HR Analytics: Job Change of Data Scientists

A cura di:
Antognozzi Chiara
Rocchi Riccardo

OBIETTIVI E MODALITÀ

IL DATASET ANALIZZATO È FRUTTO DI UN INDAGINE CONDOTTA DA UN'AZIENDA ATTIVA NEL CAMPO DI BIG DATA E DATA SCIENCE.

L'OBIETTIVO DELLA SEGUENTE ANALISI È QUELLO DI ANALIZZARE E PREDIRRE LA PROBABILITÀ DI UN CANDIDATO DI CAMBIARE IL SUO LAVORO CORRENTE, INTERPRETANDO I FATTORI CHE INCIDONO MAGGIORMENTE SULLA SUA DECISONE.

1

ANALISI ESPLORATIVA

2

REGRESSIONE LOGISTICA E MODEL SELECTION

3

ALBERO DECISIONALE E RANDOM FOREST

IL DATASET

```
> str(data_set)
'data.frame':   8841 obs. of  12 variables:
 $ city_development_index: num  0.776 0.767 0.762 0.92 0.92 0.913 0.926 0.843 0.926 0.776 ...
 $ gender                : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ relevent_experience   : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 2 2 ...
 $ enrolled_university   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ education_level       : Factor w/ 3 levels "0","1","2": 1 2 1 1 1 1 1 2 2 1 ...
 $ major_discipline      : Factor w/ 4 levels "0","1","2","3": 2 2 2 2 2 2 2 2 2 1 ...
 $ experience            : int  15 21 13 7 5 21 16 11 11 0 ...
 $ company_size          : Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 5 4 6 2 2 4 ...
 $ company_type          : Factor w/ 4 levels "0","1","2","3": 3 1 3 3 3 3 3 3 3 3 ...
 $ last_new_job          : Ord.factor w/ 3 levels "0"<"1"<"2": 3 3 3 2 2 2 3 2 2 2 ...
 $ training_hours        : num  47 8 18 46 108 23 18 68 50 65 ...
 $ target                : Factor w/ 2 levels "0.0","1.0": 1 1 2 2 1 1 1 1 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:10203] 1 3 4 6 7 10 11 14 15 17 ...
  ..- attr(*, "names")= chr [1:10203] "1" "3" "4" "6" ...
```

FONTE:
https://www.kaggle.com/.

CLEANING DATA AND SOME TRANSFORMATIONS

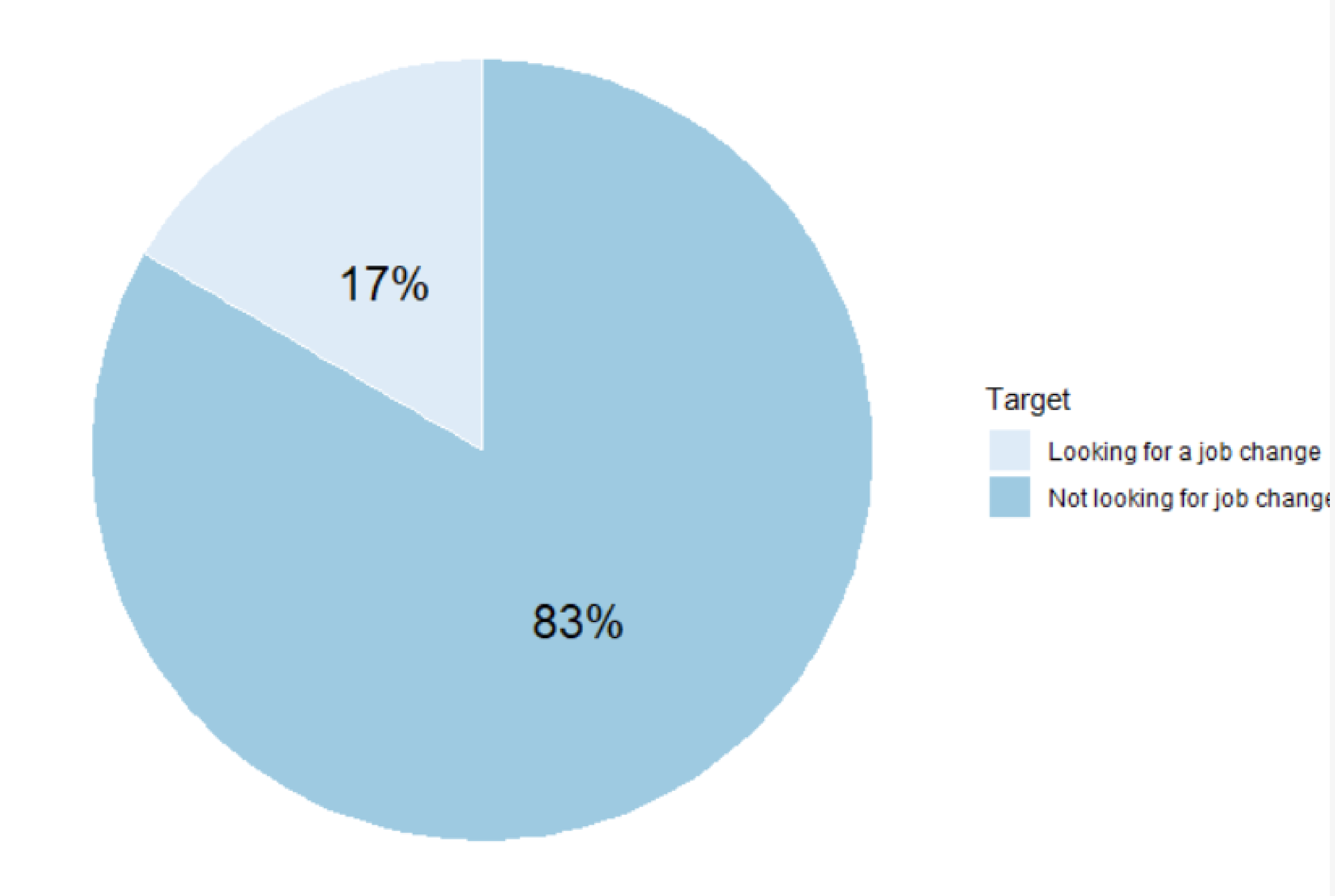
```
data_job<-na.omit(data_job)
```

```
duplicated(data_job)
```

```
data_set=unique(data_job)
```

```
#GENDER

data_set= data_set[data_set$gender!= "Other",]
p=count(data_set, vars=target)
data_set$gender[data_set$gender=="Male"]<-0
data_set$gender[data_set$gender=="Female"]<- 1

data_set$gender<-as.factor(data_set$gender)
```

