Projektbericht

Ferdinand Bubeck und Johannes Bubeck

19.01.2021

## Data Exploration

#import libraries  
library(readr)  
library(dplyr)  
library(ggplot2)  
library(tidyverse)  
  
#import CSVs  
datTest <- read\_csv("aug\_test.csv")  
  
datTrain <- read\_csv("aug\_train.csv")  
  
#join of Train & Test Data to have full Dataset for Exploration  
dat <- full\_join(datTest, datTrain)   
  
#first Impressions of Data  
head(dat)

## # A tibble: 6 x 14  
## enrollee\_id city city\_developmen~ gender relevent\_experi~ enrolled\_univer~  
## <dbl> <chr> <dbl> <chr> <chr> <chr>   
## 1 32403 city~ 0.827 Male Has relevent ex~ Full time course  
## 2 9858 city~ 0.92 Female Has relevent ex~ no\_enrollment   
## 3 31806 city~ 0.624 Male No relevent exp~ no\_enrollment   
## 4 27385 city~ 0.827 Male Has relevent ex~ no\_enrollment   
## 5 27724 city~ 0.92 Male Has relevent ex~ no\_enrollment   
## 6 217 city~ 0.899 Male No relevent exp~ Part time course  
## # ... with 8 more variables: education\_level <chr>, major\_discipline <chr>,  
## # experience <chr>, company\_size <chr>, company\_type <chr>,  
## # last\_new\_job <chr>, training\_hours <dbl>, target <dbl>

tail(dat)

## # A tibble: 6 x 14  
## enrollee\_id city city\_developmen~ gender relevent\_experi~ enrolled\_univer~  
## <dbl> <chr> <dbl> <chr> <chr> <chr>   
## 1 29754 city~ 0.92 Female Has relevent ex~ no\_enrollment   
## 2 7386 city~ 0.878 Male No relevent exp~ no\_enrollment   
## 3 31398 city~ 0.92 Male Has relevent ex~ no\_enrollment   
## 4 24576 city~ 0.92 Male Has relevent ex~ no\_enrollment   
## 5 5756 city~ 0.802 Male Has relevent ex~ no\_enrollment   
## 6 23834 city~ 0.855 <NA> No relevent exp~ no\_enrollment   
## # ... with 8 more variables: education\_level <chr>, major\_discipline <chr>,  
## # experience <chr>, company\_size <chr>, company\_type <chr>,  
## # last\_new\_job <chr>, training\_hours <dbl>, target <dbl>

summary(dat)

## enrollee\_id city city\_development\_index gender   
## Min. : 1 Length:21287 Min. :0.4480 Length:21287   
## 1st Qu.: 8554 Class :character 1st Qu.:0.7390 Class :character   
## Median :16967 Mode :character Median :0.9030 Mode :character   
## Mean :16874 Mean :0.8285   
## 3rd Qu.:25162 3rd Qu.:0.9200   
## Max. :33380 Max. :0.9490   
##   
## relevent\_experience enrolled\_university education\_level major\_discipline   
## Length:21287 Length:21287 Length:21287 Length:21287   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## experience company\_size company\_type last\_new\_job   
## Length:21287 Length:21287 Length:21287 Length:21287   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## training\_hours target   
## Min. : 1.00 Min. :0.0000   
## 1st Qu.: 23.00 1st Qu.:0.0000   
## Median : 47.00 Median :0.0000   
## Mean : 65.33 Mean :0.2493   
## 3rd Qu.: 88.00 3rd Qu.:0.0000   
## Max. :336.00 Max. :1.0000   
## NA's :2129

dim(dat)

## [1] 21287 14

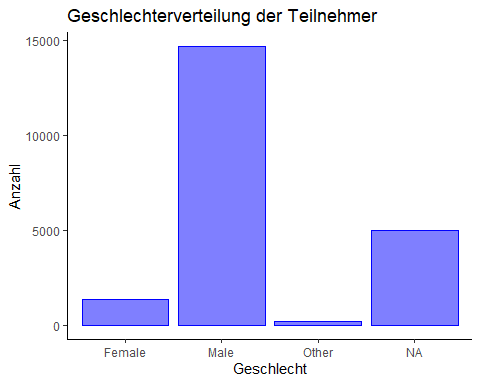
str(dat)

## tibble [21,287 x 14] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ enrollee\_id : num [1:21287] 32403 9858 31806 27385 27724 ...  
## $ city : chr [1:21287] "city\_41" "city\_103" "city\_21" "city\_13" ...  
## $ city\_development\_index: num [1:21287] 0.827 0.92 0.624 0.827 0.92 0.899 0.624 0.92 0.878 0.624 ...  
## $ gender : chr [1:21287] "Male" "Female" "Male" "Male" ...  
## $ relevent\_experience : chr [1:21287] "Has relevent experience" "Has relevent experience" "No relevent experience" "Has relevent experience" ...  
## $ enrolled\_university : chr [1:21287] "Full time course" "no\_enrollment" "no\_enrollment" "no\_enrollment" ...  
## $ education\_level : chr [1:21287] "Graduate" "Graduate" "High School" "Masters" ...  
## $ major\_discipline : chr [1:21287] "STEM" "STEM" NA "STEM" ...  
## $ experience : chr [1:21287] "9" "5" "<1" "11" ...  
## $ company\_size : chr [1:21287] "<10" NA NA "10/49" ...  
## $ company\_type : chr [1:21287] NA "Pvt Ltd" "Pvt Ltd" "Pvt Ltd" ...  
## $ last\_new\_job : chr [1:21287] "1" "1" "never" "1" ...  
## $ training\_hours : num [1:21287] 21 98 15 39 72 12 11 81 2 4 ...  
## $ target : num [1:21287] NA NA NA NA NA NA NA NA NA NA ...

