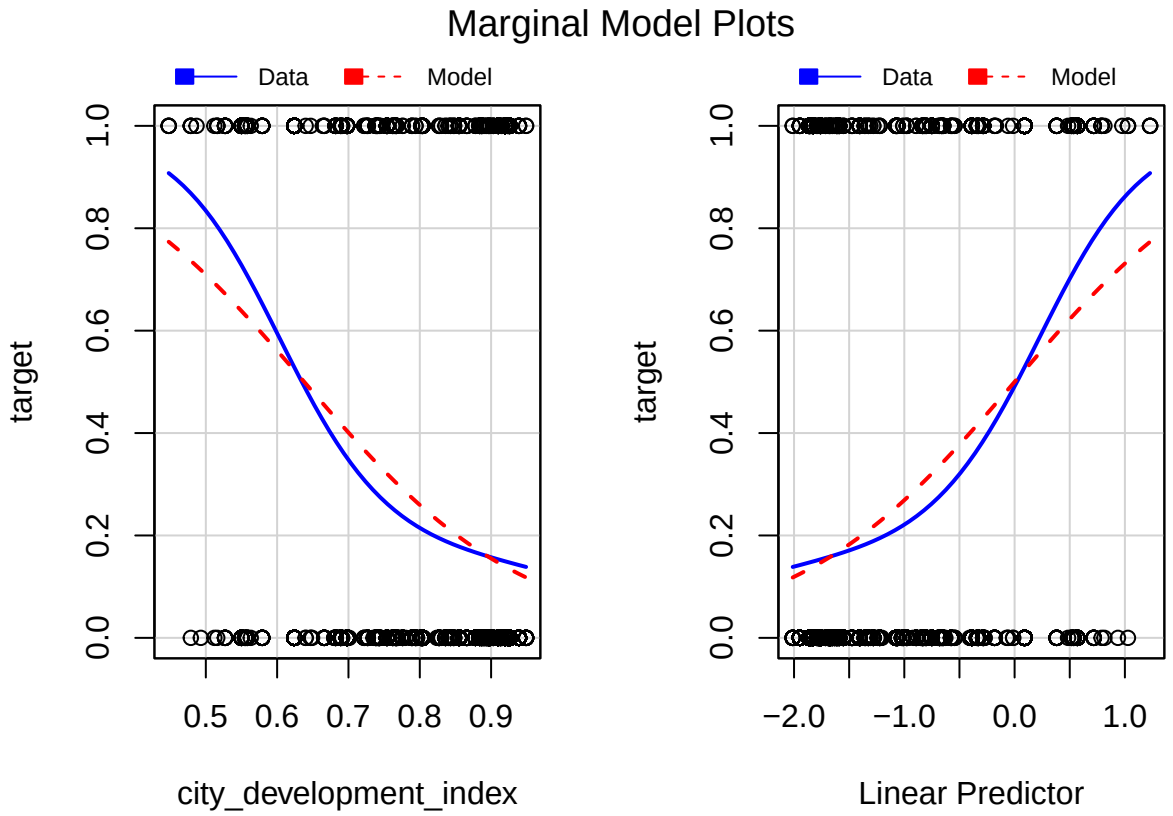
```
##
## Call:
## glm(formula = target ~ f.city_development_index, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3165  -0.6845  -0.5026  -0.4493   2.1649
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           0.32110    0.07472   4.297 1.73e-05 ***
## f.city_development_index[0.691,0.878) -1.63102    0.11783 -13.842  < 2e-16 ***
## f.city_development_index[0.878,0.920) -2.32622    0.13540 -17.181  < 2e-16 ***
## f.city_development_index[0.921,0.949] -2.56358    0.16919 -15.152  < 2e-16 ***
## f.city_development_index0.920         -1.65306    0.10798 -15.308  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4159.7  on 3700  degrees of freedom
## Residual deviance: 3628.4  on 3696  degrees of freedom
## AIC: 3638.4
```

```
##
## Number of Fisher Scoring iterations: 4
```

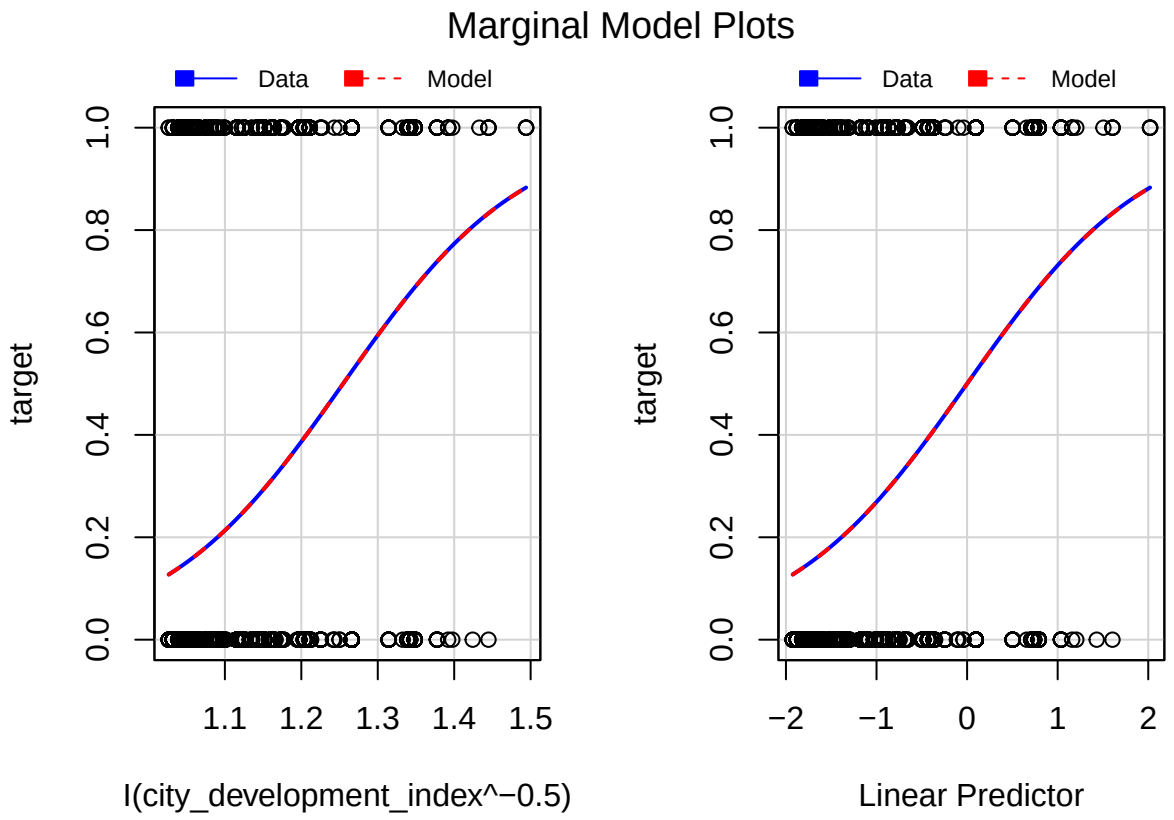
```
AIC(mncdi, mfcdi)
```

```
##       df      AIC
## mncdi  2 3696.013
## mfcdi  5 3638.376
```

```
# For the improved cdi^-0.5 - known from the marginal model plots
marginalModelPlots(mncdi)
```




```
mncdi_tr = glm(target ~ I(city_development_index^-0.5), data=df, family=binomial)
marginalModelPlots(mncdi_tr)
```




```
AIC(mncdi,mncdi_tr, mfcdi)
```

```
##          df      AIC
## mncdi     2 3696.013
## mncdi_tr  2 3685.754
## mfcdi     5 3638.376
```

```
# Discretizing the transformed index

groups = 5

df$f.city_development_index_tr = as.ordered(cut2(df$city_development_index^-0.5, g=groups, m=nrow(df)/g
mfcdi_tr = glm(target ~ f.city_development_index_tr, data=df, family=binomial)
AIC(mfcdi, mfcdi_tr)
```

```
##          df      AIC
## mfcdi     5 3638.376
## mfcdi_tr  4 3668.779
```

After having chosen the best type of variables to work with, the focus is firstly set on the two numerical variables, whose models are compared with and without interactions. As it can be seen with the deviance test, adding training hours to the model, either as an interaction or just an addition, does not yield a statistically different model, hence only the second order transformation of experience is kept.

```r
m1 = glm(target ~ training_hours, data=df, family=binomial)
m2 = glm(target ~ poly(n.experience,2), data=df, family=binomial)

m3 = glm(target ~ training_hours+poly(n.experience,2), data=df, family=binomial)
m4 = glm(target ~ training_hours*poly(n.experience,2), data=df, family=binomial)

# Gross effects
anova(m0,m1,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ 1
## Model 2: target ~ training_hours
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3700     4159.7
## 2      3699     4158.8  1  0.86562   0.3522
```

```r
anova(m0,m2,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ 1
## Model 2: target ~ poly(n.experience, 2)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3700     4159.7
## 2      3698     4010.6  2   149.09 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Net effects
anova(m1,m3,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ training_hours
## Model 2: target ~ training_hours + poly(n.experience, 2)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3699     4158.8
## 2      3697     4010.2  2   148.54 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m2,m3,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ poly(n.experience, 2)
## Model 2: target ~ training_hours + poly(n.experience, 2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3698     4010.6
## 2      3697     4010.2  1  0.31884   0.5723
```

```r
# Interaction effects
anova(m3,m4,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ training_hours + poly(n.experience, 2)
## Model 2: target ~ training_hours * poly(n.experience, 2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3697     4010.2
## 2      3695     4007.5  2   2.7093    0.258
```

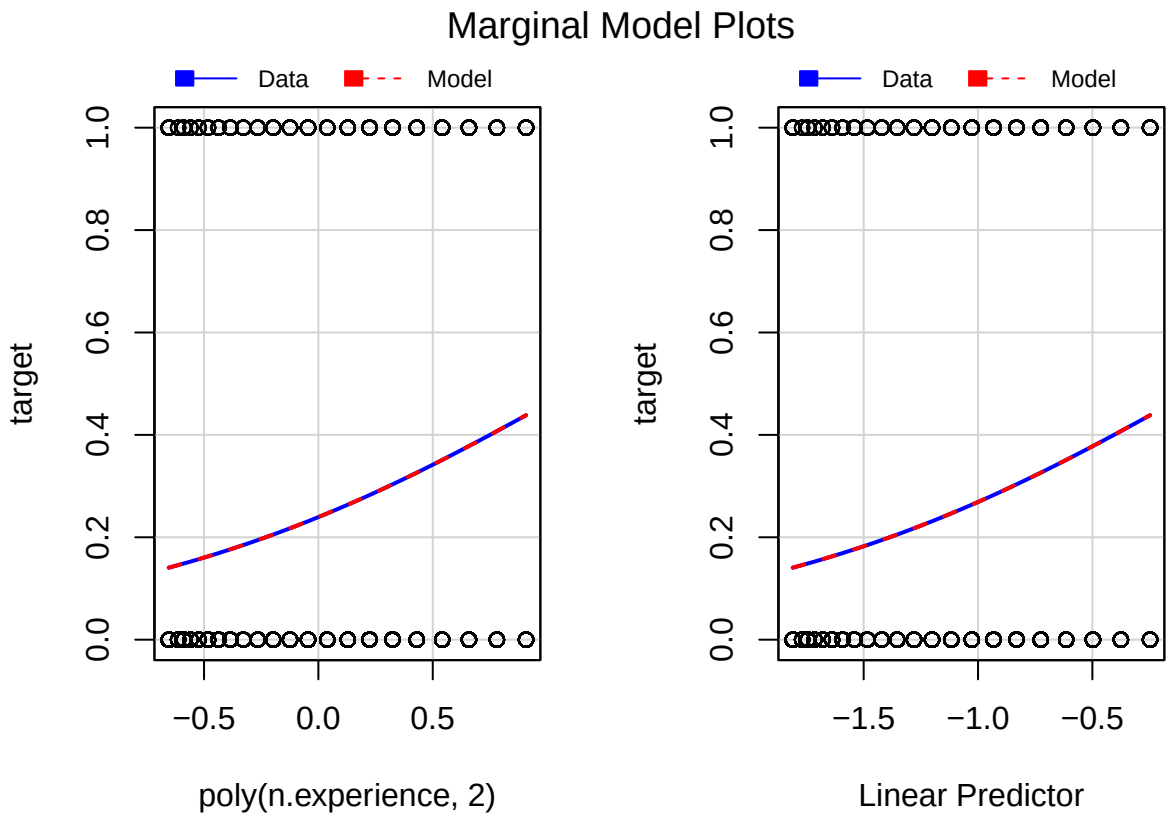
```r
AIC(m0, m1, m2, m3, m4)
```

```
##    df      AIC
## m0  1 4161.656
## m1  2 4162.790
## m2  3 4016.566
## m3  4 4018.247
## m4  6 4019.537
```