ANALISI ESPLORATIVA



TARGET

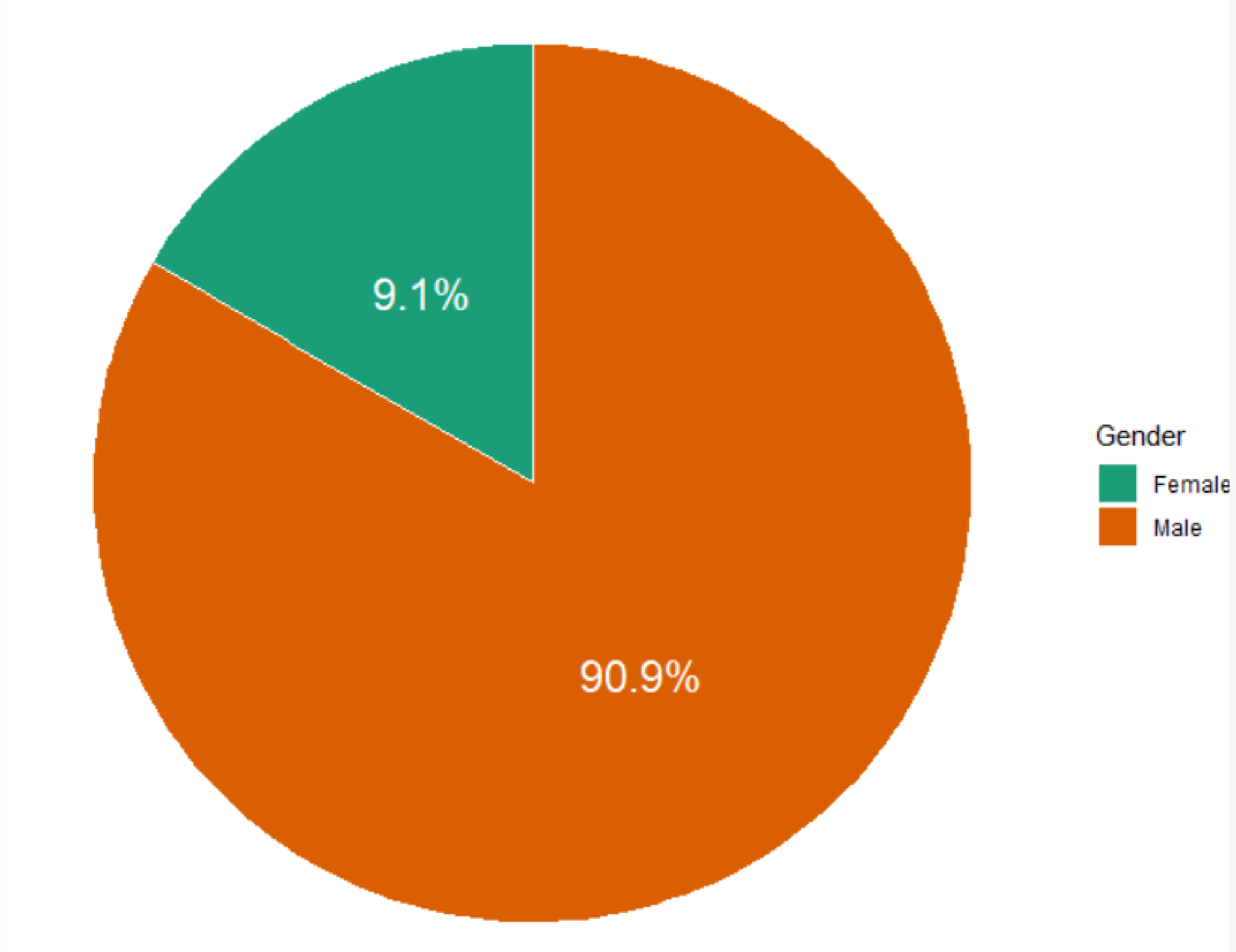
0: NON CAMBIA LAVORO

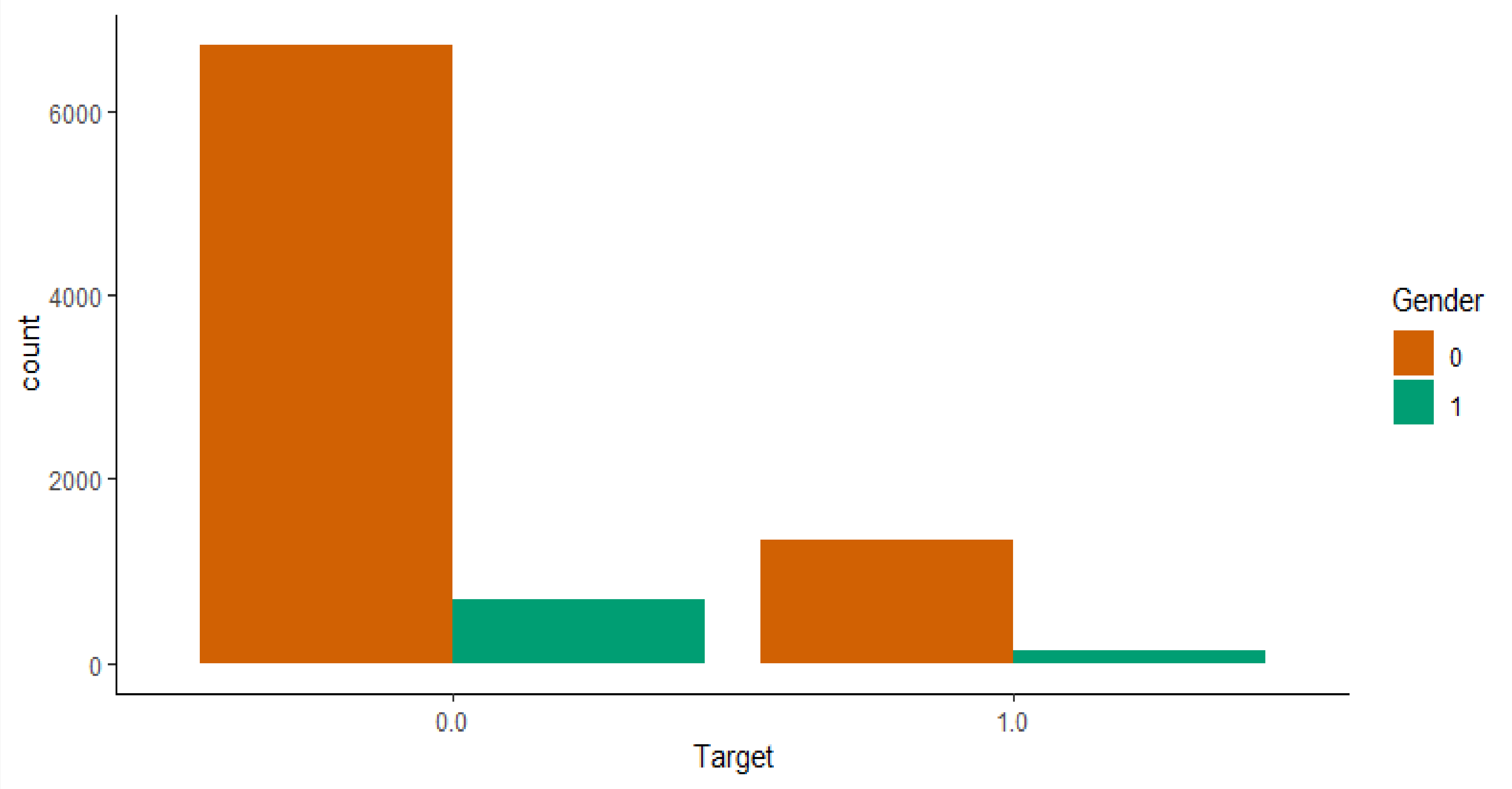
1: CAMBIA LAVORO