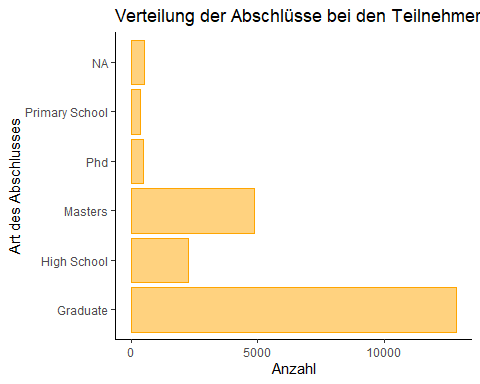
#count NAs in column gender  
sum(is.na(dat$gender))

## [1] 5016

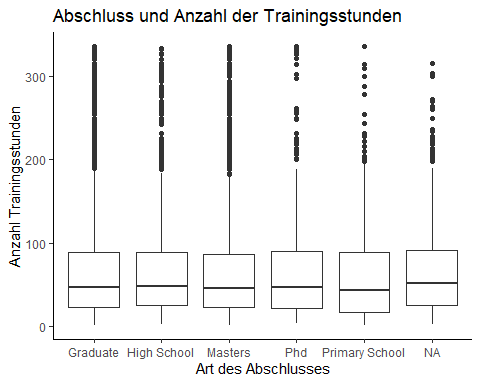
#Plot zu Geschlechterverteilung  
ggplot(data=dat, aes(x=gender))+  
 geom\_bar(color='blue', fill='blue', alpha=.5)+  
 theme\_classic()+  
 labs(title="Geschlechterverteilung der Teilnehmer")+  
 xlab("Geschlecht")+  
 ylab("Anzahl")



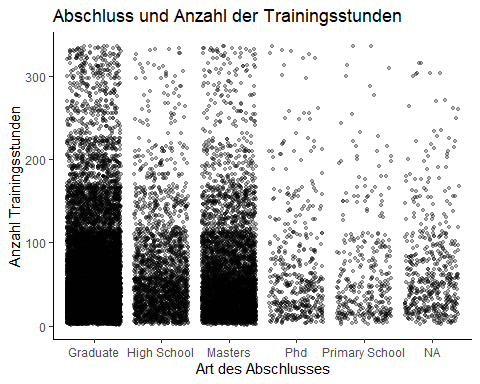
#Plot der Abschlüsse  
ggplot(data=dat, aes(x=education\_level))+  
 geom\_bar(color='orange', fill='orange', alpha=.5)+  
 coord\_flip()+  
 theme\_classic()+  
 labs(title="Verteilung der Abschlüsse bei den Teilnehmer")+  
 xlab("Art des Abschlusses")+  
 ylab("Anzahl")



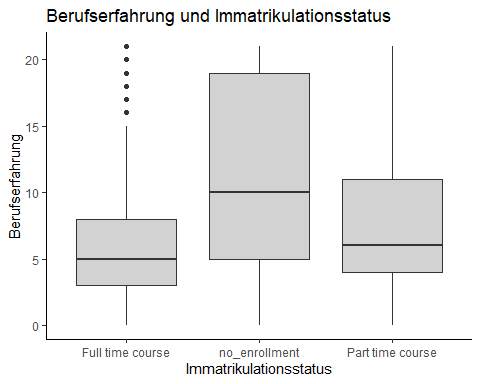
#Plot Anzahl der Trainngsstunden und Abschluss (Boxplot)  
ggplot(data=dat, aes(x=education\_level, y=training\_hours))+  
 geom\_boxplot()+  
 theme\_classic()+  
 labs(title="Abschluss und Anzahl der Trainingsstunden")+  
 xlab("Art des Abschlusses")+  
 ylab("Anzahl Trainingsstunden")



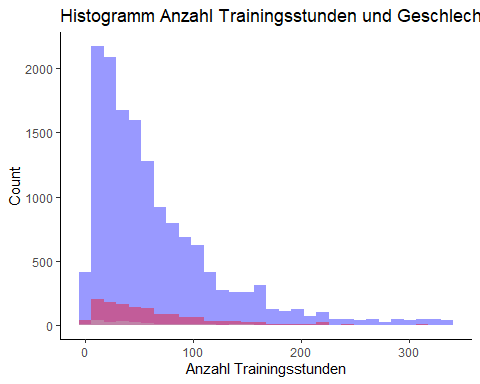
#Plot Anzahl der Trainngsstunden und Abschluss (Point/Jitter)  
ggplot(data=dat, aes(x=education\_level, y=training\_hours))+  
 geom\_jitter(color='black', alpha=.3, size=.9)+  
 theme\_classic()+  
 labs(title="Abschluss und Anzahl der Trainingsstunden")+  
 xlab("Art des Abschlusses")+  
 ylab("Anzahl Trainingsstunden")



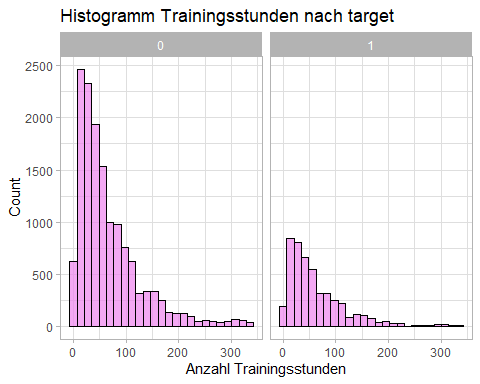
#ersetze < und > durch feste numerische Werte um typecast durchzuführen  
dat$experience <- sub("<1", "0", dat$experience)  
dat$experience <- sub(">20", "21", dat$experience)  
dat$experience <- as.numeric(dat$experience)  
  
dat %>%   
 select(enrolled\_university, experience) %>%  
 filter(!is.na(enrolled\_university),!is.na(experience)) %>%  
 ggplot(data=., aes(x=enrolled\_university, y=experience))+  
 geom\_boxplot(fill="grey", alpha=.7)+  
 theme\_classic()+  
 labs(title="Berufserfahrung und Immatrikulationsstatus")+  
 xlab("Immatrikulationsstatus")+  
 ylab("Berufserfahrung")