Assessing it with marginal model plots, no transformations are suggested, as it yields a perfect fit.

```r
marginalModelPlots(m2)
```

```
## Warning in mmps(...): Splines and/or polynomials replaced by a fitted linear
## combination
```



After that, the additive effect of variables is explored by using a step function with AIC, to be more permissive. It results in suggesting the addition of 7 of the variables to the model, which is significantly different and much better than the previous best one. Also, multicollinearity was discarded by doing a vif test.

```
df = select(df, -c("experience", "collapsed_exp", "collapsed_exp2", "city_development_index", "f.city_d
m5 = glm(target ~ poly(n.experience,2) + . - imputed, data=df, family=binomial)
maic = step(m5)
```

```
## Start:  AIC=3329.08
## target ~ poly(n.experience, 2) + (gender + relevent_experience +
##     enrolled_university + education_level + major_discipline +
##     company_size + company_type + last_new_job + training_hours +
##     imputed + n.experience + f.city_development_index) - imputed
##
##
## Step:  AIC=3329.08
## target ~ poly(n.experience, 2) + gender + relevent_experience +
##     enrolled_university + education_level + major_discipline +
##     company_size + company_type + last_new_job + training_hours +
##     f.city_development_index
##
##                            Df Deviance    AIC
## - major_discipline          5   3253.5 3323.5
## - gender                    2   3249.7 3325.7
## - company_type              5   3258.1 3328.1
## <none>                          3249.1 3329.1
## - training_hours            1   3251.4 3329.4
## - enrolled_university       2   3254.2 3330.2
## - relevent_experience       1   3252.3 3330.3
## - poly(n.experience, 2)     2   3262.6 3338.6
## - last_new_job              5   3270.1 3340.1
## - company_size              8   3277.7 3341.7
## - education_level           4   3271.0 3343.0
## - f.city_development_index  4   3641.6 3713.6
##
## Step:  AIC=3323.47
## target ~ poly(n.experience, 2) + gender + relevent_experience +
##     enrolled_university + education_level + company_size + company_type +
##     last_new_job + training_hours + f.city_development_index
##
##                            Df Deviance    AIC
## - gender                    2   3254.0 3320.0
## - company_type              5   3262.2 3322.2
## <none>                          3253.5 3323.5
## - training_hours            1   3255.6 3323.6
## - relevent_experience       1   3256.3 3324.3
## - enrolled_university       2   3258.5 3324.5
## - poly(n.experience, 2)     2   3266.8 3332.8
## - last_new_job              5   3274.3 3334.3
## - company_size              8   3282.4 3336.4
## - education_level           4   3331.3 3393.3
## - f.city_development_index  4   3650.0 3712.0
##
## Step:  AIC=3319.98
```

```
## target ~ poly(n.experience, 2) + relevent_experience + enrolled_university +
##     education_level + company_size + company_type + last_new_job +
##     training_hours + f.city_development_index
##
##                            Df Deviance    AIC
## - company_type              5   3263.1 3319.1
## <none>                          3254.0 3320.0
## - training_hours            1   3256.0 3320.0
## - relevent_experience       1   3256.8 3320.8
## - enrolled_university       2   3259.2 3321.2
## - poly(n.experience, 2)     2   3267.8 3329.8
## - last_new_job              5   3274.8 3330.8
## - company_size              8   3282.7 3332.7
## - education_level           4   3331.8 3389.8
## - f.city_development_index  4   3656.2 3714.2
##
## Step:  AIC=3319.09
## target ~ poly(n.experience, 2) + relevent_experience + enrolled_university +
##     education_level + company_size + last_new_job + training_hours +
##     f.city_development_index
##
##                            Df Deviance    AIC
## <none>                          3263.1 3319.1
## - training_hours            1   3265.2 3319.2
## - relevent_experience       1   3266.5 3320.5
## - enrolled_university       2   3269.3 3321.3
## - poly(n.experience, 2)     2   3276.6 3328.6
## - last_new_job              5   3287.6 3333.6
## - education_level           4   3338.9 3386.9
## - company_size              8   3480.2 3520.2
## - f.city_development_index  4   3665.7 3713.7
```

```
summary(maic)
```

```
##
## Call:
## glm(formula = target ~ poly(n.experience, 2) + relevent_experience +
##     enrolled_university + education_level + company_size + last_new_job +
##     training_hours + f.city_development_index, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1225  -0.6055  -0.4393  -0.1245   2.8888
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                              2.790e-01  2.526e-01   1.105
## poly(n.experience, 2)1                  -1.316e+01  3.623e+00  -3.632
## poly(n.experience, 2)2                   5.032e-01  2.870e+00   0.175
## relevent_experienceNo relevent experience  2.220e-01  1.190e-01   1.865
## enrolled_universityno_enrollment        -2.683e-01  1.172e-01  -2.290
## enrolled_universityPart time course     -4.957e-03  1.944e-01  -0.025
## education_levelHigh School              -1.326e+00  1.736e-01  -7.640
```

```
## education_levelMasters                    -2.574e-01  1.135e-01  -2.268
## education_levelPhd                        -3.121e-01  4.315e-01  -0.723
## education_levelPrimary School             -1.560e+00  4.494e-01  -3.472
## company_size10/49                          3.777e-01  2.273e-01   1.662
## company_size100-500                       -1.649e-01  2.185e-01  -0.755
## company_size1000-4999                     -1.688e-01  2.651e-01  -0.637
## company_size10000+                        -4.721e-02  2.297e-01  -0.205
## company_size50-99                         -1.346e-02  2.052e-01  -0.066
## company_size500-999                       -1.939e-01  2.731e-01  -0.710
## company_size5000-9999                      2.239e-01  3.121e-01   0.717
## company_sizeUnknown                        1.492e+00  1.919e-01   7.775
## last_new_job1                             -7.206e-02  1.512e-01  -0.477
## last_new_job2                              5.734e-02  1.713e-01   0.335
## last_new_job3                              2.024e-01  2.218e-01   0.912
## last_new_job4                              9.323e-02  2.438e-01   0.382
## last_new_jobnever                         -7.089e-01  2.005e-01  -3.535
## training_hours                            -1.099e-03  7.666e-04  -1.434
## f.city_development_index[0.691,0.878)     -1.785e+00  1.317e-01 -13.553
## f.city_development_index[0.878,0.920)     -2.325e+00  1.495e-01 -15.551
## f.city_development_index[0.921,0.949]     -2.408e+00  1.839e-01 -13.097
## f.city_development_index0.920             -1.612e+00  1.236e-01 -13.041
##                                           Pr(>|z|)
## (Intercept)                               0.269363
## poly(n.experience, 2)1                    0.000282 ***
## poly(n.experience, 2)2                    0.860834
## relevent_experienceNo relevent experience 0.062250 .
## enrolled_universityno_enrollment          0.022039 *
## enrolled_universityPart time course       0.979663
## education_levelHigh School                2.18e-14 ***
## education_levelMasters                    0.023357 *
## education_levelPhd                        0.469532
## education_levelPrimary School             0.000516 ***
## company_size10/49                         0.096562 .
## company_size100-500                       0.450324
## company_size1000-4999                     0.524243
## company_size10000+                        0.837205
## company_size50-99                         0.947729
## company_size500-999                       0.477752
## company_size5000-9999                     0.473159
## company_sizeUnknown                       7.55e-15 ***
## last_new_job1                             0.633613
## last_new_job2                             0.737852
## last_new_job3                             0.361570
## last_new_job4                             0.702214
## last_new_jobnever                         0.000407 ***
## training_hours                            0.151599
## f.city_development_index[0.691,0.878)      < 2e-16 ***
## f.city_development_index[0.878,0.920)      < 2e-16 ***
## f.city_development_index[0.921,0.949]      < 2e-16 ***
## f.city_development_index0.920              < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 4159.7  on 3700  degrees of freedom
## Residual deviance: 3263.1  on 3673  degrees of freedom
## AIC: 3319.1
##
## Number of Fisher Scoring iterations: 5
```

```
vif(maic)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## poly(n.experience, 2)    1.784647  2        1.155814
## relevent_experience      1.598800  1        1.264437
## enrolled_university      1.373758  2        1.082624
## education_level          1.359318  4        1.039119
## company_size             1.513986  8        1.026261
## last_new_job             1.818617  5        1.061632
## training_hours           1.013316  1        1.006636
## f.city_development_index 1.313912  4        1.034715
```

```
anova(m2,maic,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ poly(n.experience, 2)
## Model 2: target ~ poly(n.experience, 2) + relevent_experience + enrolled_university +
##     education_level + company_size + last_new_job + training_hours +
##     f.city_development_index
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3698     4010.6
## 2      3673     3263.1 25   747.47 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(m2,maic)
```

```
##      df      AIC
## m2    3 4016.566
## maic 28 3319.095
```

Some factors still have a lot of levels, but two of them can be further collapsed to improve the model results:

```
# Company size collapse
aux = df %>%
  mutate(across(where(is.factor), ~ as.character(.))) %>%
  mutate(company_size = case_when(company_size != "Unknown" ~ "Known",
                                  TRUE ~ company_size)) %>%
  mutate(across(where(is.character), ~ as.factor(.)))

maux = glm(formula = target ~ poly(n.experience, 2) + relevent_experience +
    enrolled_university + education_level + company_size + last_new_job +
    training_hours + f.city_development_index, family = binomial,
```

```
    data = aux)

# Education level collapse
aux2 = aux %>%
  mutate(across(where(is.factor), ~ as.character(.))) %>%
  mutate(education_level = case_when(education_level %in% c("Masters", "Phd") ~ "MastersPhd",
                                     TRUE ~ education_level)) %>%
  mutate(across(where(is.character), ~ as.factor(.)))

maux2 = glm(formula = target ~ poly(n.experience, 2) + relevent_experience +
    enrolled_university + education_level + company_size + last_new_job +
    training_hours + f.city_development_index, family = binomial,
    data = aux2)
```