ANALISI ESPLORATIVA

GENDER

1

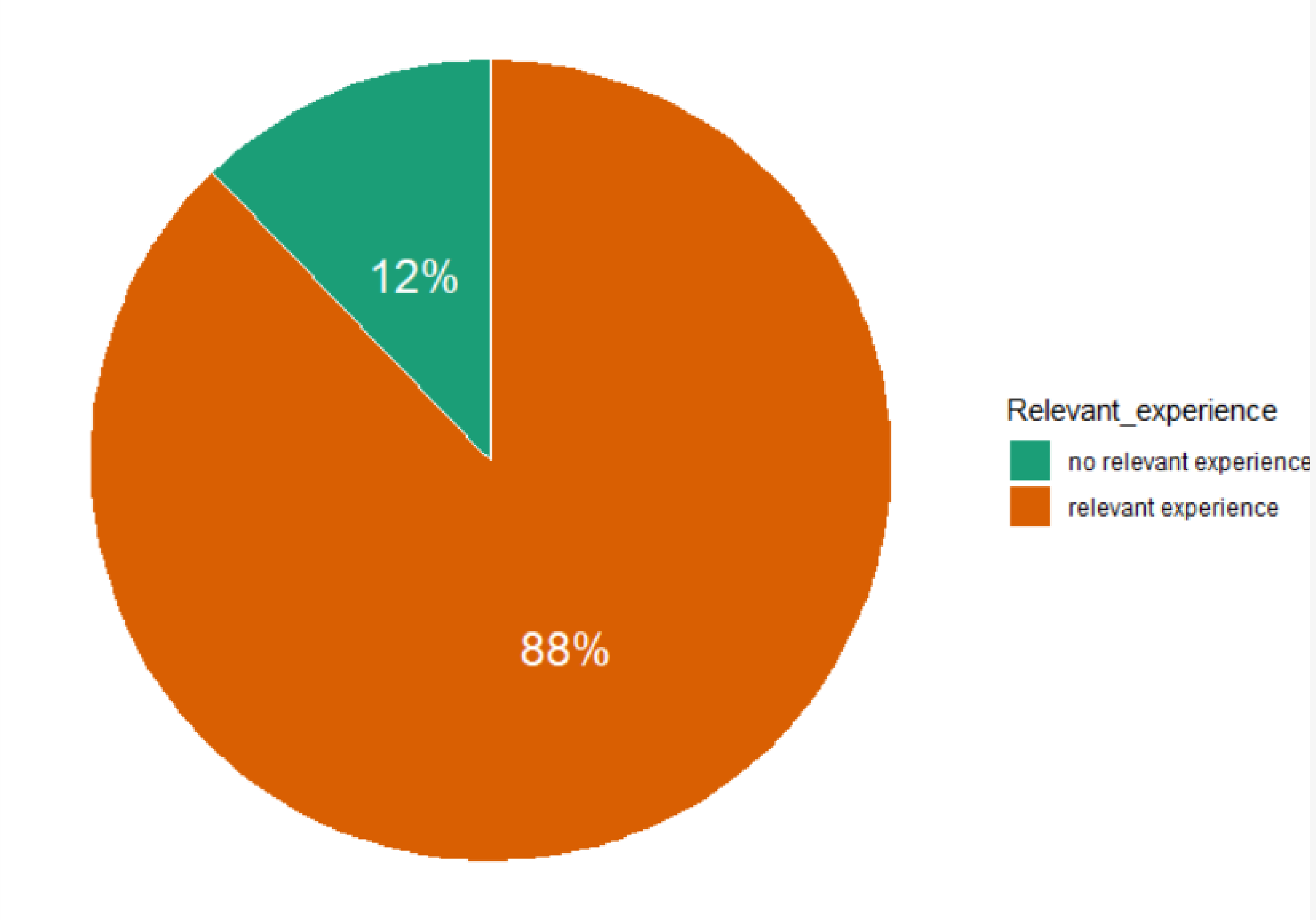


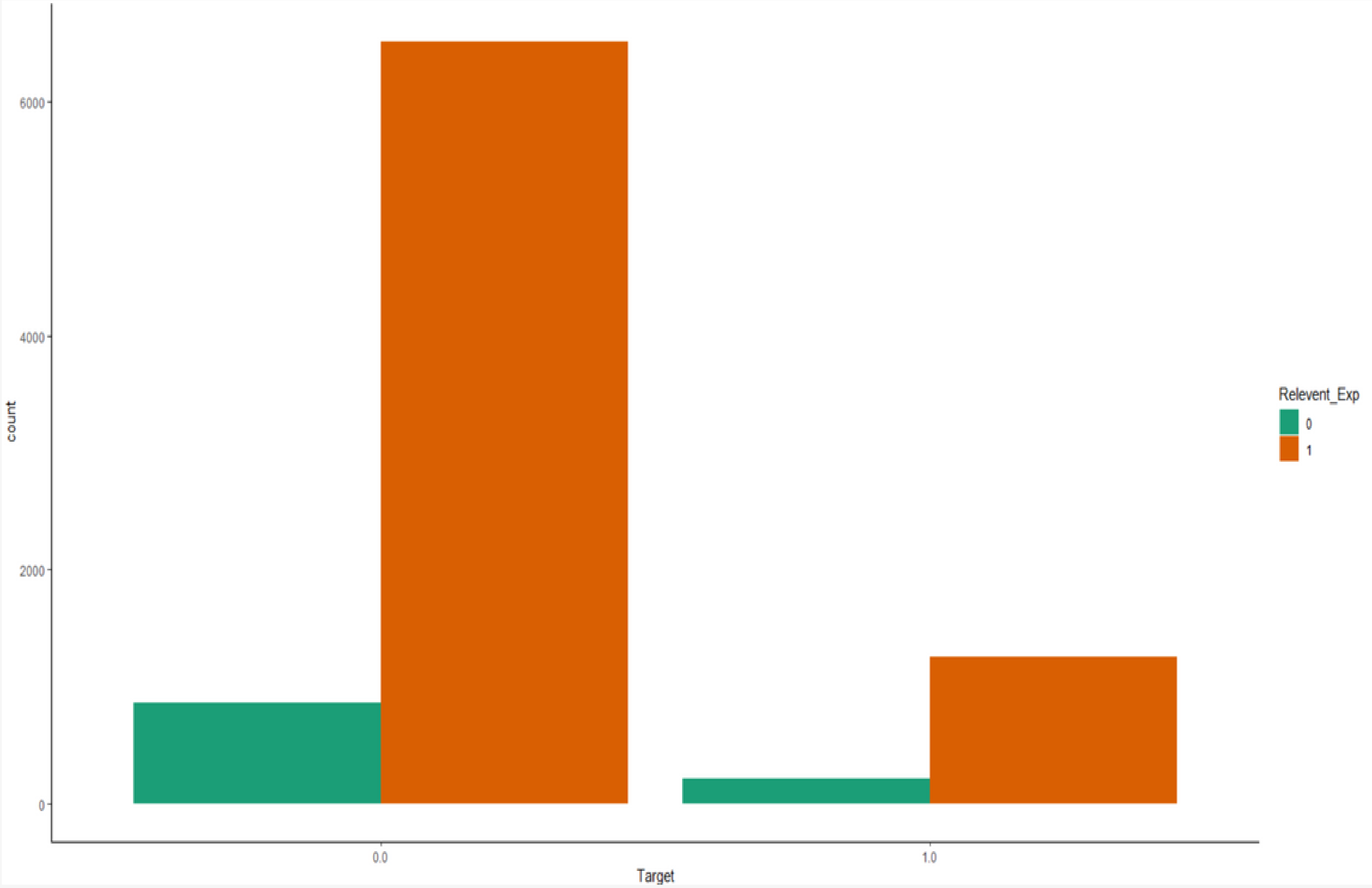


ANALISI ESPLORATIVA

RELEVANT EXPERIENCE