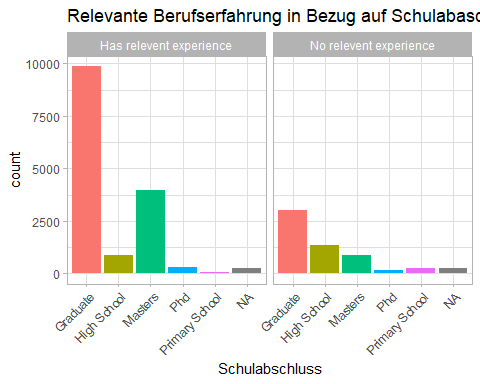
###  
ggplot()+  
 geom\_histogram(data=subset(dat, gender == "Male"), aes(x=training\_hours), fill="blue", alpha=.4)+  
 geom\_histogram(data=subset(dat, gender == "Female"), aes(x=training\_hours), fill="red", alpha=.4)+  
 geom\_histogram(data=subset(dat, gender == "Other"), aes(x=training\_hours), fill="grey", alpha=.4)+  
 theme\_classic()+  
 labs(title="Histogramm Anzahl Trainingsstunden und Geschlecht")+  
 xlab("Anzahl Trainingsstunden")+  
 ylab("Count")



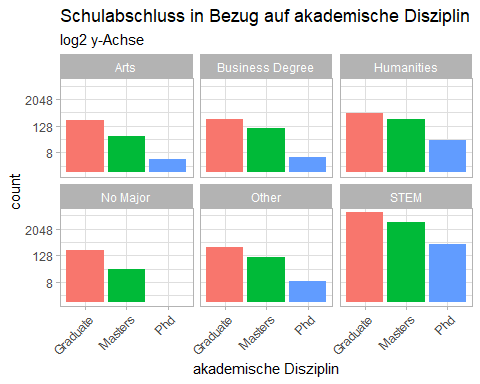
dat %>%  
 select(training\_hours, target) %>%  
 filter(!is.na(training\_hours),!is.na(target)) %>%  
 ggplot(data=., aes(x = training\_hours)) +  
 geom\_histogram(bins = 25, color = "black", fill = "violet", alpha=.7) +  
 theme\_light() +  
 facet\_wrap(vars(target))+  
 labs(title="Histogramm Trainingsstunden nach target")+  
 xlab("Anzahl Trainingsstunden")+  
 ylab("Count")



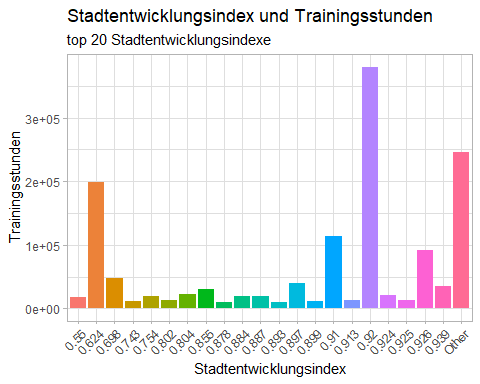
ggplot(data=dat, aes(x=as.factor(education\_level)))+  
 geom\_bar(data=dat, aes(fill=as.factor(education\_level)))+  
 facet\_wrap(vars(relevent\_experience))+  
 theme\_light()+  
 theme(legend.position = "none")+  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))+  
 labs(title="Relevante Berufserfahrung in Bezug auf Schulabaschluss")+  
 xlab("Schulabschluss")



dat %>%   
 select(major\_discipline, education\_level) %>%  
 filter(!is.na(major\_discipline),!is.na(education\_level)) %>%  
 ggplot(data=.,aes(x=education\_level))+  
 geom\_bar(aes(fill=education\_level))+  
 facet\_wrap(vars(major\_discipline))+  
 theme\_light()+  
 theme(legend.position = "none")+  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))+  
 labs(title="Schulabschluss in Bezug auf akademische Disziplin", subtitle = "log2 y-Achse")+  
 xlab("akademische Disziplin")+  
 scale\_y\_continuous(trans = 'log2')



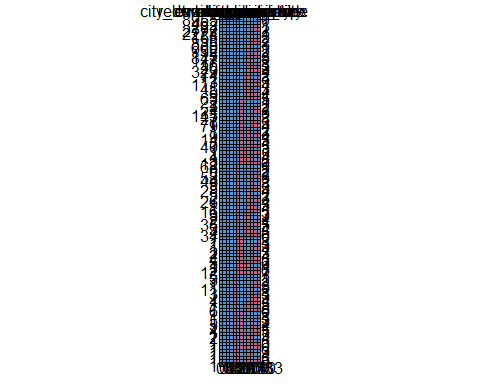
dat %>%   
 select(city\_development\_index, training\_hours)%>%  
 mutate(city\_development\_index = fct\_lump(as.factor(city\_development\_index), n = 20)) %>%  
 filter(!is.na(city\_development\_index)) %>%  
 ggplot(data= ., aes(x=city\_development\_index, y=training\_hours, fill = city\_development\_index))+  
 geom\_col()+  
 theme\_light()+  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))+  
 theme(legend.position = "none")+  
 labs(title="Stadtentwicklungsindex und Trainingsstunden", subtitle = "top 20 Stadtentwicklungsindexe")+  
 xlab("Stadtentwicklungsindex")+  
 ylab("Trainingsstunden")



library(caret)  
library(mice)  
  
#Dealing with NAs  
#detect variables with more than 5% (threshold) of NAs  
detectNA <- function(x){sum(is.na(x))/length(x)\*100}  
apply(datTrain, 2, detectNA)

## enrollee\_id city city\_development\_index   
## 0.0000000 0.0000000 0.0000000   
## gender relevent\_experience enrolled\_university   
## 23.5306399 0.0000000 2.0148241   
## education\_level major\_discipline experience   
## 2.4010857 14.6831611 0.3392839   
## company\_size company\_type last\_new\_job   
## 30.9948846 32.0492745 2.2079549   
## training\_hours target   
## 0.0000000 0.0000000

md.pattern(datTrain[,2:13])