To check for interactions the step function was used again, but this time using the BIC criterion in order to be more restrictive. The BIC criterion removed almost all the possible interactions but the one between company size and city index development (factorized).

```
mbic = step(maux, scope = . ~ .^2, k = log(nrow(df)))
```

```
## Start:  AIC=3445.42
## target ~ poly(n.experience, 2) + relevent_experience + enrolled_university +
##     education_level + company_size + last_new_job + training_hours +
##     f.city_development_index
##
##                                                Df Deviance    AIC
## + company_size:f.city_development_index         4   3171.5 3376.9
## - last_new_job                                  5   3297.0 3428.5
## - enrolled_university                           2   3279.8 3435.9
## - training_hours                                1   3274.7 3439.1
## - relevent_experience                           1   3276.0 3440.3
## - poly(n.experience, 2)                         2   3287.7 3443.9
## <none>                                              3272.9 3445.4
## + company_size:training_hours                   1   3271.6 3452.3
## + relevent_experience:training_hours            1   3272.4 3453.1
## + relevent_experience:company_size              1   3272.6 3453.4
## + enrolled_university:company_size              2   3264.7 3453.7
## + education_level:company_size                  4   3248.9 3454.3
## + poly(n.experience, 2):company_size            2   3269.7 3458.7
## + poly(n.experience, 2):training_hours          2   3270.9 3459.9
## + poly(n.experience, 2):relevent_experience     2   3271.4 3460.4
## + enrolled_university:training_hours            2   3272.5 3461.4
## + relevent_experience:enrolled_university       2   3272.5 3461.5
## + poly(n.experience, 2):enrolled_university     4   3260.8 3466.2
## + training_hours:f.city_development_index       4   3262.5 3467.9
## + relevent_experience:f.city_development_index  4   3262.6 3468.0
## + company_size:last_new_job                     5   3256.0 3469.6
## + relevent_experience:last_new_job              5   3256.7 3470.3
## + relevent_experience:education_level           4   3269.8 3475.2
## + education_level:training_hours                4   3271.5 3476.9
## + last_new_job:training_hours                   5   3266.0 3479.6
## - education_level                               4   3348.3 3488.0
## + poly(n.experience, 2):education_level         8   3258.3 3496.6
```

```
## + enrolled_university:f.city_development_index    8   3260.2 3498.5
## + poly(n.experience, 2):f.city_development_index  8   3260.5 3498.8
## + enrolled_university:education_level             8   3262.6 3500.8
## + enrolled_university:last_new_job               10   3260.4 3515.1
## + poly(n.experience, 2):last_new_job             10   3260.6 3515.4
## + education_level:f.city_development_index       16   3255.6 3559.6
## + education_level:last_new_job                   20   3229.8 3566.6
## + last_new_job:f.city_development_index          20   3246.9 3583.8
## - company_size                                    1   3480.2 3644.6
## - f.city_development_index                        4   3679.6 3819.3
##
## Step:  AIC=3376.88
## target ~ poly(n.experience, 2) + relevent_experience + enrolled_university +
##     education_level + company_size + last_new_job + training_hours +
##     f.city_development_index + company_size:f.city_development_index
##
##                                                Df Deviance    AIC
## - last_new_job                                  5   3189.6 3353.9
## - enrolled_university                           2   3179.2 3368.2
## - training_hours                                1   3172.9 3370.1
## - poly(n.experience, 2)                         2   3183.2 3372.2
## - relevent_experience                           1   3175.9 3373.1
## <none>                                              3171.5 3376.9
## + company_size:training_hours                   1   3170.9 3384.5
## + relevent_experience:training_hours            1   3171.2 3384.9
## + relevent_experience:company_size              1   3171.3 3385.0
## + enrolled_university:company_size              2   3164.3 3386.1
## + education_level:company_size                  4   3151.5 3389.8
## + poly(n.experience, 2):training_hours          2   3169.7 3391.5
## + poly(n.experience, 2):company_size            2   3170.2 3392.0
## + poly(n.experience, 2):relevent_experience     2   3170.8 3392.6
## + enrolled_university:training_hours            2   3170.9 3392.7
## + relevent_experience:enrolled_university       2   3171.1 3393.0
## + training_hours:f.city_development_index       4   3161.8 3400.1
## + poly(n.experience, 2):enrolled_university     4   3162.6 3400.9
## + relevent_experience:last_new_job              5   3159.4 3405.9
## + relevent_experience:f.city_development_index  4   3167.9 3406.2
## + relevent_experience:education_level           4   3168.1 3406.4
## + education_level:training_hours                4   3169.9 3408.2
## + company_size:last_new_job                     5   3162.9 3409.4
## + last_new_job:training_hours                   5   3164.7 3411.2
## - education_level                               4   3243.9 3416.5
## + poly(n.experience, 2):f.city_development_index  8   3159.0 3430.2
## + poly(n.experience, 2):education_level         8   3160.5 3431.7
## + enrolled_university:education_level           8   3160.6 3431.8
## + enrolled_university:f.city_development_index  8   3161.1 3432.2
## - company_size:f.city_development_index         4   3272.9 3445.4
## + poly(n.experience, 2):last_new_job           10   3158.8 3446.4
## + enrolled_university:last_new_job             10   3162.6 3450.2
## + education_level:f.city_development_index     16   3137.8 3474.6
## + education_level:last_new_job                 20   3133.3 3503.1
## + last_new_job:f.city_development_index        20   3136.3 3506.0
##
## Step:  AIC=3353.89
```

```
## target ~ poly(n.experience, 2) + relevent_experience + enrolled_university +
##     education_level + company_size + training_hours + f.city_development_index +
##     company_size:f.city_development_index
##
##                                                Df Deviance    AIC
## - enrolled_university                           2   3196.7 3344.6
## - training_hours                                1   3190.8 3346.9
## - poly(n.experience, 2)                         2   3199.8 3347.7
## - relevent_experience                           1   3191.8 3347.9
## <none>                                              3189.6 3353.9
## + relevent_experience:company_size              1   3188.8 3361.3
## + company_size:training_hours                   1   3189.0 3361.5
## + relevent_experience:training_hours            1   3189.3 3361.9
## + enrolled_university:company_size              2   3181.7 3362.4
## + education_level:company_size                  4   3166.1 3363.3
## + poly(n.experience, 2):training_hours          2   3187.7 3368.4
## + poly(n.experience, 2):company_size            2   3188.3 3369.1
## + poly(n.experience, 2):relevent_experience     2   3188.6 3369.4
## + enrolled_university:training_hours            2   3189.0 3369.8
## + relevent_experience:enrolled_university       2   3189.2 3369.9
## + training_hours:f.city_development_index       4   3179.4 3376.5
## + last_new_job                                  5   3171.5 3376.9
## + poly(n.experience, 2):enrolled_university     4   3180.8 3378.0
## + relevent_experience:education_level           4   3184.4 3381.6
## + relevent_experience:f.city_development_index  4   3185.7 3382.8
## + education_level:training_hours                4   3188.2 3385.4
## + poly(n.experience, 2):education_level         8   3177.0 3407.1
## + poly(n.experience, 2):f.city_development_index 8   3177.1 3407.1
## + enrolled_university:f.city_development_index  8   3178.2 3408.3
## + enrolled_university:education_level           8   3178.3 3408.4
## - education_level                               4   3284.0 3415.5
## - company_size:f.city_development_index         4   3297.0 3428.5
## + education_level:f.city_development_index     16   3155.9 3451.7
##
## Step:  AIC=3344.63
## target ~ poly(n.experience, 2) + relevent_experience + education_level +
##     company_size + training_hours + f.city_development_index +
##     company_size:f.city_development_index
##
##                                                Df Deviance    AIC
## - training_hours                                1   3198.0 3337.6
## - relevent_experience                           1   3201.0 3340.6
## - poly(n.experience, 2)                         2   3211.9 3343.4
## <none>                                              3196.7 3344.6
## + education_level:company_size                  4   3171.1 3351.9
## + relevent_experience:company_size              1   3196.0 3352.1
## + company_size:training_hours                   1   3196.1 3352.3
## + relevent_experience:training_hours            1   3196.5 3352.6
## + enrolled_university                           2   3189.6 3353.9
## + poly(n.experience, 2):training_hours          2   3194.7 3359.1
## + poly(n.experience, 2):company_size            2   3195.0 3359.3
## + poly(n.experience, 2):relevent_experience     2   3195.3 3359.6
## + training_hours:f.city_development_index       4   3187.0 3367.8
## + last_new_job                                  5   3179.2 3368.2
```

```
## + relevent_experience:education_level             4   3190.7 3371.4
## + relevent_experience:f.city_development_index    4   3192.3 3373.1
## + education_level:training_hours                  4   3195.1 3375.9
## + poly(n.experience, 2):education_level           8   3183.2 3396.8
## + poly(n.experience, 2):f.city_development_index  8   3183.3 3397.0
## - education_level                                 4   3291.8 3406.9
## - company_size:f.city_development_index           4   3303.3 3418.3
## + education_level:f.city_development_index       16   3161.9 3441.3
##
## Step:  AIC=3337.65
## target ~ poly(n.experience, 2) + relevent_experience + education_level +
##     company_size + f.city_development_index + company_size:f.city_development_index
##
##                                                Df Deviance    AIC
## - relevent_experience                           1   3202.3 3333.7
## - poly(n.experience, 2)                         2   3213.4 3336.7
## <none>                                              3198.0 3337.6
## + training_hours                                1   3196.7 3344.6
## + education_level:company_size                  4   3172.3 3344.8
## + relevent_experience:company_size              1   3197.2 3345.1
## + enrolled_university                           2   3190.8 3346.9
## + poly(n.experience, 2):company_size            2   3196.2 3352.3
## + poly(n.experience, 2):relevent_experience     2   3196.5 3352.6
## + last_new_job                                  5   3180.7 3361.4
## + relevent_experience:education_level           4   3191.9 3364.4
## + relevent_experience:f.city_development_index  4   3193.5 3366.0
## + poly(n.experience, 2):education_level         8   3184.4 3389.8
## + poly(n.experience, 2):f.city_development_index 8  3184.5 3389.9
## - education_level                               4   3293.1 3399.9
## - company_size:f.city_development_index         4   3304.9 3411.7
## + education_level:f.city_development_index     16   3163.1 3434.2
##
## Step:  AIC=3333.73
## target ~ poly(n.experience, 2) + education_level + company_size +
##     f.city_development_index + company_size:f.city_development_index
##
##                                                Df Deviance    AIC
## <none>                                              3202.3 3333.7
## + relevent_experience                           1   3198.0 3337.6
## - poly(n.experience, 2)                         2   3224.9 3339.9
## + training_hours                                1   3201.0 3340.6
## + enrolled_university                           2   3193.1 3341.0
## + education_level:company_size                  4   3176.6 3341.0
## + poly(n.experience, 2):company_size            2   3200.3 3348.2
## + last_new_job                                  5   3188.2 3360.8
## + poly(n.experience, 2):education_level         8   3188.8 3386.0
## + poly(n.experience, 2):f.city_development_index 8  3188.9 3386.1
## - education_level                               4   3293.1 3391.7
## - company_size:f.city_development_index         4   3307.4 3406.0
## + education_level:f.city_development_index     16   3168.1 3431.0
```

```
mbic2 = step(maux2, scope = . ~ .^2, k = log(nrow(df)))
```

```
## Start:  AIC=3437.21
```

```