1

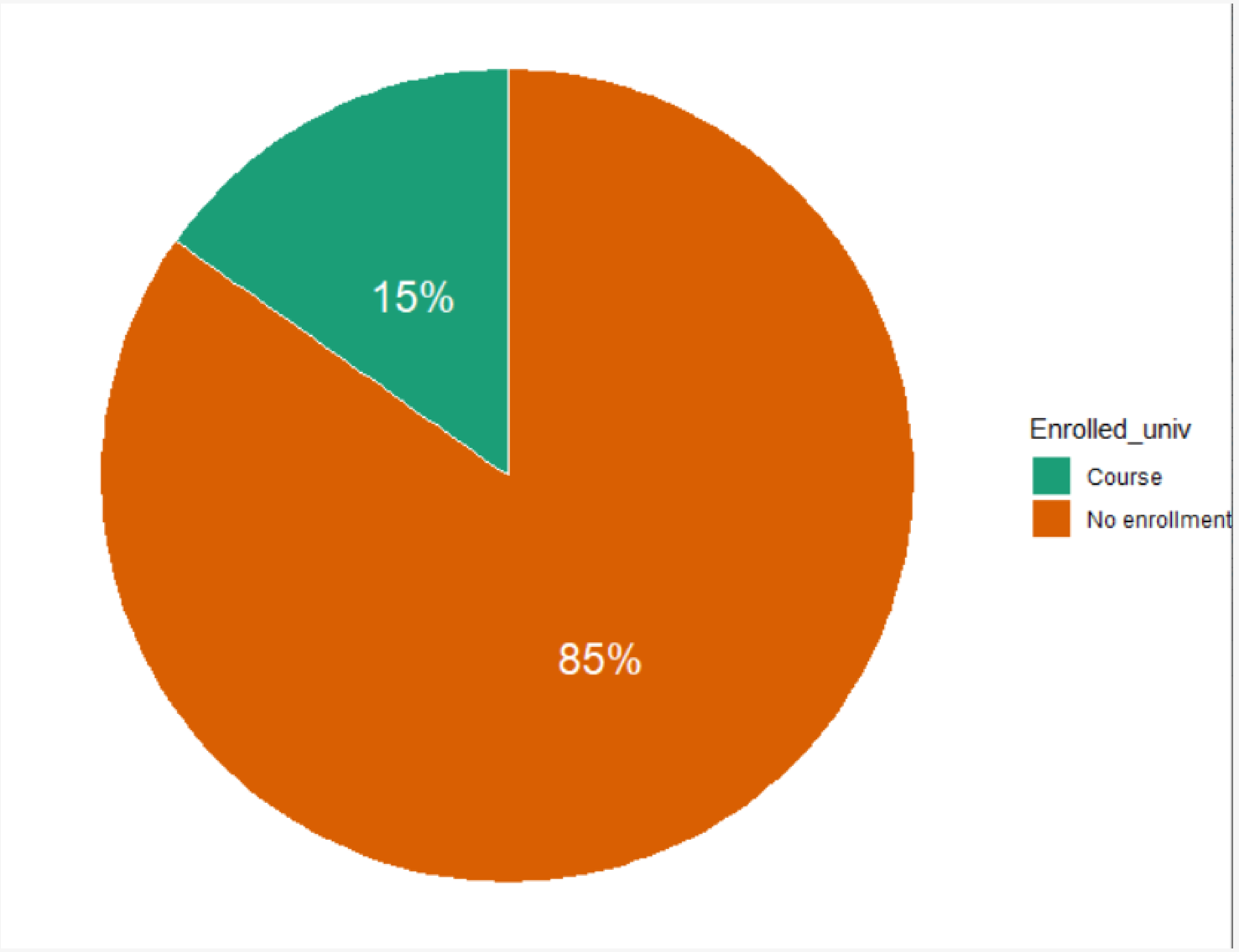


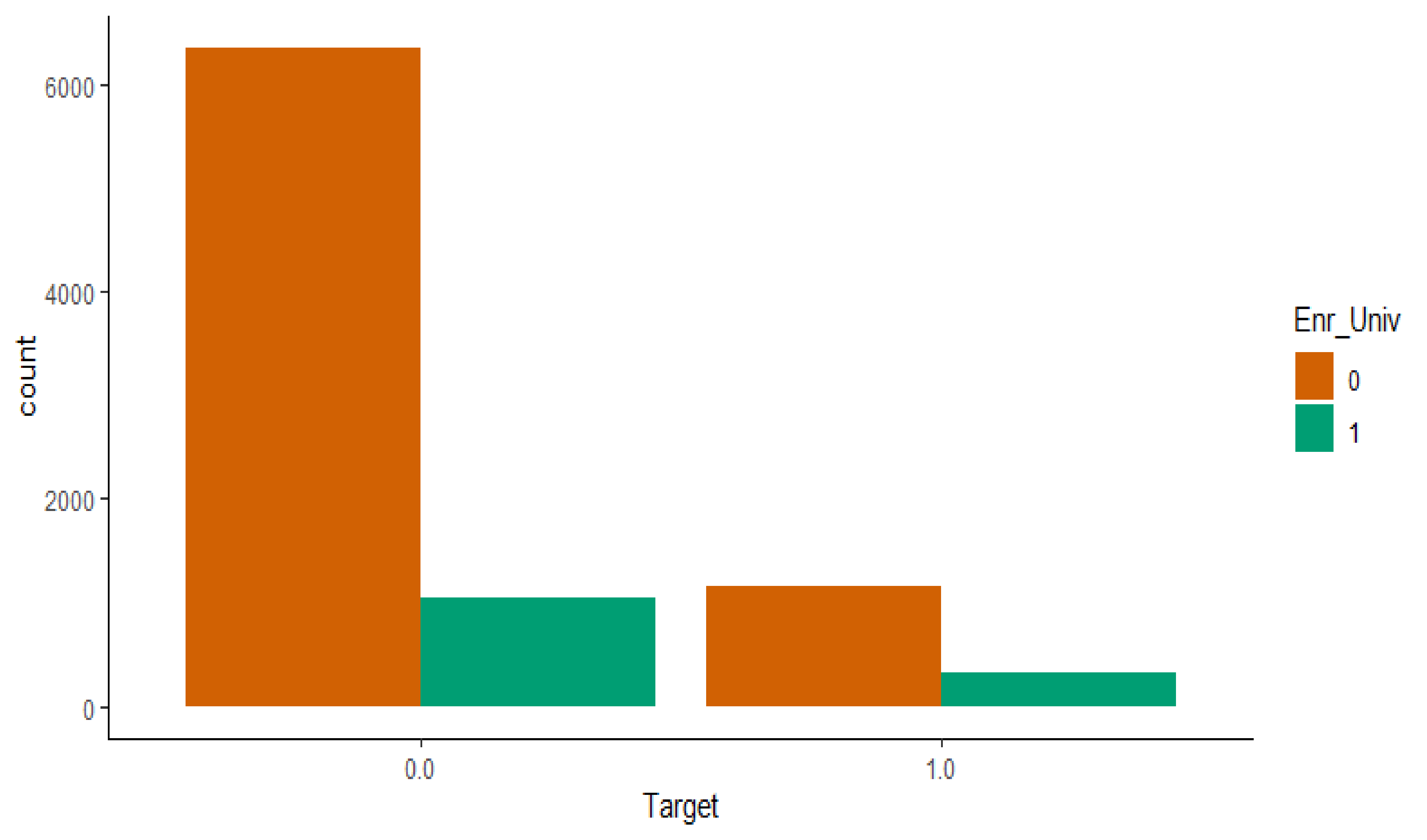


ANALISI ESPLORATIVA

ENROLLED UNIVERSITY

1