## city city\_development\_index relevent\_experience training\_hours experience  
## 8955 1 1 1 1 1  
## 462 1 1 1 1 1  
## 283 1 1 1 1 1  
## 2777 1 1 1 1 1  
## 2224 1 1 1 1 1  
## 158 1 1 1 1 1  
## 98 1 1 1 1 1  
## 835 1 1 1 1 1  
## 660 1 1 1 1 1  
## 52 1 1 1 1 1  
## 115 1 1 1 1 1  
## 847 1 1 1 1 1  
## 177 1 1 1 1 1  
## 16 1 1 1 1 1  
## 26 1 1 1 1 1  
## 329 1 1 1 1 1  
## 74 1 1 1 1 1  
## 11 1 1 1 1 1  
## 13 1 1 1 1 1  
## 111 1 1 1 1 1  
## 45 1 1 1 1 1  
## 5 1 1 1 1 1  
## 9 1 1 1 1 1  
## 63 1 1 1 1 1  
## 22 1 1 1 1 1  
## 3 1 1 1 1 1  
## 4 1 1 1 1 1  
## 21 1 1 1 1 1  
## 115 1 1 1 1 1  
## 27 1 1 1 1 1  
## 7 1 1 1 1 1  
## 79 1 1 1 1 1  
## 1 1 1 1 1 1  
## 1 1 1 1 1 1  
## 9 1 1 1 1 1  
## 14 1 1 1 1 1  
## 3 1 1 1 1 1  
## 40 1 1 1 1 1  
## 1 1 1 1 1 1  
## 1 1 1 1 1 1  
## 4 1 1 1 1 1  
## 3 1 1 1 1 1  
## 12 1 1 1 1 1  
## 62 1 1 1 1 1  
## 9 1 1 1 1 1  
## 5 1 1 1 1 1  
## 53 1 1 1 1 1  
## 44 1 1 1 1 1  
## 9 1 1 1 1 1  
## 5 1 1 1 1 1  
## 28 1 1 1 1 1  
## 5 1 1 1 1 1  
## 2 1 1 1 1 1  
## 26 1 1 1 1 1  
## 1 1 1 1 1 1  
## 1 1 1 1 1 1  
## 16 1 1 1 1 1  
## 3 1 1 1 1 1  
## 2 1 1 1 1 1  
## 2 1 1 1 1 1  
## 35 1 1 1 1 1  
## 3 1 1 1 1 1  
## 4 1 1 1 1 1  
## 34 1 1 1 1 1  
## 1 1 1 1 1 1  
## 1 1 1 1 1 1  
## 1 1 1 1 1 1  
## 1 1 1 1 1 1  
## 2 1 1 1 1 1  
## 2 1 1 1 1 1  
## 4 1 1 1 1 1  
## 5 1 1 1 1 1  
## 1 1 1 1 1 1  
## 2 1 1 1 1 1  
## 12 1 1 1 1 1  
## 5 1 1 1 1 0  
## 3 1 1 1 1 0  
## 1 1 1 1 1 0  
## 1 1 1 1 1 0  
## 11 1 1 1 1 0  
## 3 1 1 1 1 0  
## 1 1 1 1 1 0  
## 4 1 1 1 1 0  
## 1 1 1 1 1 0  
## 1 1 1 1 1 0  
## 6 1 1 1 1 0  
## 1 1 1 1 1 0  
## 1 1 1 1 1 0  
## 5 1 1 1 1 0  
## 5 1 1 1 1 0  
## 1 1 1 1 1 0  
## 4 1 1 1 1 0  
## 1 1 1 1 1 0  
## 2 1 1 1 1 0  
## 1 1 1 1 1 0  
## 1 1 1 1 1 0  
## 1 1 1 1 1 0  
## 1 1 1 1 1 0  
## 1 1 1 1 1 0  
## 1 1 1 1 1 0  
## 1 1 1 1 1 0  
## 1 1 1 1 1 0  
## 0 0 0 0 65  
## enrolled\_university last\_new\_job education\_level major\_discipline gender  
## 8955 1 1 1 1 1  
## 462 1 1 1 1 1  
## 283 1 1 1 1 1  
## 2777 1 1 1 1 1  
## 2224 1 1 1 1 0  
## 158 1 1 1 1 0  
## 98 1 1 1 1 0  
## 835 1 1 1 1 0  
## 660 1 1 1 0 1  
## 52 1 1 1 0 1  
## 115 1 1 1 0 1  
## 847 1 1 1 0 1  
## 177 1 1 1 0 0  
## 16 1 1 1 0 0  
## 26 1 1 1 0 0  
## 329 1 1 1 0 0  
## 74 1 1 0 0 1  
## 11 1 1 0 0 1  
## 13 1 1 0 0 1  
## 111 1 1 0 0 1  
## 45 1 1 0 0 0  
## 5 1 1 0 0 0  
## 9 1 1 0 0 0  
## 63 1 1 0 0 0  
## 22 1 0 1 1 1  
## 3 1 0 1 1 1  
## 4 1 0 1 1 1  
## 21 1 0 1 1 1  
## 115 1 0 1 1 0  
## 27 1 0 1 1 0  
## 7 1 0 1 1 0  
## 79 1 0 1 1 0  
## 1 1 0 1 0 1  
## 1 1 0 1 0 1  
## 9 1 0 1 0 1  
## 