## target ~ poly(n.experience, 2) + relevent_experience + enrolled_university +
##     education_level + company_size + last_new_job + training_hours +
##     f.city_development_index
##
##                                                Df Deviance    AIC
## + company_size:f.city_development_index         4   3171.5 3368.7
## - last_new_job                                  5   3297.0 3420.2
## - enrolled_university                           2   3279.8 3427.7
## - training_hours                                1   3274.8 3430.9
## - relevent_experience                           1   3276.0 3432.1
## - poly(n.experience, 2)                         2   3287.9 3435.8
## <none>                                              3272.9 3437.2
## + education_level:company_size                  3   3253.0 3442.0
## + company_size:training_hours                   1   3271.6 3444.1
## + relevent_experience:training_hours            1   3272.4 3444.9
## + relevent_experience:company_size              1   3272.6 3445.2
## + enrolled_university:company_size              2   3264.7 3445.4
## + poly(n.experience, 2):company_size            2   3269.7 3450.5
## + poly(n.experience, 2):training_hours          2   3270.9 3451.7
## + poly(n.experience, 2):relevent_experience     2   3271.4 3452.2
## + enrolled_university:training_hours            2   3272.5 3453.2
## + relevent_experience:enrolled_university       2   3272.5 3453.3
## + poly(n.experience, 2):enrolled_university     4   3260.8 3458.0
## + relevent_experience:education_level           3   3270.1 3459.1
## + training_hours:f.city_development_index       4   3262.5 3459.7
## + relevent_experience:f.city_development_index  4   3262.6 3459.8
## + education_level:training_hours                3   3272.0 3461.0
## + company_size:last_new_job                     5   3256.0 3461.4
## + relevent_experience:last_new_job              5   3256.7 3462.1
## + last_new_job:training_hours                   5   3266.0 3471.4
## + poly(n.experience, 2):education_level         6   3261.9 3475.5
## + enrolled_university:education_level           6   3267.3 3481.0
## - education_level                               3   3348.3 3488.0
## + enrolled_university:f.city_development_index  8   3260.2 3490.2
## + poly(n.experience, 2):f.city_development_index 8  3260.5 3490.5
## + enrolled_university:last_new_job             10   3260.4 3506.9
## + poly(n.experience, 2):last_new_job           10   3260.6 3507.1
## + education_level:last_new_job                 15   3237.8 3525.4
## + education_level:f.city_development_index     12   3264.4 3527.4
## + last_new_job:f.city_development_index        20   3246.9 3575.6
## - company_size                                  1   3481.0 3637.1
## - f.city_development_index                      4   3680.0 3811.4
##
## Step:  AIC=3368.69
## target ~ poly(n.experience, 2) + relevent_experience + enrolled_university +
##     education_level + company_size + last_new_job + training_hours +
##     f.city_development_index + company_size:f.city_development_index
##
##                                                Df Deviance    AIC
## - last_new_job                                  5   3189.6 3345.7
## - enrolled_university                           2   3179.3 3360.0
## - training_hours                                1   3172.9 3361.9
## - poly(n.experience, 2)                         2   3183.2 3364.0
## - relevent_experience                           1   3176.0 3365.0
```

```
## <none>                                                 3171.5 3368.7
## + company_size:training_hours                      1   3170.9 3376.3
## + relevent_experience:training_hours               1   3171.3 3376.7
## + relevent_experience:company_size                 1   3171.4 3376.8
## + enrolled_university:company_size                 2   3164.3 3378.0
## + education_level:company_size                     3   3156.6 3378.4
## + poly(n.experience, 2):training_hours             2   3169.7 3383.3
## + poly(n.experience, 2):company_size               2   3170.2 3383.8
## + poly(n.experience, 2):relevent_experience        2   3170.8 3384.5
## + enrolled_university:training_hours               2   3170.9 3384.5
## + relevent_experience:enrolled_university          2   3171.2 3384.8
## + relevent_experience:education_level              3   3168.6 3390.4
## + training_hours:f.city_development_index          4   3161.9 3391.9
## + education_level:training_hours                   3   3170.4 3392.3
## + poly(n.experience, 2):enrolled_university        4   3162.7 3392.7
## + relevent_experience:last_new_job                 5   3159.4 3397.7
## + relevent_experience:f.city_development_index     4   3168.0 3398.0
## + company_size:last_new_job                        5   3162.9 3401.2
## + last_new_job:training_hours                      5   3164.8 3403.0
## + poly(n.experience, 2):education_level            6   3164.1 3410.6
## + enrolled_university:education_level              6   3166.2 3412.7
## - education_level                                  3   3243.9 3416.5
## + poly(n.experience, 2):f.city_development_index   8   3159.1 3422.0
## + enrolled_university:f.city_development_index     8   3161.1 3424.0
## - company_size:f.city_development_index            4   3272.9 3437.2
## + poly(n.experience, 2):last_new_job              10   3158.8 3438.2
## + enrolled_university:last_new_job                10   3162.7 3442.1
## + education_level:f.city_development_index        12   3147.4 3443.2
## + education_level:last_new_job                    15   3142.5 3463.0
## + last_new_job:f.city_development_index           20   3136.4 3497.9
##
## Step:  AIC=3345.72
## target ~ poly(n.experience, 2) + relevent_experience + enrolled_university +
##     education_level + company_size + training_hours + f.city_development_index +
##     company_size:f.city_development_index
##
##                                                   Df Deviance    AIC
## - enrolled_university                              2   3196.8 3336.4
## - training_hours                                   1   3190.8 3338.7
## - poly(n.experience, 2)                            2   3199.8 3339.5
## - relevent_experience                              1   3191.9 3339.8
## <none>                                                 3189.6 3345.7
## + education_level:company_size                     3   3171.1 3351.9
## + relevent_experience:company_size                 1   3188.8 3353.1
## + company_size:training_hours                      1   3189.0 3353.4
## + relevent_experience:training_hours               1   3189.4 3353.7
## + enrolled_university:company_size                 2   3181.7 3354.3
## + poly(n.experience, 2):training_hours             2   3187.7 3360.2
## + poly(n.experience, 2):company_size               2   3188.4 3360.9
## + poly(n.experience, 2):relevent_experience        2   3188.7 3361.2
## + enrolled_university:training_hours               2   3189.1 3361.6
## + relevent_experience:enrolled_university          2   3189.2 3361.7
## + relevent_experience:education_level              3   3184.9 3365.7
## + training_hours:f.city_development_index          4   3179.4 3368.4
```

```
## + last_new_job                                      5   3171.5 3368.7
## + education_level:training_hours                    3   3188.6 3369.4
## + poly(n.experience, 2):enrolled_university        4   3180.8 3369.8
## + relevent_experience:f.city_development_index     4   3185.7 3374.7
## + poly(n.experience, 2):education_level            6   3180.8 3386.2
## + enrolled_university:education_level              6   3183.1 3388.5
## + poly(n.experience, 2):f.city_development_index  8   3177.2 3399.0
## + enrolled_university:f.city_development_index     8   3178.3 3400.1
## - education_level                                   3   3284.0 3415.5
## + education_level:f.city_development_index        12   3164.8 3419.5
## - company_size:f.city_development_index             4   3297.0 3420.2
##
## Step:  AIC=3336.44
## target ~ poly(n.experience, 2) + relevent_experience + education_level +
##     company_size + training_hours + f.city_development_index +
##     company_size:f.city_development_index
##
##                                                    Df Deviance    AIC
## - training_hours                                    1   3198.0 3329.5
## - relevent_experience                               1   3201.1 3332.6
## - poly(n.experience, 2)                             2   3212.0 3335.2
## <none>                                                  3196.8 3336.4
## + education_level:company_size                      3   3176.3 3340.6
## + relevent_experience:company_size                  1   3196.0 3343.9
## + company_size:training_hours                       1   3196.2 3344.1
## + relevent_experience:training_hours                1   3196.6 3344.5
## + enrolled_university                               2   3189.6 3345.7
## + poly(n.experience, 2):training_hours              2   3194.8 3350.9
## + poly(n.experience, 2):company_size                2   3195.0 3351.1
## + poly(n.experience, 2):relevent_experience         2   3195.3 3351.4
## + relevent_experience:education_level               3   3191.1 3355.4
## + training_hours:f.city_development_index           4   3187.1 3359.6
## + education_level:training_hours                    3   3195.7 3360.0
## + last_new_job                                      5   3179.3 3360.0
## + relevent_experience:f.city_development_index     4   3192.4 3364.9
## + poly(n.experience, 2):education_level            6   3186.9 3375.9
## + poly(n.experience, 2):f.city_development_index  8   3183.4 3388.8
## - education_level                                   3   3291.8 3406.9
## + education_level:f.city_development_index        12   3170.7 3408.9
## - company_size:f.city_development_index             4   3303.3 3410.1
##
## Step:  AIC=3329.45
## target ~ poly(n.experience, 2) + relevent_experience + education_level +
##     company_size + f.city_development_index + company_size:f.city_development_index
##
##                                                    Df Deviance    AIC
## - relevent_experience                               1   3202.4 3325.6
## - poly(n.experience, 2)                             2   3213.5 3328.5
## <none>                                                  3198.0 3329.5
## + education_level:company_size                      3   3177.6 3333.7
## + training_hours                                    1   3196.8 3336.4
## + relevent_experience:company_size                  1   3197.3 3336.9
## + enrolled_university                               2   3190.8 3338.7
## + poly(n.experience, 2):company_size                2   3196.2 3344.1
```

```
## + poly(n.experience, 2):relevent_experience       2   3196.6 3344.5
## + relevent_experience:education_level             3   3192.3 3348.4
## + last_new_job                                    5   3180.7 3353.2
## + relevent_experience:f.city_development_index    4   3193.5 3357.9
## + poly(n.experience, 2):education_level           6   3188.2 3368.9
## + poly(n.experience, 2):f.city_development_index  8   3184.5 3381.7
## - education_level                                 3   3293.1 3399.9
## + education_level:f.city_development_index       12   3171.9 3401.9
## - company_size:f.city_development_index           4   3304.9 3403.5
##
## Step:  AIC=3325.61
## target ~ poly(n.experience, 2) + education_level + company_size +
##     f.city_development_index + company_size:f.city_development_index
##
##                                                Df Deviance    AIC
## <none>                                              3202.4 3325.6
## + relevent_experience                           1   3198.0 3329.5
## + education_level:company_size                  3   3182.3 3330.2
## - poly(n.experience, 2)                         2   3224.9 3331.7
## + training_hours                                1   3201.1 3332.6
## + enrolled_university                           2   3193.2 3332.8
## + poly(n.experience, 2):company_size            2   3200.4 3340.1
## + last_new_job                                  5   3188.3 3352.7
## + poly(n.experience, 2):education_level         6   3193.0 3365.5
## + poly(n.experience, 2):f.city_development_index  8   3189.1 3378.1
## - education_level                               3   3293.1 3391.7
## - company_size:f.city_development_index         4   3307.4 3397.8
## + education_level:f.city_development_index      12   3176.9 3398.8
```

```
AIC(maic, maux, maux2, mbic, mbic2)
```

```
##       df      AIC
## maic  28 3319.095
## maux  21 3314.880
## maux2 20 3312.883
## mbic  16 3234.265
## mbic2 15 3232.367
```

```
summary(mbic2)
```

```
##
## Call:
## glm(formula = target ~ poly(n.experience, 2) + education_level +
##     company_size + f.city_development_index + company_size:f.city_development_index,
##     family = binomial, data = aux2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5973  -0.5575  -0.4225  -0.1512   2.7851
##
## Coefficients:
##                                                      Estimate Std. Error
## (Intercept)                                           0.24114    0.09933
```

```
## poly(n.experience, 2)1                                          -14.75756    3.17608
## poly(n.experience, 2)2                                            0.39121    2.85354
## education_levelHigh School                                       -1.31802    0.16475
## education_levelMastersPhd                                        -0.26115    0.11179
## education_levelPrimary School                                    -1.88282    0.43061
## company_sizeUnknown                                               0.39811    0.17016
## f.city_development_index[0.691,0.878)                            -2.00154    0.16743
## f.city_development_index[0.878,0.920)                            -2.38832    0.18318
## f.city_development_index[0.921,0.949]                            -2.52068    0.22452
## f.city_development_index0.920                                    -2.50964    0.16767
## company_sizeUnknown:f.city_development_index[0.691,0.878)         1.04754    0.25453
## company_sizeUnknown:f.city_development_index[0.878,0.920)         0.74709    0.28744
## company_sizeUnknown:f.city_development_index[0.921,0.949]         0.81220    0.35769
## company_sizeUnknown:f.city_development_index0.920                 2.47892    0.24958
##                                                                 z value Pr(>|z|)
## (Intercept)                                                       2.428  0.01519 *
## poly(n.experience, 2)1                                           -4.646 3.38e-06 ***
## poly(n.experience, 2)2                                            0.137  0.89095
## education_levelHigh School                                       -8.000 1.24e-15 ***
## education_levelMastersPhd                                        -2.336  0.01949 *
## education_levelPrimary School                                    -4.372 1.23e-05 ***
## company_sizeUnknown                                               2.340  0.01930 *
## f.city_development_index[0.691,0.878)                           -11.955  < 2e-16 ***
## f.city_development_index[0.878,0.920)                           -13.038  < 2e-16 ***
## f.city_development_index[0.921,0.949]                           -11.227  < 2e-16 ***
## f.city_development_index0.920                                   -14.968  < 2e-16 ***
## company_sizeUnknown:f.city_development_index[0.691,0.878)         4.116 3.86e-05 ***
## company_sizeUnknown:f.city_development_index[0.878,0.920)         2.599  0.00935 **
## company_sizeUnknown:f.city_development_index[0.921,0.949]         2.271  0.02317 *
## company_sizeUnknown:f.city_development_index0.920                 9.932  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4159.7  on 3700  degrees of freedom
## Residual deviance: 3202.4  on 3686  degrees of freedom
## AIC: 3232.4
##
## Number of Fisher Scoring iterations: 5
```

```
df = aux2
mb = mbic2
```