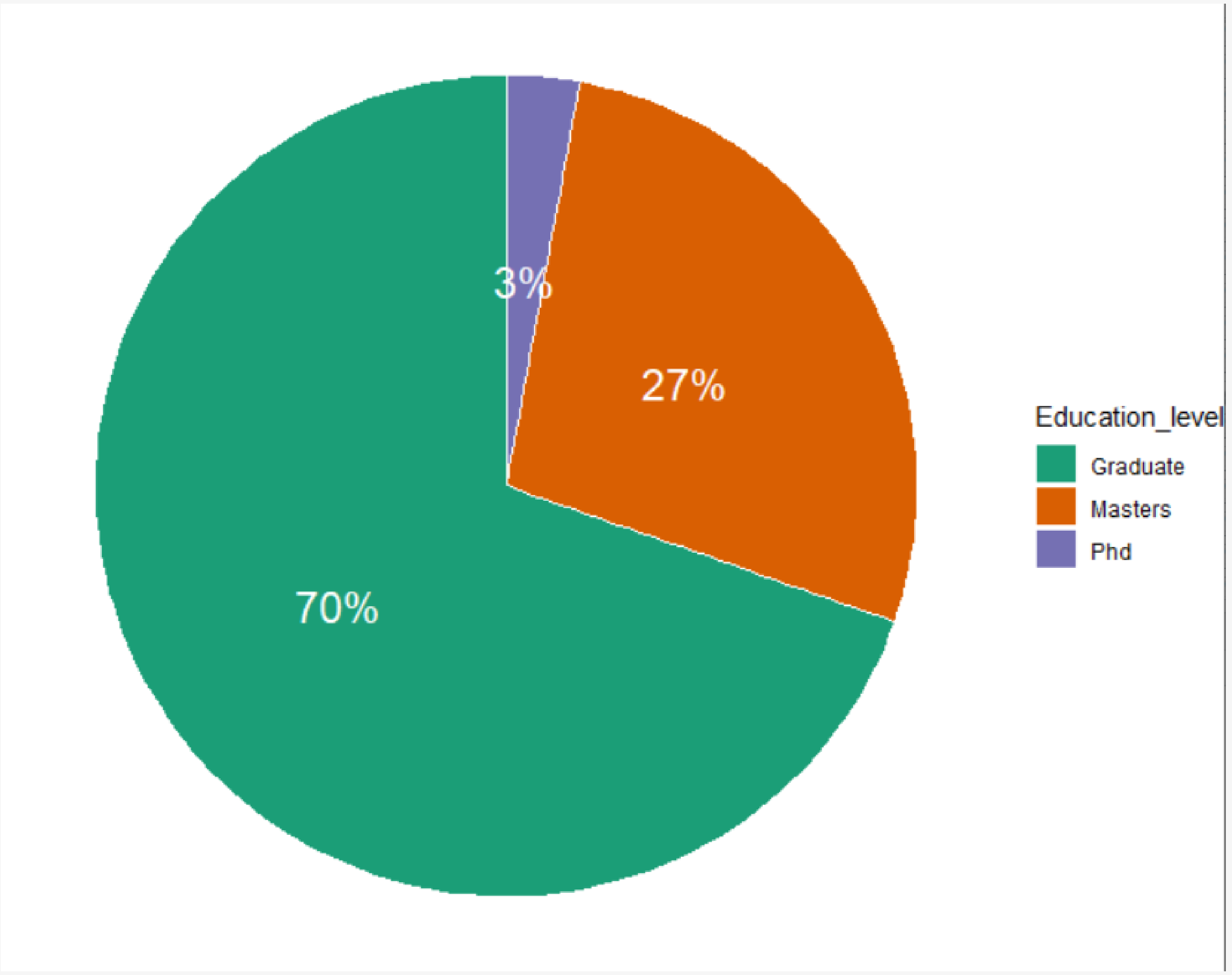


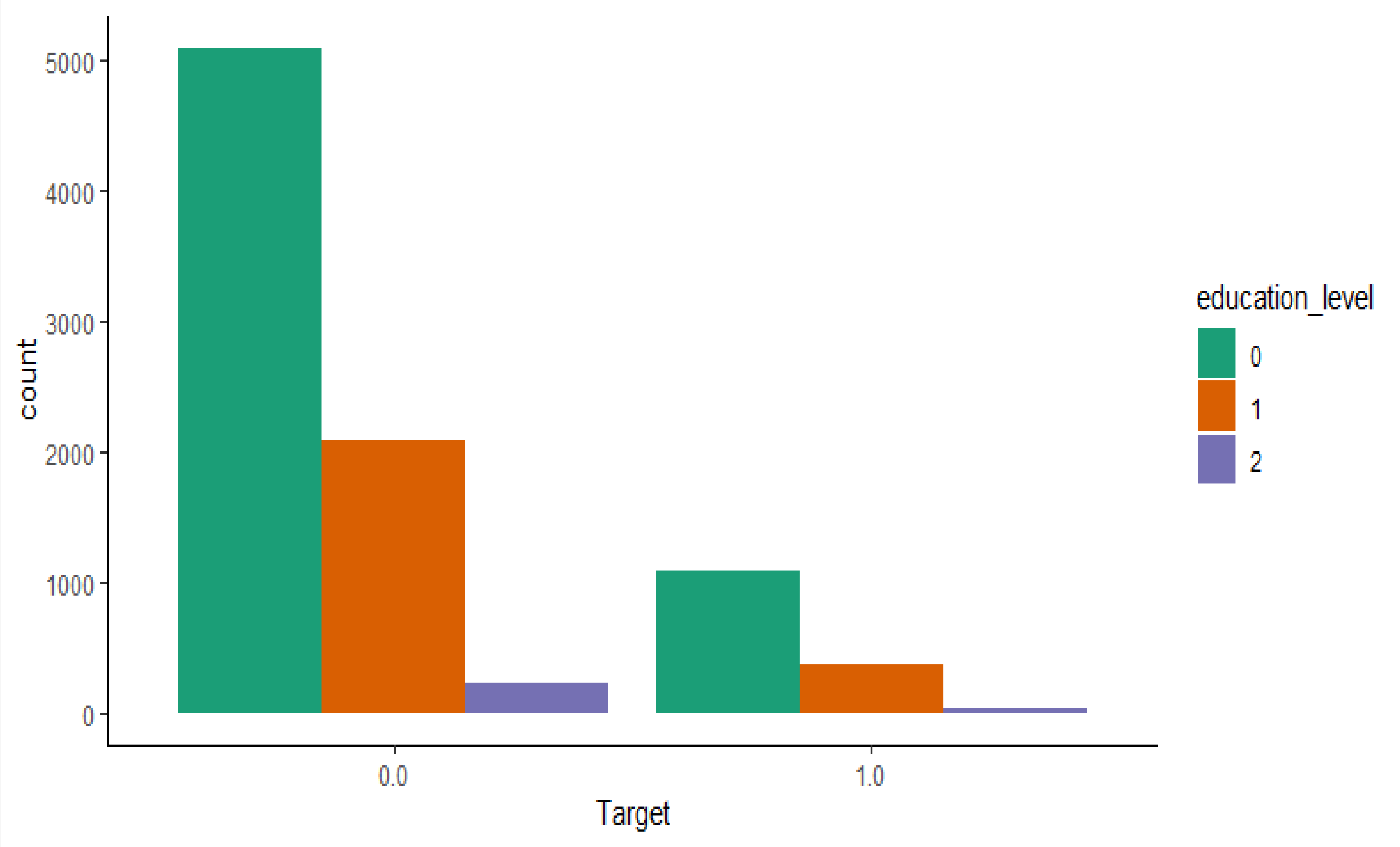


ANALISI ESPLORATIVA

EDUCATION LEVEL

1

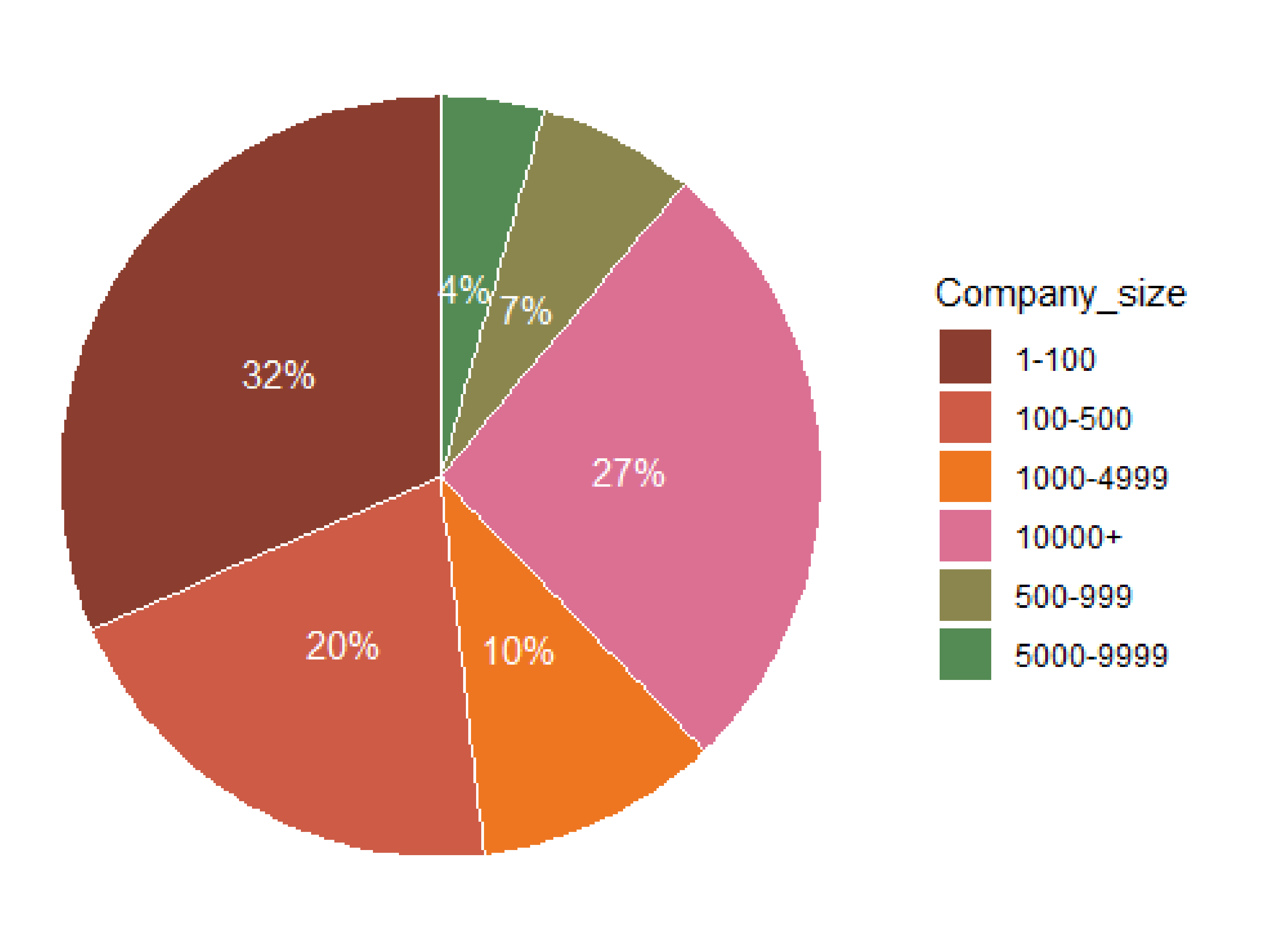