14 1 0 1 0 0  
## 3 1 0 1 0 0  
## 40 1 0 1 0 0  
## 1 1 0 0 0 1  
## 1 1 0 0 0 1  
## 4 1 0 0 0 0  
## 3 1 0 0 0 0  
## 12 1 0 0 0 0  
## 62 0 1 1 1 1  
## 9 0 1 1 1 1  
## 5 0 1 1 1 1  
## 53 0 1 1 1 1  
## 44 0 1 1 1 0  
## 9 0 1 1 1 0  
## 5 0 1 1 1 0  
## 28 0 1 1 1 0  
## 5 0 1 1 0 1  
## 2 0 1 1 0 1  
## 26 0 1 1 0 1  
## 1 0 1 1 0 0  
## 1 0 1 1 0 0  
## 16 0 1 1 0 0  
## 3 0 1 0 0 1  
## 2 0 1 0 0 1  
## 2 0 1 0 0 1  
## 35 0 1 0 0 1  
## 3 0 1 0 0 0  
## 4 0 1 0 0 0  
## 34 0 1 0 0 0  
## 1 0 0 1 1 1  
## 1 0 0 1 1 1  
## 1 0 0 1 1 1  
## 1 0 0 1 1 1  
## 2 0 0 1 1 0  
## 2 0 0 1 1 0  
## 4 0 0 1 1 0  
## 5 0 0 1 0 0  
## 1 0 0 0 0 1  
## 2 0 0 0 0 1  
## 12 0 0 0 0 0  
## 5 1 1 1 1 1  
## 3 1 1 1 1 1  
## 1 1 1 1 1 1  
## 1 1 1 1 1 1  
## 11 1 1 1 1 0  
## 3 1 1 1 1 0  
## 1 1 1 1 1 0  
## 4 1 1 1 1 0  
## 1 1 1 1 0 0  
## 1 1 1 0 0 0  
## 6 1 1 0 0 0  
## 1 1 0 1 1 1  
## 1 1 0 1 1 1  
## 5 1 0 1 1 0  
## 5 1 0 1 1 0  
## 1 1 0 1 1 0  
## 4 1 0 1 1 0  
## 1 1 0 1 0 1  
## 2 1 0 1 0 0  
## 1 1 0 1 0 0  
## 1 1 0 1 0 0  
## 1 1 0 0 0 1  
## 1 0 1 1 1 0  
## 1 0 1 1 0 0  
## 1 0 1 0 0 1  
## 1 0 1 0 0 1  
## 1 0 0 1 1 0  
## 386 423 460 2813 4508  
## company\_size company\_type   
## 8955 1 1 0  
## 462 1 0 1  
## 283 0 1 1  
## 2777 0 0 2  
## 2224 1 1 1  
## 158 1 0 2  
## 98 0 1 2  
## 835 0 0 3  
## 660 1 1 1  
## 52 1 0 2  
## 115 0 1 2  
## 847 0 0 3  
## 177 1 1 2  
## 16 1 0 3  
## 26 0 1 3  
## 329 0 0 4  
## 74 1 1 2  
## 11 1 0 3  
## 13 0 1 3  
## 111 0 0 4  
## 45 1 1 3  
## 5 1 0 4  
## 9 0 1 4  
## 63 0 0 5  
## 22 1 1 1  
## 3 1 0 2  
## 4 0 1 2  
## 21 0 0 3  
## 115 1 1 2  
## 27 1 0 3  
## 7 0 1 3  
## 79 0 0 4  
## 1 1 1 2  
## 1 1 0 3  
## 9 0 0 4  
## 14 1 1 3  
## 3 1 0 4  
## 40 0 0 5  
## 1 1 1 3  
## 1 0 0 5  
## 4 1 1 4  
## 3 1 0 5  
## 12 0 0 6  
## 62 1 1 1  
## 9 1 0 2  
## 5 0 1 2  
## 53 0 0 3  
## 44 1 1 2  
## 9 1 0 3  
## 5 0 1 3  
## 28 0 0 4  
## 5 1 1 2  
## 2 0 1 3  
## 26 0 0 4  
## 1 1 1 3  
## 1 0 1 4  
## 16 0 0 5  
## 3 1 1 3  
## 2 1 0 4  
## 2 0 1 4  
## 35 0 0 5  
## 3 1 1 4  
## 4 0 1 5  
## 34 0 0 6  
## 1 1 1 2  
## 1 1 0 3  
## 1 0 1 3  
## 1 0 0 4  
## 2 1 1 3  
## 2 1 0 4  
## 4 0 0 5  
## 5 0 0 6  
## 1 1 1 4  
## 2 0 0 6  
## 12 0 0 7  
## 5 1 1 1  
## 3 1 0 2  
## 1 0 1 2  
## 1 0 0 3  
## 11 1 1 2  
## 3 1 0 3  
## 1 0 1 3  
## 4 0 0 4  
## 1 0 0 5  
## 1 1 0 5  
## 6 0 0 6  
## 1 1 1 2  
## 1 1 0 3  
## 5 1 1 3  
## 5 1 0 4  
## 1 0 1 4  
## 4 0 0 5  
## 1 0 0 5  
## 2 1 1 4  
## 1 1 0 5  
## 1 0 0 6  
## 1 0 0 6  
## 1 1 1 3  
## 1 1 1 4  
## 1 1 0 5  
## 1 0 0 6  
## 1 1 0 5  
## 5938 6140 20733