Henceforth, the model resulting from the bic step is the one studied.

Model Interpretation ### Explain

The model formula is as follows:

$\text{logit}(\pi_{ijk}) = \eta + \beta_1 experience + \beta_2 experience^2 + \alpha_i + \nu_j + \kappa_k + \nu\kappa_{jk}$

To interpret it, it has to be stated that the reference level is Graduate, Known company Size and from the quantile of cities with poorest development.

Some of the coefficients can be interpreted as follows: - When considering experience, all else equal, the log odds decrease during the first 20 years, and then start increasing. - Considering education level, all else equal, having low levels of education (high/primary school) decrease considerably the log odds, compared to the reference level(graduate), and Masters or PhD decrease it slightly. - For city development, it can be said: - Company_size Known: the log odds are reduced by increasing order of city development (-2,-2.38,-2.5,-2.52), all else equal. Hence the odds of changing jobs are higher for people from less developed cities from known company_size - Company_size Unknown: to assess this case, the value of the interaction coefficient must be added respectively. All else equal, the same conclusion as before can be reached, but in this case the decrease in log odds is much more smaller.

A similar argumentation could be done for company size.

Hence, people that live in an underdeveloped city and not have not reported working for company are more prone to change jobs, as well as people with a Graduate or Masters/PhD.

```
summary(mb)
```

```
##
## Call:
## glm(formula = target ~ poly(n.experience, 2) + education_level +
##     company_size + f.city_development_index + company_size:f.city_development_index,
##     family = binomial, data = aux2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5973  -0.5575  -0.4225  -0.1512   2.7851
##
## Coefficients:
##                                                                  Estimate Std. Error
## (Intercept)                                                       0.24114    0.09933
## poly(n.experience, 2)1                                          -14.75756    3.17608
## poly(n.experience, 2)2                                            0.39121    2.85354
## education_levelHigh School                                       -1.31802    0.16475
## education_levelMastersPhd                                        -0.26115    0.11179
## education_levelPrimary School                                    -1.88282    0.43061
## company_sizeUnknown                                               0.39811    0.17016
## f.city_development_index[0.691,0.878)                            -2.00154    0.16743
## f.city_development_index[0.878,0.920)                            -2.38832    0.18318
## f.city_development_index[0.921,0.949]                            -2.52068    0.22452
## f.city_development_index0.920                                    -2.50964    0.16767
## company_sizeUnknown:f.city_development_index[0.691,0.878)         1.04754    0.25453
## company_sizeUnknown:f.city_development_index[0.878,0.920)         0.74709    0.28744
## company_sizeUnknown:f.city_development_index[0.921,0.949]         0.81220    0.35769
## company_sizeUnknown:f.city_development_index0.920                 2.47892    0.24958
##                                                                  z value Pr(>|z|)
## (Intercept)                                                        2.428  0.01519 *
## poly(n.experience, 2)1                                            -4.646 3.38e-06 ***
## poly(n.experience, 2)2                                             0.137  0.89095
## education_levelHigh School                                        -8.000 1.24e-15 ***
## education_levelMastersPhd                                         -2.336  0.01949 *
## education_levelPrimary School                                     -4.372 1.23e-05 ***
## company_sizeUnknown                                                2.340  0.01930 *
## f.city_development_index[0.691,0.878)                            -11.955  < 2e-16 ***
## f.city_development_index[0.878,0.920)                            -13.038  < 2e-16 ***
## f.city_development_index[0.921,0.949]                            -11.227  < 2e-16 ***
```

```
## f.city_development_index0.920                                -14.968  < 2e-16 ***
## company_sizeUnknown:f.city_development_index[0.691,0.878)   4.116 3.86e-05 ***
## company_sizeUnknown:f.city_development_index[0.878,0.920)   2.599  0.00935 **
## company_sizeUnknown:f.city_development_index[0.921,0.949]   2.271  0.02317 *
## company_sizeUnknown:f.city_development_index0.920           9.932  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4159.7  on 3700  degrees of freedom
## Residual deviance: 3202.4  on 3686  degrees of freedom
## AIC: 3232.4
##
## Number of Fisher Scoring iterations: 5
```

```
coef(mb)
```

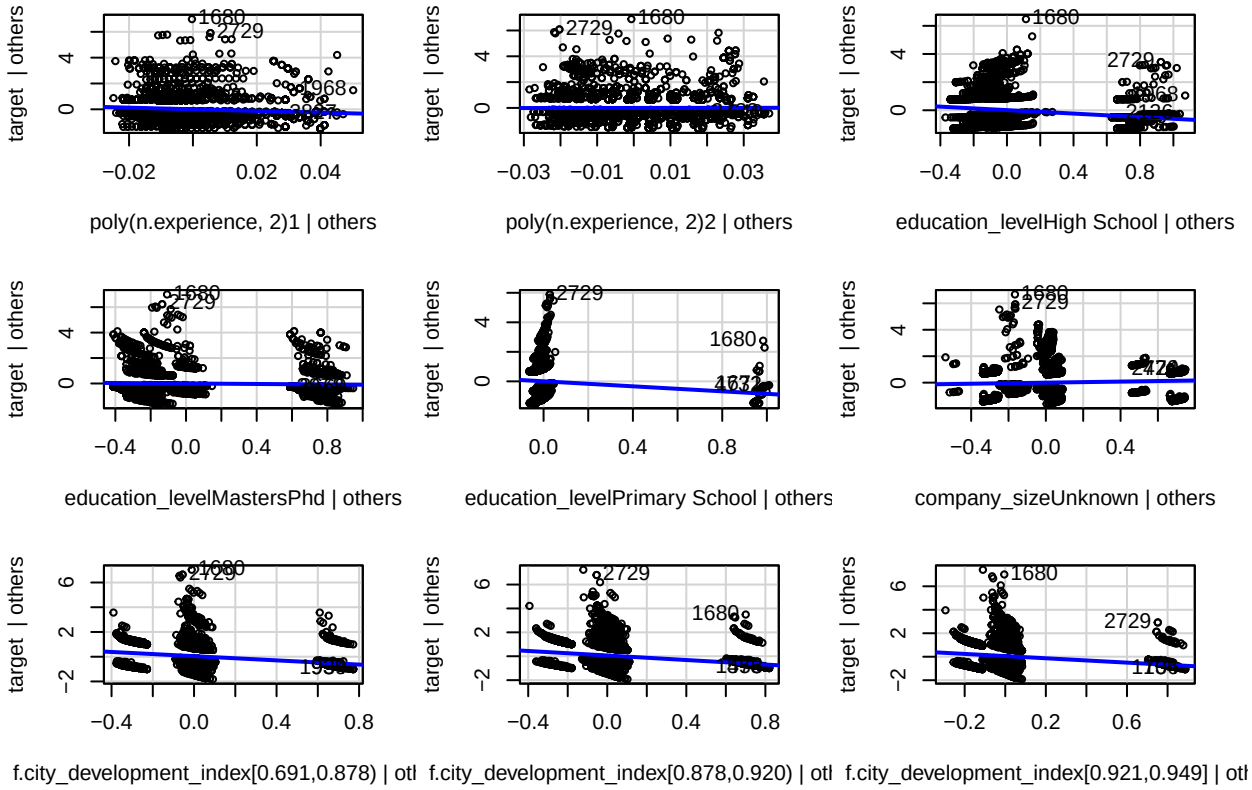
```
##                                                (Intercept)
##                                                  0.2411406
##                                     poly(n.experience, 2)1
##                                                -14.7575586
##                                     poly(n.experience, 2)2
##                                                  0.3912115
##                                  education_levelHigh School
##                                                 -1.3180183
##                                   education_levelMastersPhd
##                                                 -0.2611538
##                               education_levelPrimary School
##                                                 -1.8828246
##                                         company_sizeUnknown
##                                                  0.3981144
##                       f.city_development_index[0.691,0.878)
##                                                 -2.0015402
##                       f.city_development_index[0.878,0.920)
##                                                 -2.3883160
##                       f.city_development_index[0.921,0.949]
##                                                 -2.5206811
##                                f.city_development_index0.920
##                                                 -2.5096351
## company_sizeUnknown:f.city_development_index[0.691,0.878)
##                                                  1.0475353
## company_sizeUnknown:f.city_development_index[0.878,0.920)
##                                                  0.7470938
## company_sizeUnknown:f.city_development_index[0.921,0.949]
##                                                  0.8121991
##         company_sizeUnknown:f.city_development_index0.920
##                                                  2.4789165
```

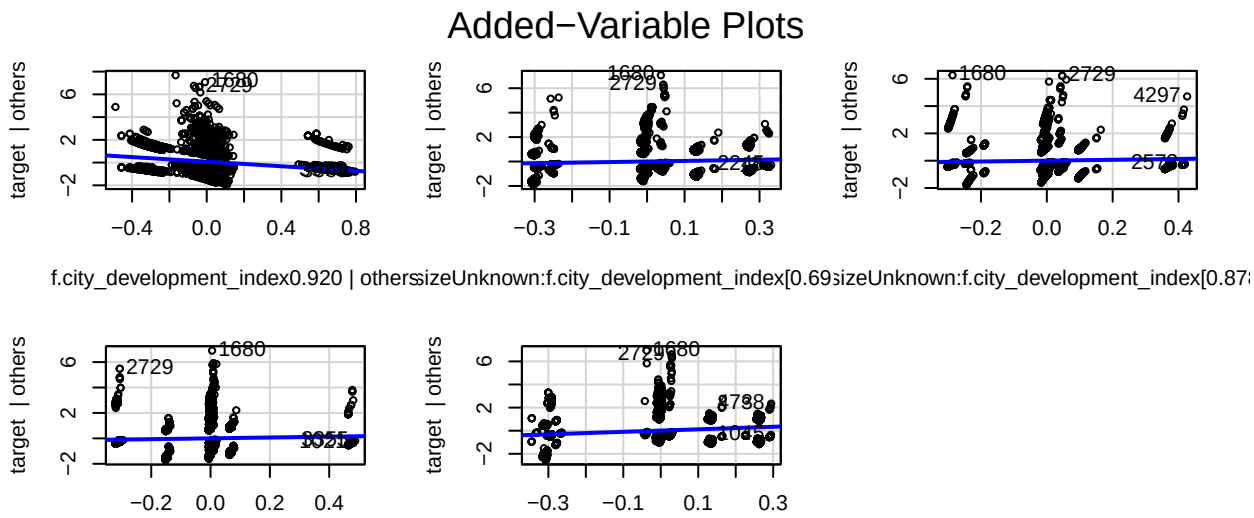
Model Diagnostics

The added variable plots, which can bee seen in the Annex, do not show any alarming behavior and the same can be said for the MarginalModelPlots. Regarding the residualPlots, there is not much to be said as

overall the behavior is acceptable. In the allEffects plots, it can be observed how the probabilities of wanting to change jobs decreses with experience, increases for higher levels of education, and is sligthly higher when the company information is not provided. There is a clear difference in developed countries where company information is provided or not.

```
avPlots(mb)
```