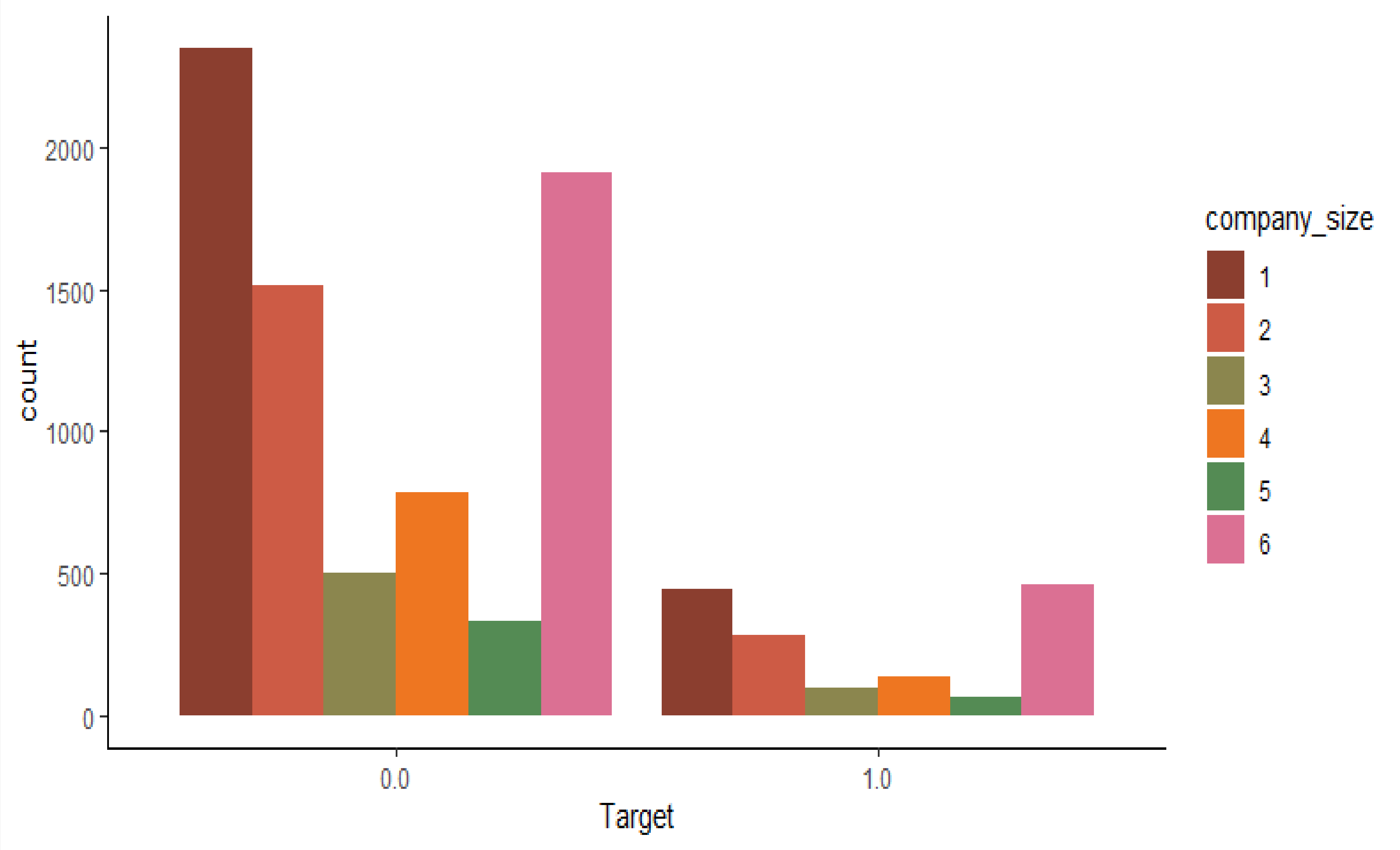


ANALISI ESPLORATIVA

COMPANY SIZE

1

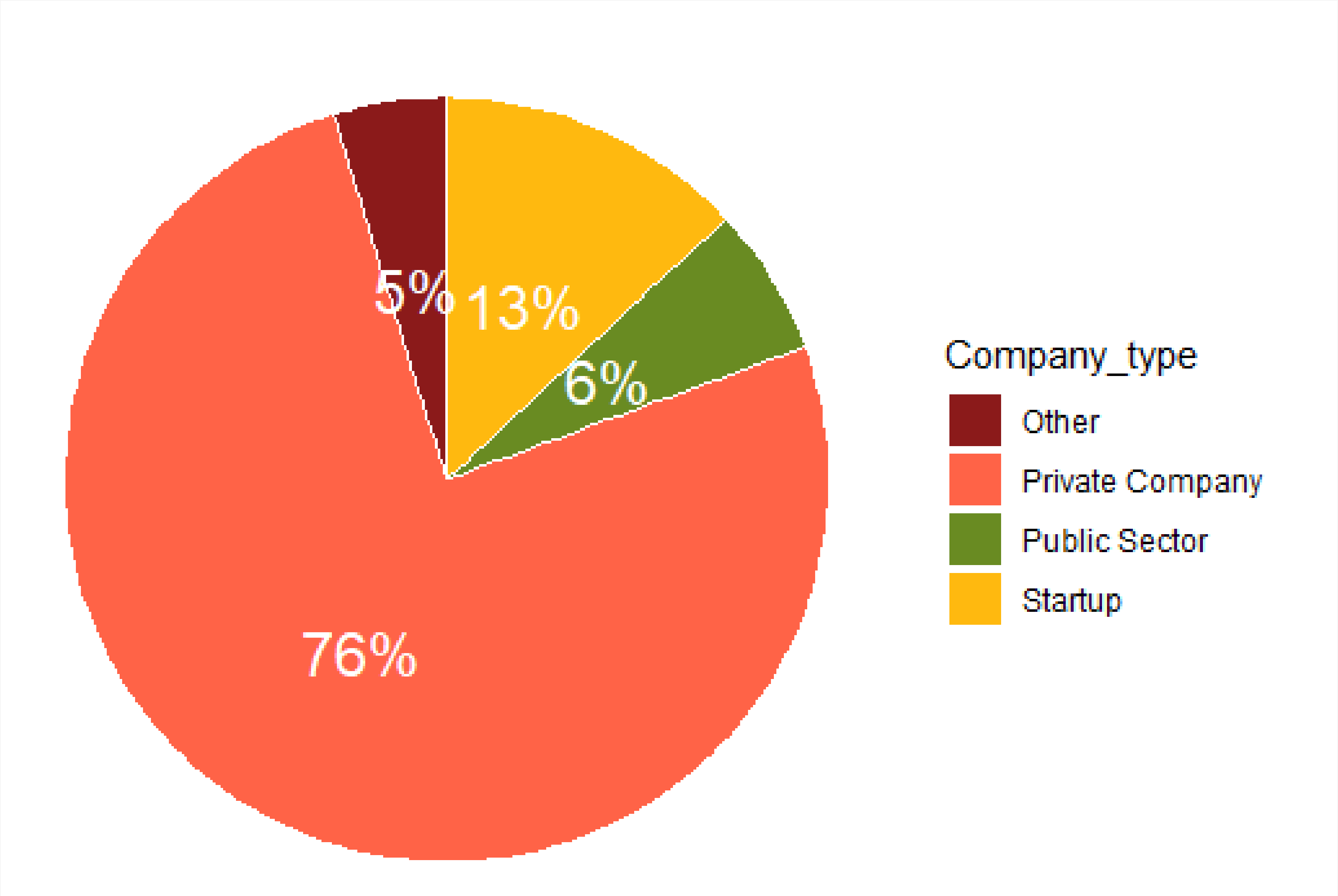


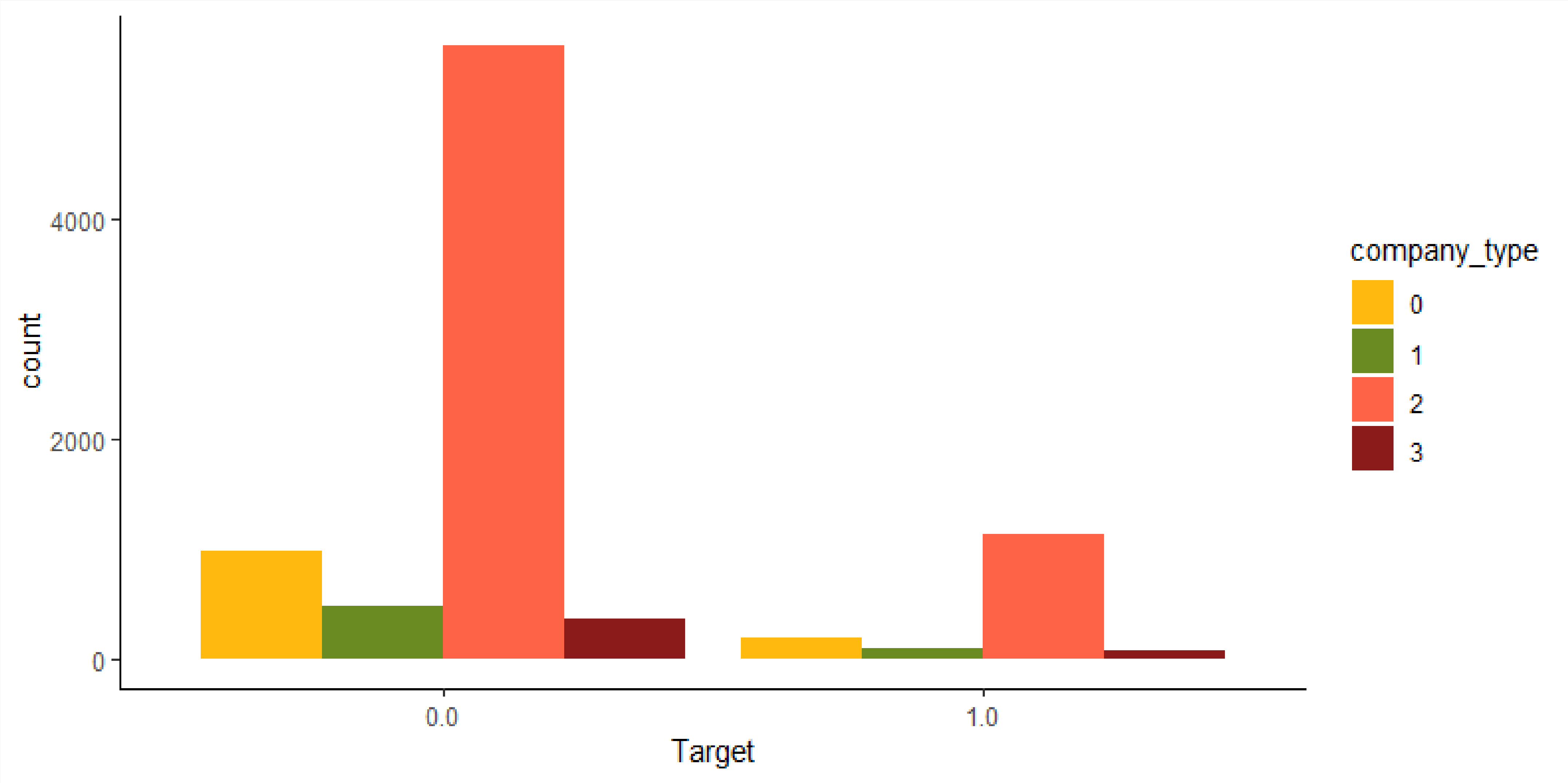