#Data Preparation  
datTrain.casted <- datTrain %>% select(-enrollee\_id, -city) %>%  
 mutate(city\_development\_index = as.numeric(city\_development\_index),  
 target = as.factor(target)) %>% as.data.frame() %>%  
 mutate\_if(is.character, as.factor) #converting all variables with type=char to factor  
  
   
  
#Data Imputation  
impute <- mice(datTrain.casted, m = 1, remove.constant = FALSE)

##   
## iter imp variable  
## 1 1 gender enrolled\_university education\_level major\_discipline experience company\_size company\_type last\_new\_job  
## 2 1 gender enrolled\_university education\_level major\_discipline experience company\_size company\_type last\_new\_job  
## 3 1 gender enrolled\_university education\_level major\_discipline experience company\_size company\_type last\_new\_job  
## 4 1 gender enrolled\_university education\_level major\_discipline experience company\_size company\_type last\_new\_job  
## 5 1 gender enrolled\_university education\_level major\_discipline experience company\_size company\_type last\_new\_job

impute$loggedEvents

## it im dep meth  
## 1 1 1 major\_discipline polyreg  
## 2 2 1 major\_discipline polyreg  
## 3 3 1 major\_discipline polyreg  
## 4 4 1 major\_discipline polyreg  
## 5 5 1 major\_discipline polyreg  
## out  
## 1 education\_levelHigh School, education\_levelPrimary School  
## 2 education\_levelHigh School, education\_levelPrimary School  
## 3 education\_levelHigh School, education\_levelPrimary School  
## 4 education\_levelHigh School, education\_levelPrimary School  
## 5 education\_levelHigh School, education\_levelPrimary School

datTrain.imputed <- complete(impute)  
  
apply(datTrain.imputed, 2, detectNA)