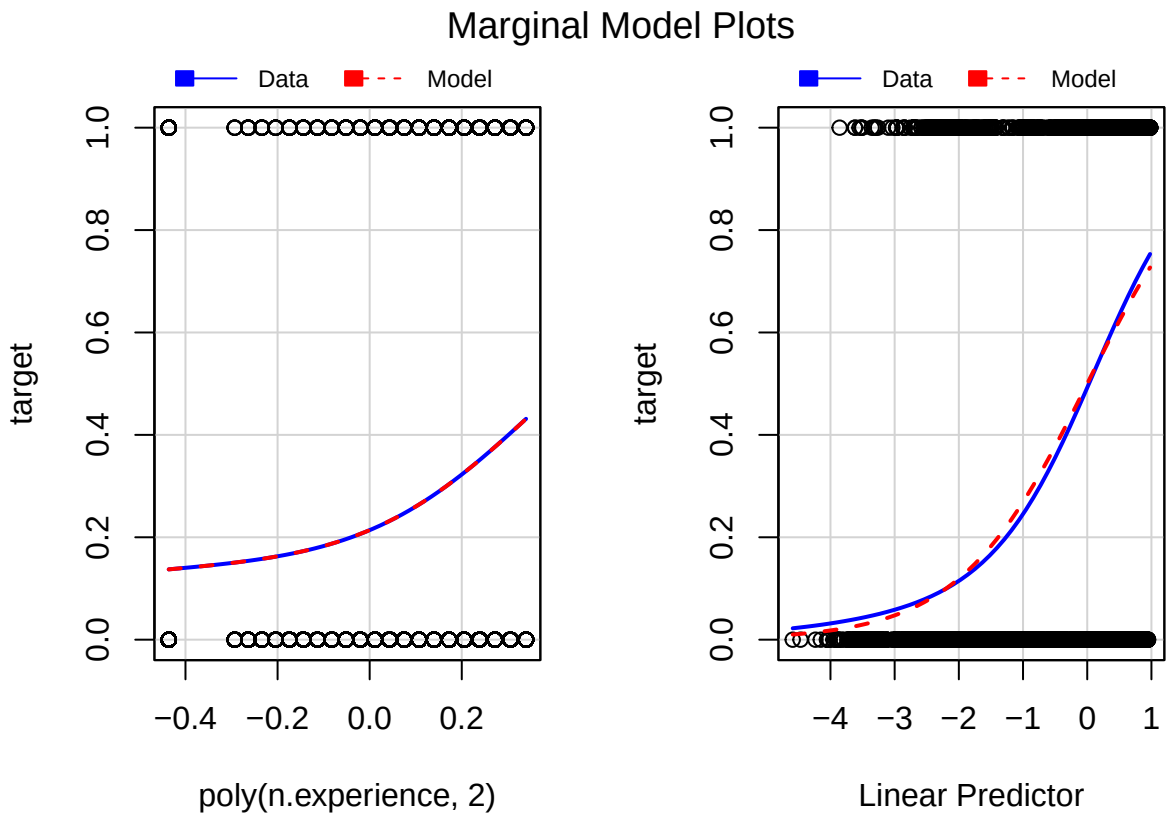





```
marginalModelPlots(mb)
```

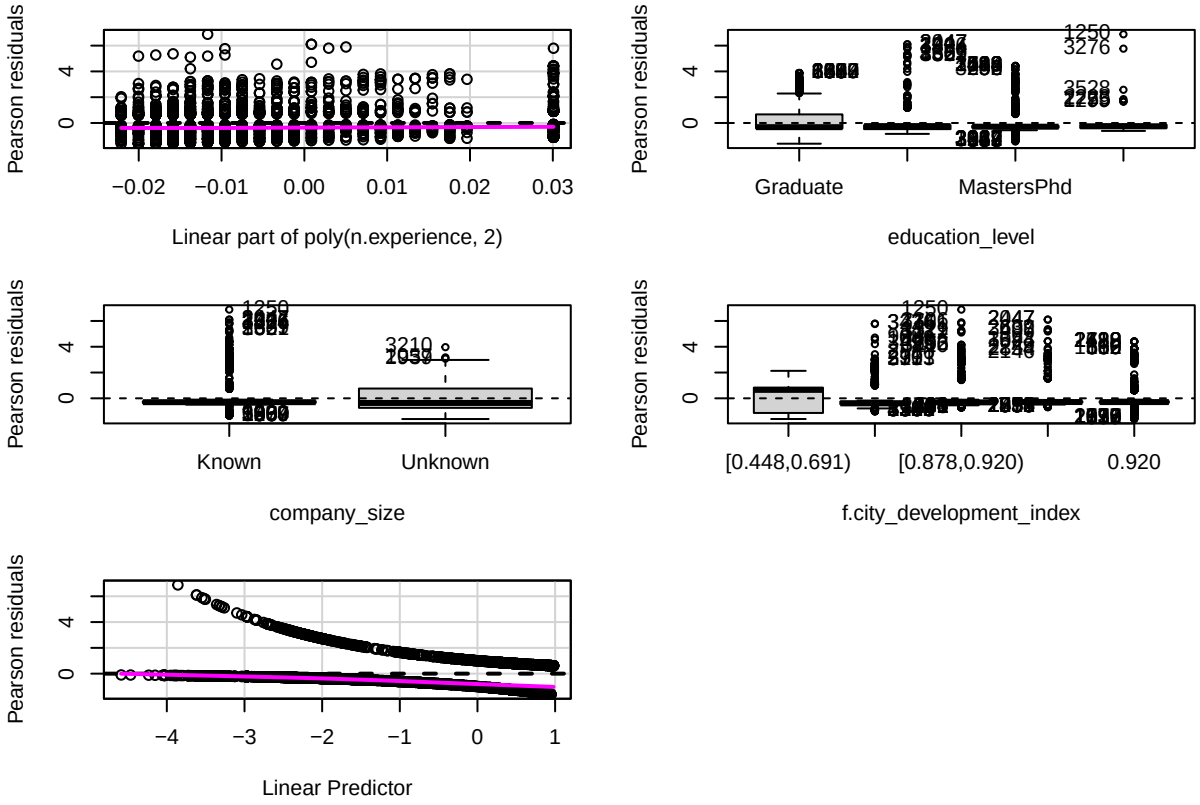
```
## Warning in mmps(...): Splines and/or polynomials replaced by a fitted linear
## combination
```

```
## Warning in mmps(...): Interactions and/or factors skipped
```

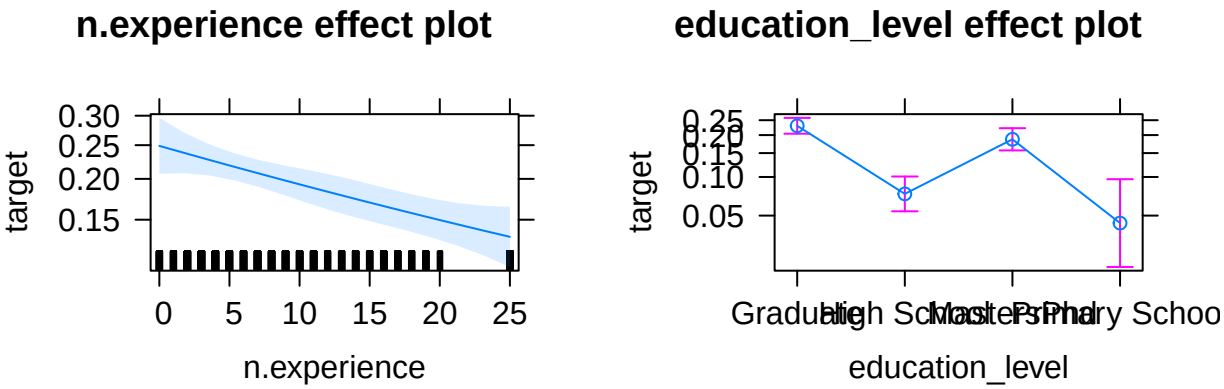


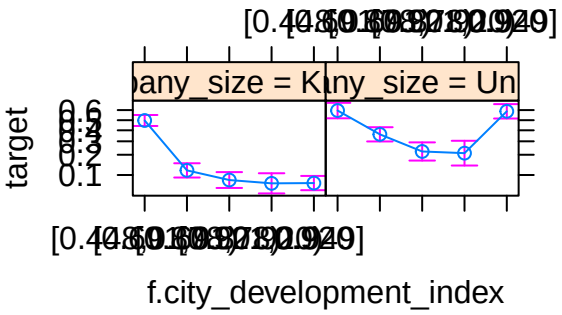

```
residualPlots(mb)
```

```
## Warning in residualPlots.default(model, ...): No possible lack-of-fit tests
```



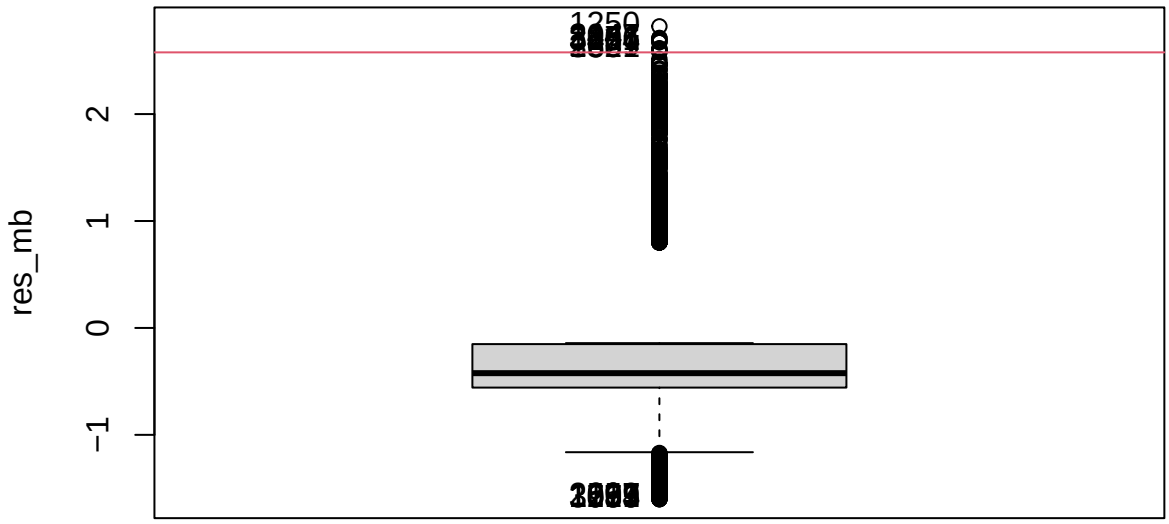

```
plot(allEffects(mb))
```





Studentized Residuals: Some outliers have been found in them, but they do not have,in any case, a value greater than 3. It can be seen that they are all people who want to change jobs, and contrary to the whole dataset, the vast majority of them are only high school graduates (hence no major). Also, all of them had specified the company size. Moreover, when performing an outlierTest only one observation is detected, but it will also be detected and removed when assessing the cook's distance.

```
n = dim(df)[1]
p = mb$rank
res_mb = rstudent(mb)
cut_off = qt(0.995,n-p-1)

ls = Boxplot(res_mb)
abline(h=cut_off,col=2)
abline(h=-cut_off,col=2)
```




```
nrow(df[which(abs(res_mb)>cut_off),])
```

```
## [1] 12
```

```
aux = df[which(abs(res_mb)>cut_off),]
summary(aux)
```

```
##     gender               relevent_experience       enrolled_university
##  Female:0   Has relevent experience:7        Full time course:1
##  Male  :8   No relevent experience :5        no_enrollment   :7
##  Other :4                                    Part time course:4
##
##
##
##        education_level        major_discipline  company_size
##  Graduate      : 0     Arts           : 0      Known  :12
##  High School   :10     Business Degree: 0      Unknown: 0
##  MastersPhd    : 0     Humanities     : 0
##  Primary School: 2     No Major       :12
##                        Other          : 0
##                        STEM           : 0
##               company_type last_new_job training_hours   target  imputed
##  Early Stage Startup:1     >4   :2      Min.   : 25.00   0: 0   Mode :logical
##  Funded Startup     :0     1    :5      1st Qu.: 35.75   1:12   FALSE:12
##  NGO                :0     2    :3      Median : 67.50
```

```
##  Other              :0     3    :1      Mean   : 65.08
##  Public Sector      :2     4    :1      3rd Qu.: 95.00
##  Pvt Ltd            :9     never:0      Max.   :105.00
##   n.experience     f.city_development_index
##  Min.   : 1.000   [0.448,0.691):0
##  1st Qu.: 4.500   [0.691,0.878):2
##  Median : 6.000   [0.878,0.920):5
##  Mean   : 8.333   [0.921,0.949]:5
##  3rd Qu.:11.250   0.920        :0
##  Max.   :25.000
```

```
outlierTest(mb) # The outlier is already taken into account in the cooks distance
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 1680 2.820526          0.0047945           NA
```

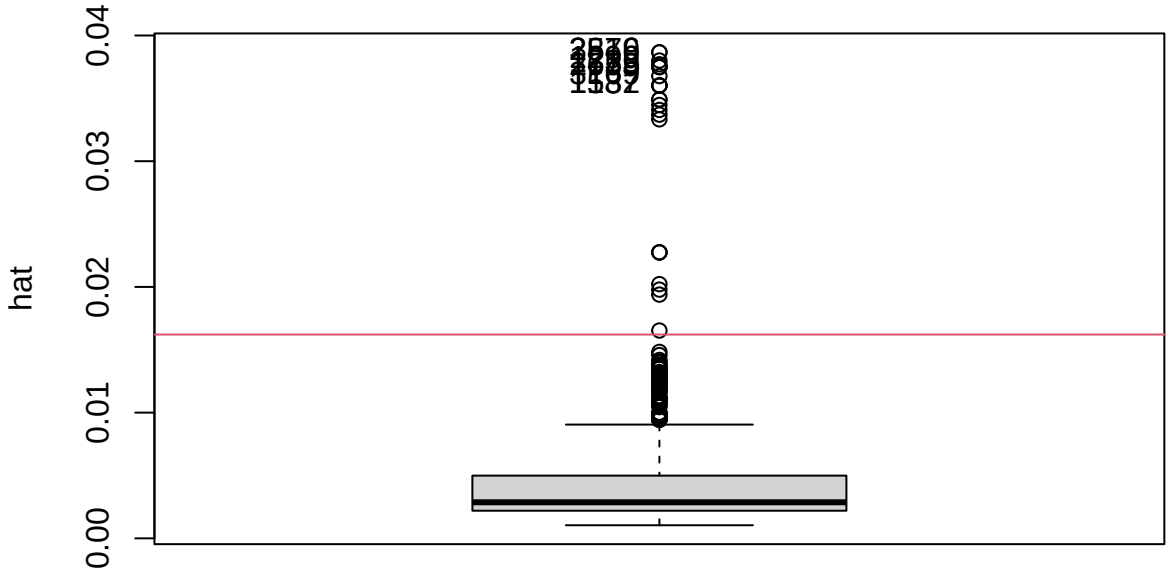
```
# prop.table(table(aux$gender)); prop.table(table(df$gender))
# prop.table(table(aux$relevent_experience)); prop.table(table(df$relevent_experience))
# prop.table(table(aux$enrolled_university)); prop.table(table(df$enrolled_university))
# prop.table(table(aux$education_level)); prop.table(table(df$education_level))
# prop.table(table(aux$major_discipline)); prop.table(table(df$major_discipline))
# prop.table(table(aux$company_size)); prop.table(table(df$company_size))
# prop.table(table(aux$company_type)); prop.table(table(df$company_type))
# prop.table(table(aux$last_new_job)); prop.table(table(df$last_new_job))
# prop.table(table(aux$training_hours)); prop.table(table(df$training_hours))
# prop.table(table(aux$n.experience)); prop.table(table(df$n.experience))
# prop.table(table(aux$f.city_development_index)); prop.table(table(df$f.city_development_index))
# prop.table(table(aux$target)); prop.table(table(df$target))
```