ANALISI ESPLORATIVA

COMPANY TYPE

1

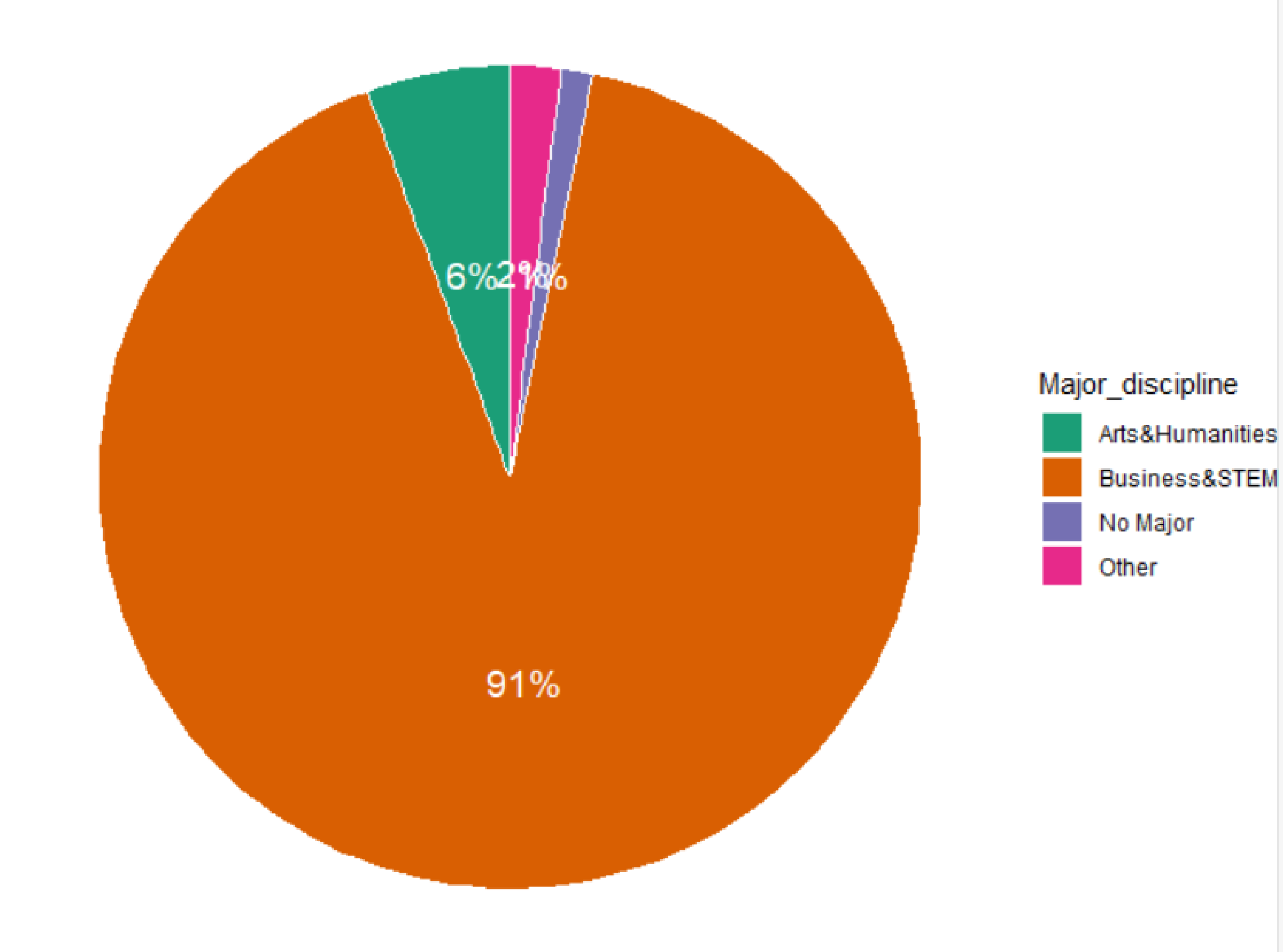


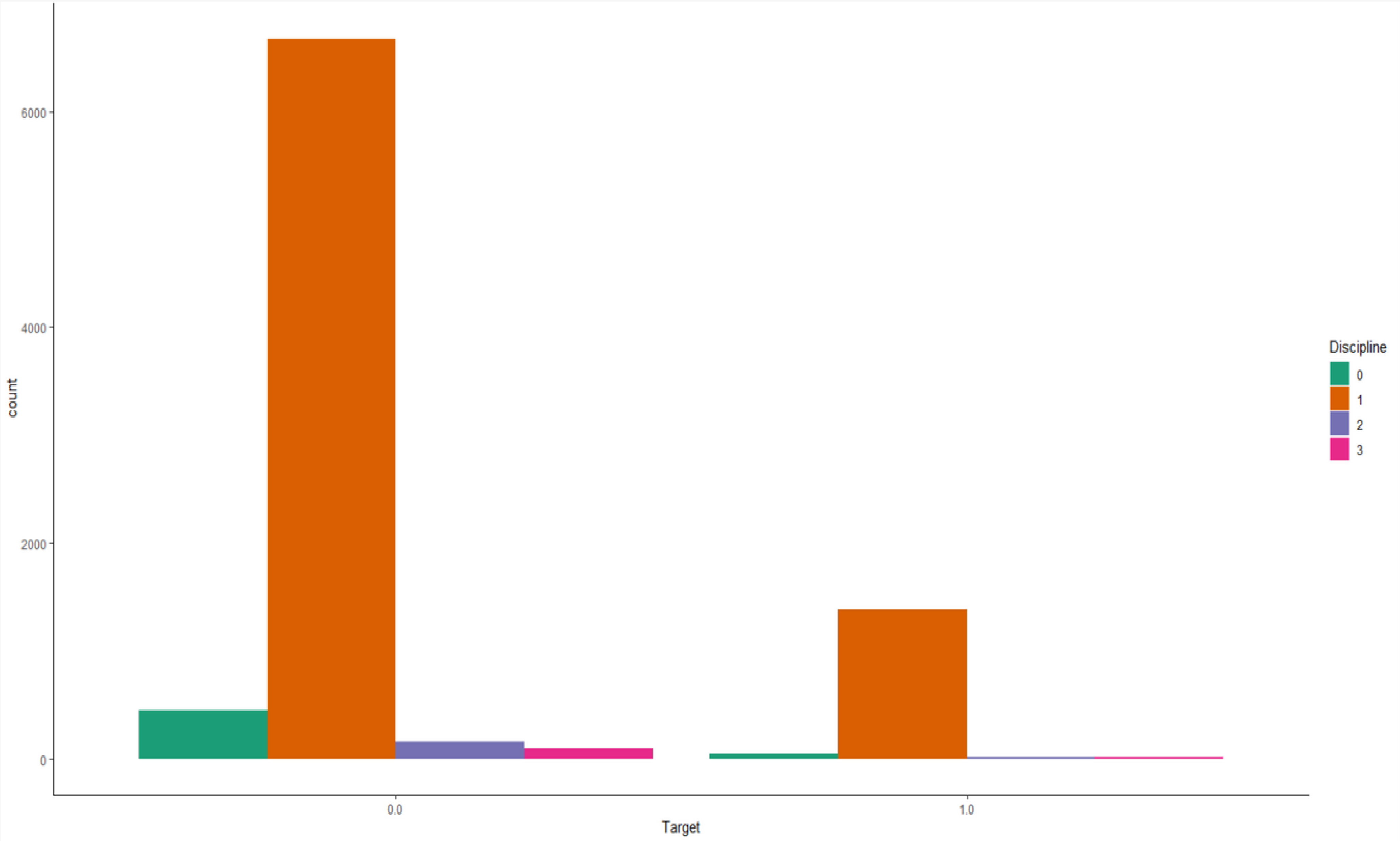