## city\_development\_index gender relevent\_experience   
## 0 0 0   
## enrolled\_university education\_level major\_discipline   
## 0 0 0   
## experience company\_size company\_type   
## 0 0 0   
## last\_new\_job training\_hours target   
## 0 0 0

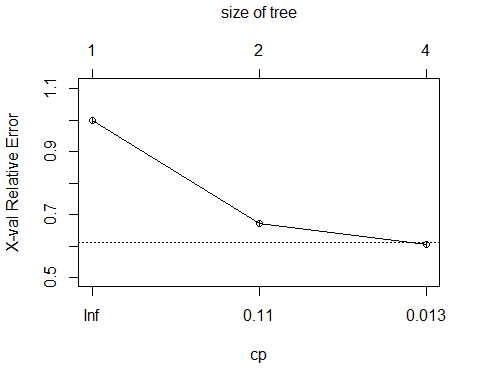
#Data Balancing  
library(groupdata2)  
  
target.balancing <- datTrain.imputed %>%   
 group\_by(target) %>%  
 summarise(no\_rows = length(target))  
  
datTrain.balanced <- upsample(datTrain.imputed, cat\_col = "target")  
  
datTrain.balanced %>%   
 group\_by(target) %>%  
 summarise(no\_rows = length(target))

## # A tibble: 2 x 2  
## target no\_rows  
## <fct> <int>  
## 1 0 14381  
## 2 1 14381

#Test / Training Split  
set.seed(3024)  
trainingRows <- sort(sample(nrow(datTrain.balanced), nrow(datTrain.balanced)\*.7))  
  
Train <- datTrain.balanced[trainingRows,]  
Test <- datTrain.balanced[-trainingRows,]  
  
###Decision tree  
  
library(rpart)  
library(rpart.plot)  
  
set.seed(39)  
dt.fit <- rpart(target~ ., data=Train, cp=0.005)  
  
printcp(dt.fit)

##   
## Classification tree:  
## rpart(formula = target ~ ., data = Train, cp = 0.005)  
##   
## Variables actually used in tree construction:  
## [1] city\_development\_index education\_level relevent\_experience   
##   
## Root node error: 10036/20133 = 0.49849  
##   
## n= 20133   
##   
## CP nsplit rel error xerror xstd  
## 1 0.326425 0 1.00000 1.00000 0.0070691  
## 2 0.034625 1 0.67358 0.67358 0.0066769  
## 3 0.005000 3 0.60432 0.60432 0.0064866

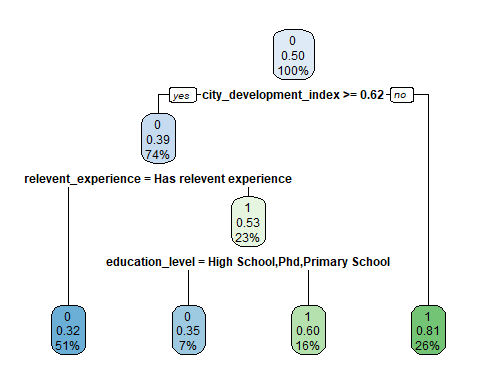
plotcp(dt.fit)



#search cp with lowest Cross-Validation Error  
xerror.min <- dt.fit$cptable[which.min(dt.fit$cptable[,4]),]  
cp.best <- xerror.min[1]  
  
#Pruning with best cp  
dt.pruned <- prune(dt.fit, cp=cp.best)  
  
summary(dt.pruned)

## Call:  
## rpart(formula = target ~ ., data = Train, cp = 0.005)  
## n= 20133   
##   
## CP nsplit rel error xerror xstd  
## 1 0.32642487 0 1.0000000 1.0000000 0.007069059  
## 2 0.03462535 1 0.6735751 0.6735751 0.006676869  
## 3 0.00500000 3 0.6043244 0.6043244 0.006486594  
##   
## Variable importance  
## city\_development\_index relevent\_experience education\_level   
## 69 13 8   
## experience enrolled\_university last\_new\_job   
## 4 3 2   
## company\_type   
## 1   
##   
## Node number 1: 20133 observations, complexity param=0.3264249  
## predicted class=0 expected loss=0.4984851 P(node) =1  
## class counts: 10097 10036  
## probabilities: 0.502 0.498   
## left son=2 (14915 obs) right son=3 (5218 obs)  
## Primary splits:  
## city\_development\_index < 0.6245 to the right, improve=1401.5890, (0 missing)  
## experience splits as RLRLLLLLLLLLLRLRRRRRRL, improve= 380.7607, (0 missing)  
## enrolled\_university splits as RLL, improve= 298.6551, (0 missing)  
## relevent\_experience splits as LR, improve= 209.0044, (0 missing)  
## education\_level splits as RLLLL, improve= 117.5548, (0 missing)  
## Surrogate splits:  
## experience splits as RLLLLLLLLLLLLLLLLLLLLL, agree=0.744, adj=0.014, (0 split)  
##   
## Node number 2: 14915 observations, complexity param=0.03462535  
## predicted class=0 expected loss=0.3881328 P(node) =0.7408235  
## class counts: 9126 5789  
## probabilities: 0.612 0.388   
## left son=4 (10276 obs) right son=5 (4639 obs)  
## Primary splits:  
## relevent\_experience splits as LR, improve=273.7800, (0 missing)  
## enrolled\_university splits as RLL, improve=232.2027, (0 missing)  
## experience splits as RLRLLLLLLLLLLRLRRLLRLL, improve=107.2240, (0 missing)  
## city\_development\_index < 0.7915 to the right, improve=106.5384, (0 missing)  
## company\_type splits as LLRRRL, improve=102.7185, (0 missing)  
## Surrogate splits:  
## enrolled\_university splits as RLL, agree=0.769, adj=0.257, (0 split)  
## experience splits as RLRLLLLLLLLLLRLRLLLLLL, agree=0.750, adj=0.196, (0 split)  
## last\_new\_job splits as LLLLLR, agree=0.747, adj=0.186, (0 split)  
## education\_level splits as LRLLR, agree=0.723, adj=0.110, (0 split)  
## company\_type splits as LLLLRL, agree=0.707, adj=0.059, (0 split)  
##   
## Node number 3: 5218 observations  
## predicted class=1 expected loss=0.1860866 P(node) =0.2591765  
## class counts: 971 4247  
## probabilities: 0.186 0.814   
##   
## Node number 4: 10276 observations  
## predicted class=0 expected loss=0.3237641 P(node) =0.5104058  
## class counts: 6949 3327  
## probabilities: 0.676 0.324   
##   
## Node number 5: 4639 observations, complexity param=0.03462535  
## predicted class=1 expected loss=0.4692822 P(node) =0.2304177  
## class counts: 2177 2462  
## probabilities: 0.469 0.531   
## left son=10 (1326 obs) right son=11 (3313 obs)  
## Primary splits:  
## education\_level splits as RLRLL, improve=127.53000, (0 missing)  
## enrolled\_university splits as RLR, improve= 68.95016, (0 missing)  
## city\_development\_index < 0.922 to the right, improve= 64.44139, (0 missing)  
## last\_new\_job splits as LRRRRL, improve= 42.11660, (0 missing)  
## company\_type splits as LLLRRL, improve= 25.96188, (0 missing)  
## Surrogate splits:  
## city\_development\_index < 0.6425 to the left, agree=0.714, adj=0.001, (0 split)  
##   
## Node number 10: 1326 observations  
## predicted class=0 expected loss=0.3453997 P(node) =0.06586202  
## class counts: 868 458  
## probabilities: 0.655 0.345   
##   
## Node number 11: 3313 observations  
## predicted class=1 expected loss=0.3951102 P(node) =0.1645557  
## class counts: 1309 2004  
## probabilities: 0.395 0.605