Hat Values: The cut off for this assessment has been 4 times the mean,as the dataset can be considered big enough. Regarding the description of the observations that fall under the criterion, there is a large proportion of people with no experience, with only primary school education (hence no major) and in this case all of them have not specified the company. As the had values indicate the leverage, these outliers have not been removed as the overall effect will be assessed with Cook's distance, taking into account discrepancy.

```
hat = hatvalues(mb)
hat_cut = 4*p/n

Boxplot(hat)
```

```
##  [1] 2876 3519 1299 1335   55 1679 2775 3169 1132 1587
```

```
abline(h=hat_cut,col=2)
```




```
sum(hat>hat_cut)
```

```
## [1] 25
```

```
aux = df[which(hat>hat_cut),]
summary(aux)
```

```
##     gender                relevent_experience       enrolled_university
##  Female: 1   Has relevent experience: 1       Full time course: 5
##  Male  :15   No relevent experience :24       no_enrollment   :19
##  Other : 9                                    Part time course: 1
##
##
##
##        education_level        major_discipline  company_size
##  Graduate      : 1     Arts           : 0      Known  : 0
##  High School   : 0     Business Degree: 0      Unknown:25
##  MastersPhd    : 0     Humanities     : 0
##  Primary School:24     No Major       :24
##                        Other          : 0
##                        STEM           : 1
##               company_type last_new_job training_hours   target  imputed
##  Early Stage Startup: 0    >4   : 0     Min.   :  6.00   0:20   Mode :logical
##  Funded Startup     : 0    1    : 2     1st Qu.: 17.00   1: 5   FALSE:22
##  NGO                : 0    2    : 2     Median : 25.00          TRUE :3
```

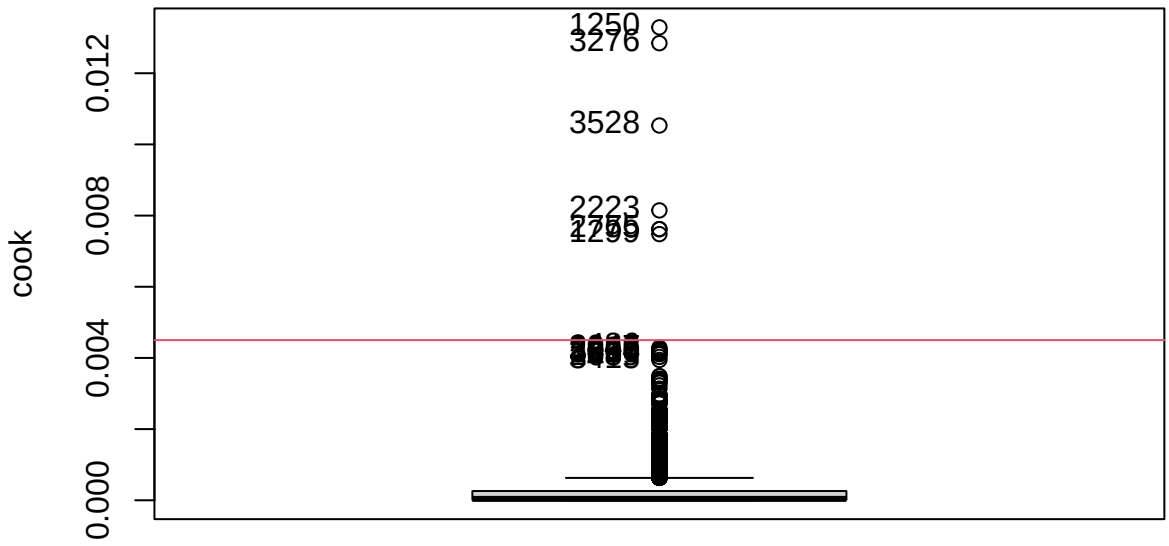
```
##  Other              :23    3    : 0     Mean   : 51.24
##  Public Sector      : 0    4    : 0     3rd Qu.: 67.00
##  Pvt Ltd            : 2    never:21     Max.   :210.00
##   n.experience  f.city_development_index
##  Min.   :0.0   [0.448,0.691): 6
##  1st Qu.:2.0   [0.691,0.878): 6
##  Median :3.0   [0.878,0.920): 0
##  Mean   :3.8   [0.921,0.949]: 1
##  3rd Qu.:5.0   0.920        :12
##  Max.   :9.0
```

Cook's distance: For the Cook's distance criterion, a threshold had to be defined to match the need of our model as a group of observations can clearly be seen as outlier far from the main group of observations. As before, it can be seen that the proportions for people with no experience and with primary school education. Moreover, all of them want to change jobs. Fortunately, none of the resulting influential observations is one of the ones that were imputed in the previous steps.

```
cook = cooks.distance(mb)
lc = Boxplot(cook, id=list(n=18))
cook_cut = 0.0045
nrow(df[which(cook>cook_cut),])
```

```
## [1] 7
```

```
abline(h=cook_cut, col=2)
```

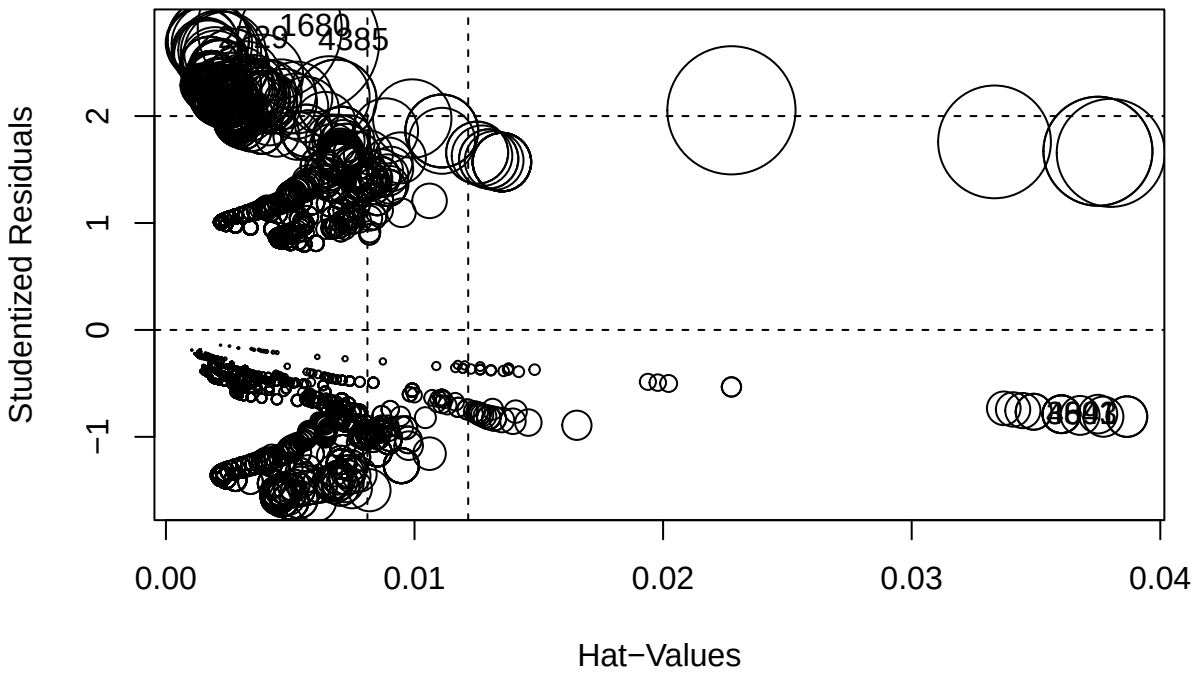


```
aux = df[which(cook>cook_cut),]
summary(aux)
```

```
##     gender               relevent_experience       enrolled_university
##  Female:0   Has relevent experience:1        Full time course:2
##  Male  :5   No relevent experience :6        no_enrollment   :5
##  Other :2                                    Part time course:0
##
##
##
##        education_level        major_discipline  company_size
##  Graduate      :0      Arts           :0       Known  :2
##  High School   :0      Business Degree:0       Unknown:5
##  MastersPhd    :0      Humanities     :0
##  Primary School:7      No Major       :7
##                        Other          :0
##                        STEM           :0
##               company_type last_new_job training_hours   target  imputed
##  Early Stage Startup:0     >4   :0      Min.   :  6.00   0:0    Mode :logical
##  Funded Startup     :0     1    :1      1st Qu.: 21.00   1:7    FALSE:7
##  NGO                :0     2    :0      Median : 32.00
##  Other              :5     3    :0      Mean   : 64.43
##  Public Sector      :0     4    :1      3rd Qu.: 80.50
##  Pvt Ltd            :2     never:5      Max.   :210.00
##   n.experience    f.city_development_index
##  Min.   :1.000   [0.448,0.691):3
##  1st Qu.:3.500   [0.691,0.878):2
##  Median :4.000   [0.878,0.920):1
##  Mean   :4.571   [0.921,0.949]:0
##  3rd Qu.:5.500   0.920        :1
##  Max.   :9.000
```

The influential data can be clearly seen with the help of an influence plot:

```
influencePlot(mb)
```




```
##          StudRes         Hat       CookD
## 1680  2.8205255 0.004174910 0.013289045
## 2729  2.7114695 0.001701840 0.004241244
## 3843 -0.8106938 0.038658716 0.001055543
## 4385  2.6940306 0.005731646 0.012844534
## 4691 -0.8106938 0.038658716 0.001055543
```

Reevaluate the model

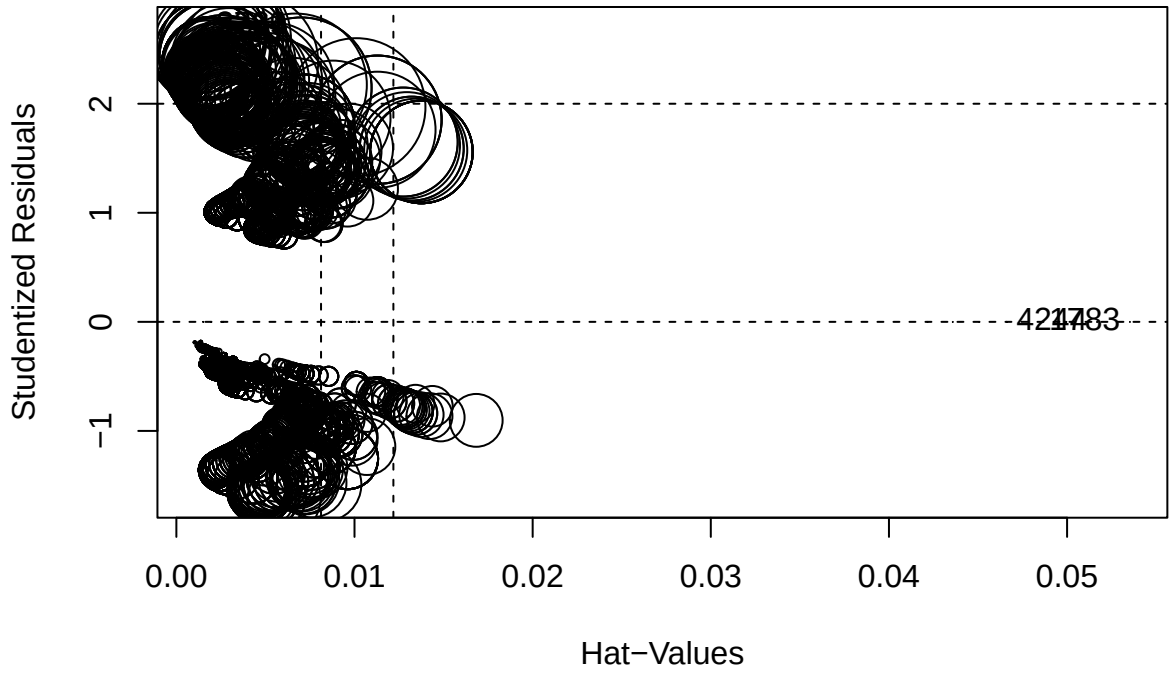
The outliers detected with the Cook's distance method have been removed from the dataset, and the model has been reevaluated without those observations. The two models have been evaluated firstly with AIC, knowing that it is not a strictly accurate comparison as the number of observations differs from one model to the other. Since it only differs by 7 observations some general intuition of the behaviour can be obtained. Thus, it can be seen that the new reevaluated model seems to be better, and the influencePlot results are much more accurate.

```
daux = df
df = df[which(cook<cook_cut),]

mbest = glm(formula = target ~ poly(n.experience, 2) + education_level +
    company_size + f.city_development_index + company_size:f.city_development_index,
    family = binomial, data = df)

AIC(mb, mbest)
```

```
## Warning in AIC.default(mb, mbest): models are not all fitted to the same number
## of observations
```