rpart.plot(dt.pruned)



#Prediction & Performance  
#Train Data  
train.pred <- predict(dt.pruned, newdata = Train, type = "class")  
train.confMatrix <- table(train.pred, Train[, 12])  
print(train.confMatrix)

##   
## train.pred 0 1  
## 0 7817 3785  
## 1 2280 6251

train.accuracy <- sum(diag(train.confMatrix))/sum(train.confMatrix)  
print(train.accuracy)

## [1] 0.6987533

#Test Data  
test.pred <- predict(dt.pruned, newdata = Test, type = "class")  
test.confMatrix <- table(test.pred, Test[, 12])  
print(test.confMatrix)

##   
## test.pred 0 1  
## 0 3272 1660  
## 1 1012 2685

test.accuracy <- sum(diag(test.confMatrix))/sum(test.confMatrix)  
print(test.accuracy)

## [1] 0.6903465

###glm  
#dummy variable  
  
dummy\_target <- as.numeric(Train[12] == 1)  
dat.dummy\_target <- cbind(Train,dummy\_target)  
  
set.seed(4)  
glm.fit <- glm(target ~ ., family = binomial, Train)  
summary(glm.fit)

##   
## Call:  
## glm(formula = target ~ ., family = binomial, data = Train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4399 -0.9265 -0.4851 0.9452 2.2432   
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) 5.8168648 0.2166719 26.846  
## city\_development\_index -5.8543468 0.1359752 -43.055  
## genderMale -0.2221090 0.0550014 -4.038  
## genderOther -0.1435268 0.1421903 -1.009  
## relevent\_experienceNo relevent experience 0.5975576 0.0426092 14.024  
## enrolled\_universityno\_enrollment -0.4139532 0.0437657 -9.458  
## enrolled\_universityPart time course -0.4771810 0.0711134 -6.710  
## education\_levelHigh School -0.8301602 0.0576484 -14.400  
## education\_levelMasters -0.3079838 0.0389226 -7.913  
## education\_levelPhd -0.7841340 0.1245006 -6.298  
## education\_levelPrimary School -1.1850776 0.1415536 -8.372  
## major\_disciplineBusiness Degree 0.2495140 0.1625282 1.535  
## major\_disciplineHumanities -0.0530453 0.1446462 -0.367  
## major\_disciplineNo Major -0.2115293 0.1811200 -1.168  
## major\_disciplineOther -0.1563483 0.1591599 -0.982  
## major\_disciplineSTEM -0.1082671 0.1249672 -0.866  
## experience>20 -0.4451846 0.1097677 -4.056  
## experience1 0.0498592 0.1285748 0.388  
## experience10 -0.5818854 0.1198217 -4.856  
## experience11 -0.2799909 0.1277456 -2.192  
## experience12 -0.6665005 0.1397792 -4.768  
## experience13 -0.6538347 0.1492851 -4.380  
## experience14 -0.5560289 0.1349068 -4.122  
## experience15 -0.6163734 0.1320196 -4.669  
## experience16 -0.9592687 0.1505769 -6.371  
## experience17 -0.6298331 0.1576111 -3.996  
## experience18 -0.6718318 0.1738735 -3.864  
## experience19 -0.4483955 0.1670768 -2.684  
## experience2 -0.4640734 0.1135401 -4.087  
## experience20 -0.1439359 0.2011071 -0.716  
## experience3 -0.3152526 0.1106820 -2.848  
## experience4 -0.2690743 0.1103625 -2.438  
## experience5 -0.5085675 0.1113180 -4.569  
## experience6 -0.4431160 0.1142951 -3.877  
## experience7 -0.3629950 0.1167790 -3.108  
## experience8 -0.5098286 0.1234902 -4.128  
## experience9 -0.5928314 0.1200976 -4.936  
## company\_size10/49 0.3686967 0.0662959 5.561  
## company\_size100-500 -0.0237965 0.0632974 -0.376  
## company\_size1000-4999 0.1402222 0.0722746 1.940  
## company\_size10000+ 0.3706318 0.0648016 5.719  
## company\_size50-99 0.0699619 0.0598299 1.169  
## company\_size500-999 0.0368845 0.0785771 0.469  
## company\_size5000-9999 0.1260724 0.0906983 1.390  
## company\_typeFunded Startup -0.2615524 0.0939998 -2.782  
## company\_typeNGO 0.0479092 0.1034399 0.463  
## company\_typeOther 0.6053999 0.1695719 3.570  
## company\_typePublic Sector 0.3040805 0.0918710 3.310  
## company\_typePvt Ltd -0.1085088 0.0741413 -1.464  
## last\_new\_job1 -0.0648884 0.0510954 -1.270  
## last\_new\_job2 0.0487677 0.0588721 0.828  
## last\_new\_job3 0.0955230 0.0788126 1.212  
## last\_new\_job4 0.1231432 0.0783200 1.572  
## last\_new\_jobnever -0.2975641 0.0675085 -4.408  
## training\_hours -0.0009267 0.0002657 -3.487  
## Pr(>|z|)   
## (Intercept) < 2e-16 \*\*\*  
## city\_development\_index < 2e-16 \*\*\*  
## genderMale 5.39e-05 \*\*\*  
## genderOther 0.312783   
## relevent\_experienceNo relevent experience < 2e-16 \*\*\*  
## enrolled\_universityno\_enrollment < 2e-16 \*\*\*  
## enrolled\_universityPart time course 1.94e-11 \*\*\*  
## education\_levelHigh School < 2e-16 \*\*\*  
## education\_levelMasters 2.52e-15 \*\*\*  
## education\_levelPhd 3.01e-10 \*\*\*  
## education\_levelPrimary School < 2e-16 \*\*\*  
## major\_disciplineBusiness Degree 0.124734   
## major\_disciplineHumanities 0.713825   
## major\_disciplineNo Major 0.242849   
## major\_disciplineOther 0.325935   
## major\_disciplineSTEM 0.386291   
## experience>20 5.00e-05 \*\*\*  
## experience1 0.698176   
## experience10 1.20e-06 \*\*\*  
## experience11 0.028395 \*   
## experience12 1.86e-06 \*\*\*  
## experience13 1.19e-05 \*\*\*  
## experience14 3.76e-05 \*\*\*  
## experience15 3.03e-06 \*\*\*  
## experience16 1.88e-10 \*\*\*  
## experience17 6.44e-05 \*\*\*  
## experience18 0.000112 \*\*\*  
## experience19 0.007280 \*\*   
## experience2 4.36e-05 \*\*\*  
## experience20 0.474166   
## experience3 0.004396 \*\*   
## experience4 0.014765 \*   
## experience5 4.91e-06 \*\*\*  
## experience6 0.000106 \*\*\*  
## experience7 0.001881 \*\*   
## experience8 3.65e-05 \*\*\*  
## experience9 7.96e-07 \*\*\*  
## company\_size10/49 2.68e-08 \*\*\*  
## company\_size100-500 0.706956   
## company\_size1000-4999 0.052364 .   
## company\_size10000+ 1.07e-08 \*\*\*  
## company\_size50-99 0.242264   
## company\_size500-999 0.638780   
## company\_size5000-9999 0.164523   
## company\_typeFunded Startup 0.005395 \*\*   
## company\_typeNGO 0.643250   
## company\_typeOther 0.000357 \*\*\*  
## company\_typePublic Sector 0.000933 \*\*\*  
## company\_typePvt Ltd 0.143320   
## last\_new\_job1 0.204104   
## last\_new\_job2 0.407463   
## last\_new\_job3 0.225502   
## last\_new\_job4 0.115879   
## last\_new\_jobnever 1.04e-05 \*\*\*  
## training\_hours 0.000488 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 27910 on 20132 degrees of freedom  
## Residual deviance: 23843 on 20078 degrees of freedom  
## AIC: 23953  
##   
## Number of Fisher Scoring iterations: 4

varImp(glm.fit, scale = FALSE)

## Overall  
## city\_development\_index 43.0545269  
## genderMale 4.0382456  
## genderOther 1.0093991  
## relevent\_experienceNo relevent experience 14.0241426  
## enrolled\_universityno\_enrollment 9.4583961  
## enrolled\_universityPart time course 6.7101412  
## education\_levelHigh School 14.4004135  
## education\_levelMasters 7.9127141  
## education\_levelPhd 6.2982345  
## education\_levelPrimary School 8.3719334  
## major\_disciplineBusiness Degree 1.5352046  
## major\_disciplineHumanities 0.3667242  
## major\_disciplineNo Major 1.1678960  
## major\_disciplineOther 0.9823348  
## major\_disciplineSTEM 0.8663638  
## experience>20 4.0556983  
## experience1 0.3877837  
## experience10 4.8562602  
## experience11 2.1917848  
## experience12 4.7682378  
## experience13 4.3797703  
## experience14 4.1215782  
## experience15 4.6688011  
## experience16 6.3706235  
## experience17 3.9961228  
## experience18 3.8639123  
## experience19 2.6837684  
## experience2 4.0873083  
## experience20 0.7157176  
## experience3 2.8482720  
## experience4 2.4380958  
## experience5 4.5686018  
## experience6 3.8769469  
## experience7 3.1083936  
## experience8 4.1284958  
## experience9 4.9362479  
## company\_size10/49 5.5613816  
## company\_size100-500 0.3759478  
## company\_size1000-4999 1.9401304  
## company\_size10000+ 5.7194843  
## company\_size50-99 1.1693475  
## company\_size500-999 0.4694055  
## company\_size5000-9999 1.3900190  
## company\_typeFunded Startup 2.7824776  
## company\_typeNGO 0.4631595  
## company\_typeOther 3.5701664  
## company\_typePublic Sector 3.3098625  
## company\_typePvt Ltd 1.4635399  
## last\_new\_job1 1.2699457  
## last\_new\_job2 0.8283667  
## last\_new\_job3 1.2120267  
## last\_new\_job4 1.5723097  
## last\_new\_jobnever 4.4078002  
## training\_hours 3.4872043

glm.fit.predicted <- predict(   
 object = glm.fit,   
 data = Test,   
 type = "response"  
)  
  
plot(  
 x = dat.dummy\_target$city\_development\_index,   
 y = dat.dummy\_target$dummy\_target,   
 col = "red"  
)  
lines(Train$city\_development\_index, glm.fit.predicted, col="blue")