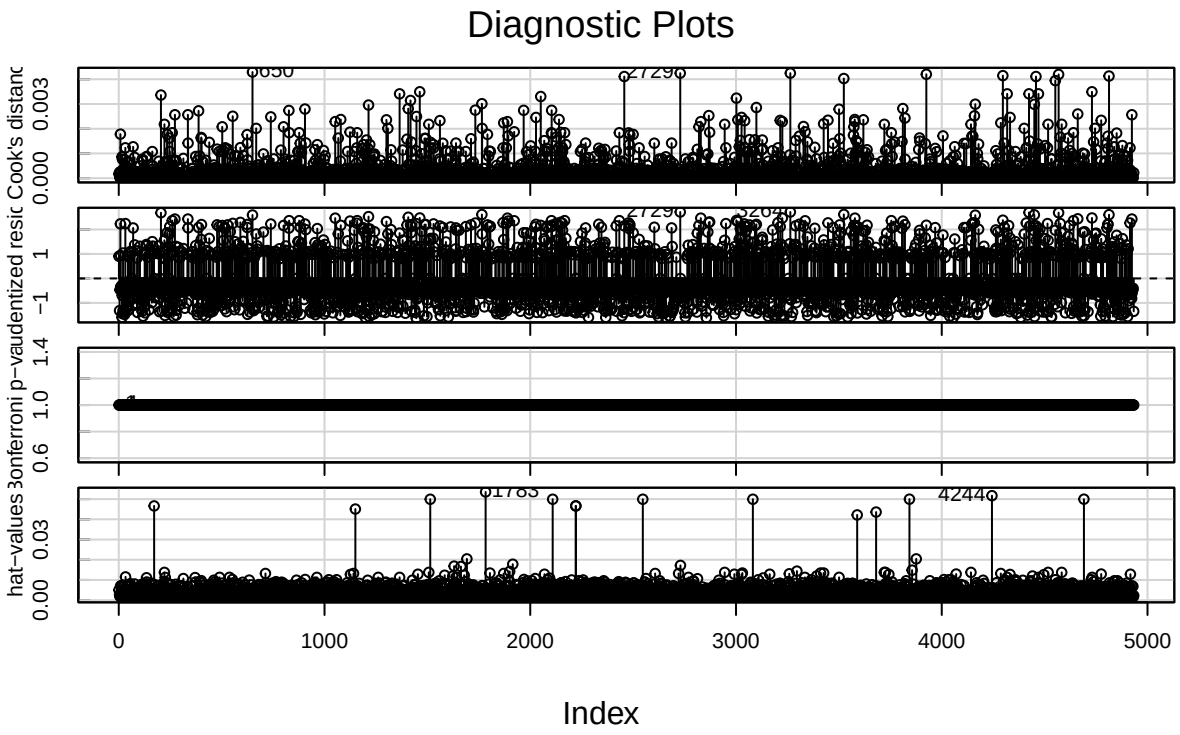
```
##       df      AIC
## mb    15 3232.367
## mbest 15 3189.407
```

```
influencePlot(mbest)
```




```
##            StudRes         Hat        CookD
## 650   2.5908367509 0.002393039 4.291953e-03
## 1783 -0.0006945885 0.053528433 9.345292e-10
## 2729  2.7135899915 0.001694517 4.247732e-03
## 3264  2.7135899915 0.001694517 4.247732e-03
## 4244 -0.0006821939 0.051686671 8.679453e-10
```

```
influenceIndexPlot(mbest)
```



Model Performance Evaluation

Now that the model has been improved, the ROC curve can be assessed to further diagnose its performance. The area under the curve (AUC) is computed for both models, which also serves as an indicator to compare them. First, the curve for the reevaluated model can be depicted, then the AUC's are displayed.

```
library(pROC)
```

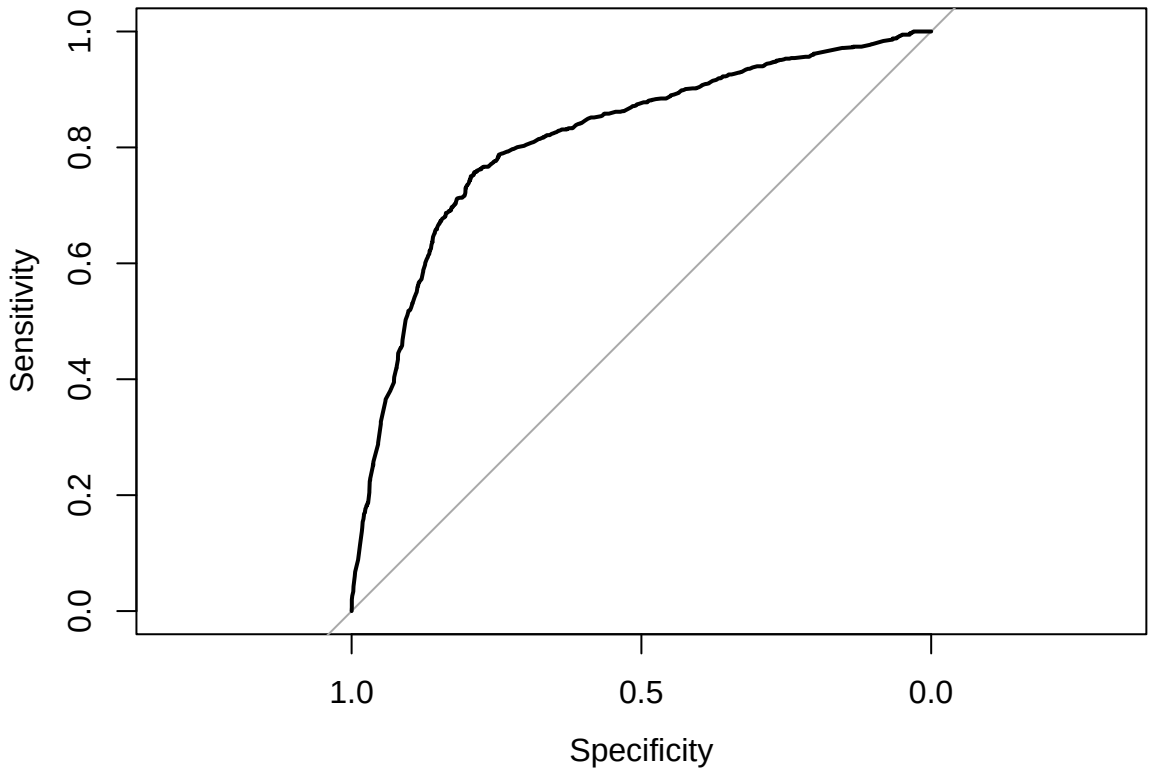
```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following object is masked from 'package:colorspace':
##
##     coords

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
prob=predict(mbest, type=c("response"))
df$prob=prob
g = roc(target ~ prob, data = df)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(g)
```



The AUC has been assessed for the working set, where a 0.5% improvement can be seen from the reevaluated model, which supports the previous conclusion that the removing the influential observations was beneficial. Regarding its value, 81.3% can be considered a very good model, even more so considering the imbalance in the response variable.

```
# Model
prob = predict(mb, type=c("response"))
auc(daux$target, prob)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Area under the curve: 0.8094
```

```
# Reevaluated model
prob = predict(mbest, type=c("response"))
auc(df$target, prob)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
## Area under the curve: 0.8133
```

Lastly, the confusion matrix is also assessed on the test set. First of all, the same transformations that have been done during the modelling steps have to be applied to the test set.

```
test = test %>%
  mutate(across(where(is.factor), ~ as.character(.))) %>%
  mutate(n.experience = case_when(experience == "<1" ~ "0",
                                  experience == ">20" ~ "25",
                                  TRUE ~ experience)) %>%
  mutate(n.experience = as.integer(n.experience)) %>%
  mutate(company_size = case_when(company_size != "Unknown" ~ "Known",
                                    TRUE ~ company_size)) %>%
  mutate(education_level = case_when(education_level %in% c("Masters", "Phd") ~ "MastersPhd",
                                    TRUE ~ education_level)) %>%
  mutate(across(where(is.character), ~ as.factor(.))) %>%
  select(., -c("experience"))

table(df$f.city_development_index)
```

```
##
## [0.448,0.691) [0.691,0.878) [0.878,0.920) [0.921,0.949]         0.920
##           732           718           749           500           995
```

```
test$f.city_development_index = as.ordered(cut2(test$city_development_index, cuts = c(0.691, 0.878, 0.9
table(test$f.city_development_index)
```

```
##
## [0.479,0.691) [0.691,0.878) [0.878,0.920)         0.920 [0.921,0.949]
##           269           250           250           298           165
```

```
levels(test$f.city_development_index) = c('[0.448,0.691)', '[0.691,0.878)', '[0.878,0.920)','0.920', '[
```

Then, the confusion matrix on the test set can be depicted in order to better assess the model. It can be observed how the model does not overfit the training data and it is much better than a random model (as already seen with AUC). Notice how the positive response (1) is in this case the negative one, and vice versa. As such, the specificity indicates how well it is being predicted that someone will change its job (0.52). This is decent accounting for the fact that the target is imbalanced. The measures related to the person not wanting to change job are all good. Overall, the model has a good accuracy and balanced accuracy.

```
prob = predict(mbest, newdata = test, type = "response")
test$prob = ifelse(prob<0.5,0,1)
confusionMatrix(data = as.factor(test$prob), reference = test$target)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 787 149
##          1 138 158
##
##                Accuracy : 0.767
##                  95% CI : (0.7424, 0.7904)
##     No Information Rate : 0.7508
##     P-Value [Acc > NIR] : 0.09885
##
##                   Kappa : 0.3699
##
##  Mcnemar's Test P-Value : 0.55500
##
##             Sensitivity : 0.8508
##             Specificity : 0.5147
##          Pos Pred Value : 0.8408
##          Neg Pred Value : 0.5338
##              Prevalence : 0.7508
##          Detection Rate : 0.6388
##    Detection Prevalence : 0.7597
##       Balanced Accuracy : 0.6827
##
##        'Positive' Class : 0
##
```

Continuing with the predictive power of the model, a Hoslem test has been run and the null hypothesis has been clearly rejected, stating that the model does not fit well the data. Regarding some Pseduo R^2 metrics, which have to be assessed with caution, they are not very promising as well. All of this results could well be from the

```
library(ResourceSelection)
```

```
## Warning: package 'ResourceSelection' was built under R version 4.1.2
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
hoslem.test(test$target, prob)
```

```
## Warning in Ops.factor(1, y): '-' not meaningful for factors
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  test$target, prob
## X-squared = 1232, df = 8, p-value < 2.2e-16
```

```
prob = predict(mbest, newdata = test, type = "response")
test$prob = ifelse(prob<0.5,0,1)
confusionMatrix(data = as.factor(test$prob), reference = test$target)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 787 149
##          1 138 158
##
##                Accuracy : 0.767
##                  95% CI : (0.7424, 0.7904)
##     No Information Rate : 0.7508
##     P-Value [Acc > NIR] : 0.09885
##
##                   Kappa : 0.3699
##
##  Mcnemar's Test P-Value : 0.55500
##
##             Sensitivity : 0.8508
##             Specificity : 0.5147
##          Pos Pred Value : 0.8408
##          Neg Pred Value : 0.5338
##              Prevalence : 0.7508
##          Detection Rate : 0.6388
##    Detection Prevalence : 0.7597
##       Balanced Accuracy : 0.6827
##
##        'Positive' Class : 0
##
```

To wrap up, the model has been used to predict the most representative individual (i.e experience on the mean, Graduate, having reported the company size and from a very developed city). It can be seen that the model predicts very strongly (0.09) that he/she is not going to change jobs. Alternatively, it can be seen that by maintaining the same parameters for the individual and changing its city to a not very developed one, the prediction from the model gets over 0.55, a very significant increase, as was found when exploring the model equations.

```
newdata = data.frame(n.experience = c(mean(df$n.experience),mean(df$n.experience)), education_level = c

predict(mbest,newdata,type='response')
```

```
##          1          2
## 0.09306754 0.55686657
```

Overall, the project has proved the importance of the initial data treatment and how the decisions made by the data scientist that face the problems affect the final outcome of the model. It helped to consolidate the general workflow of creating a model from scratch and the assessments that have to be done in each of the steps.