ANALISI ESPLORATIVA

MAJOR DISCIPLINE

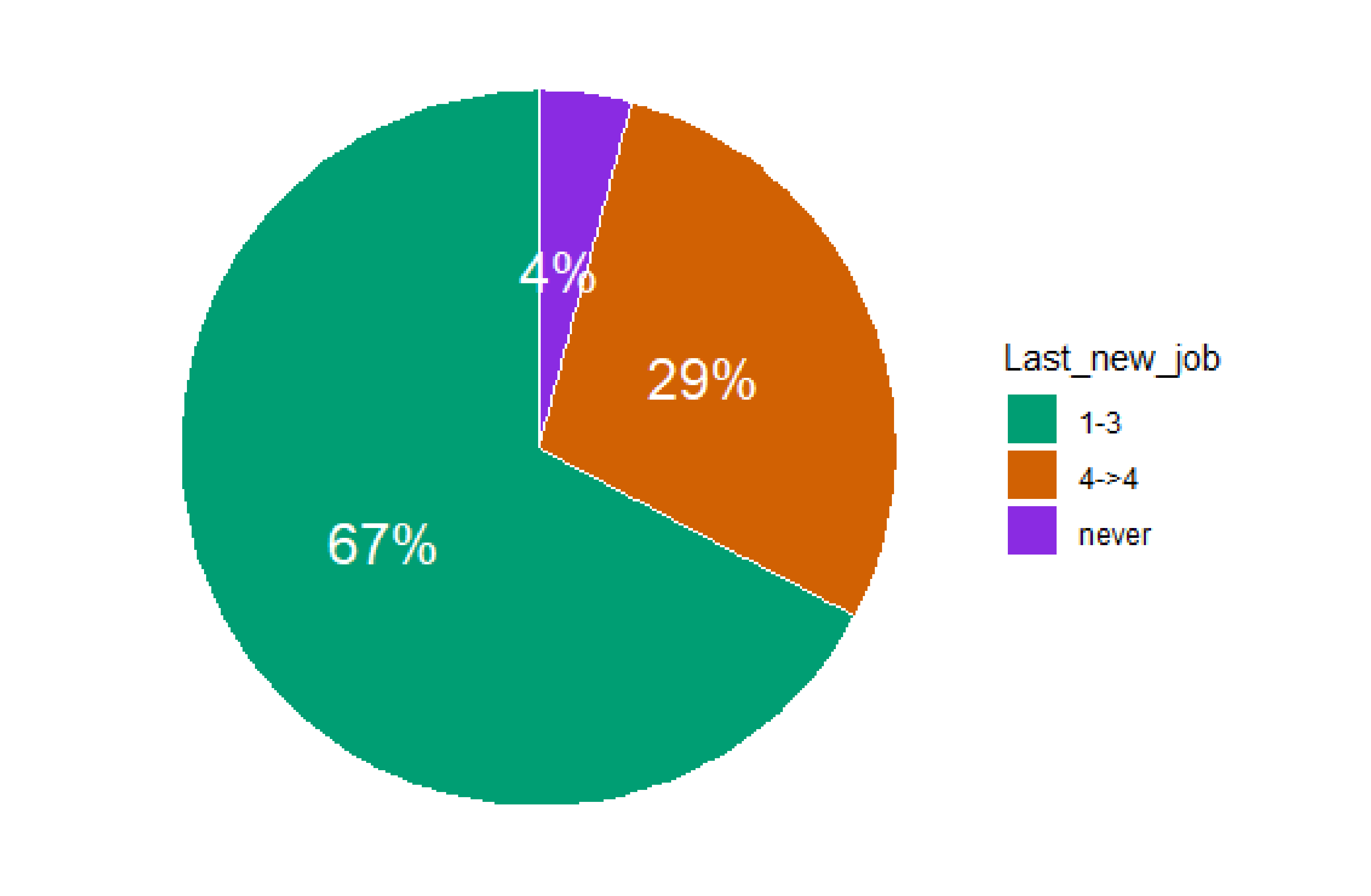
1

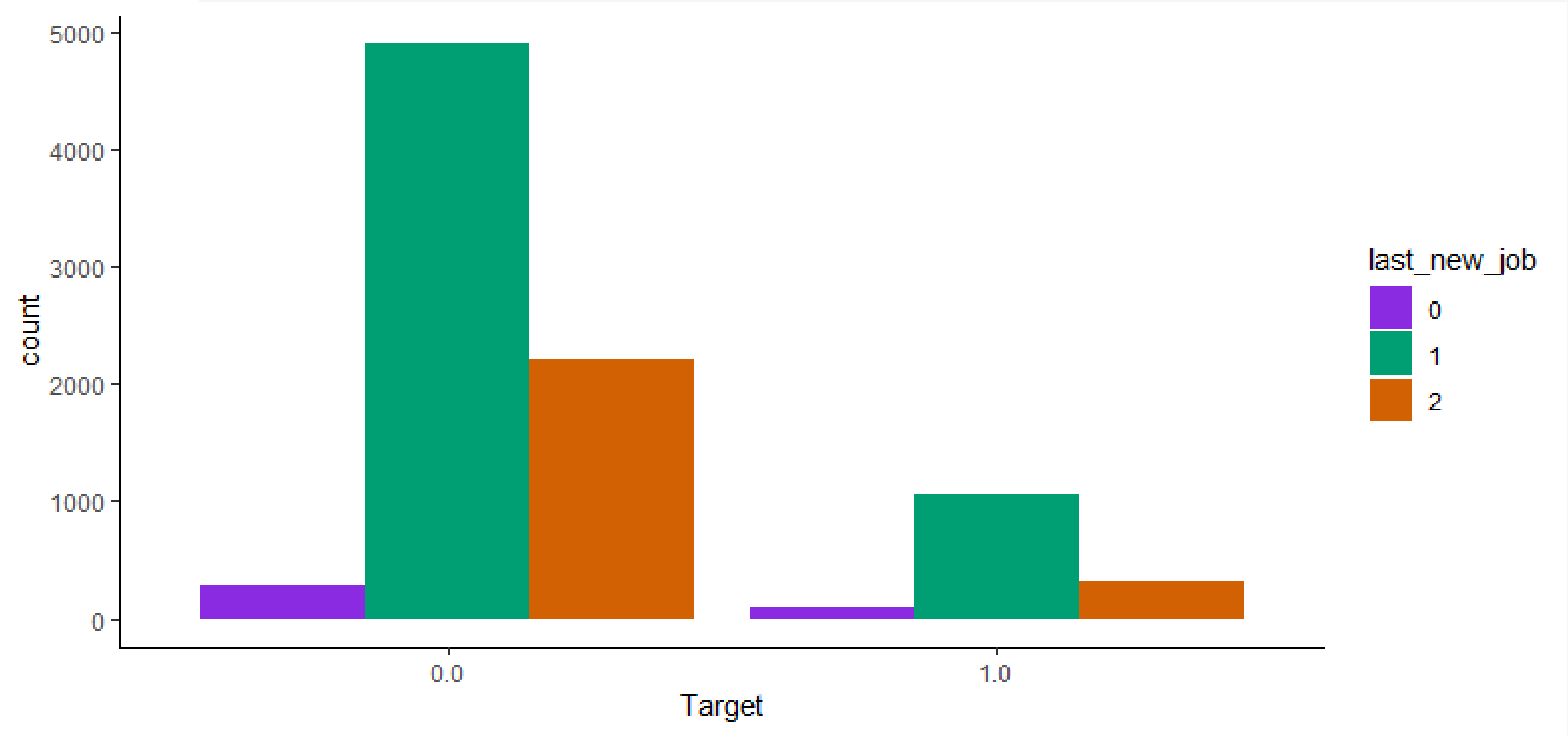




ANALISI ESPLORATIVA

LAST NEW JOB

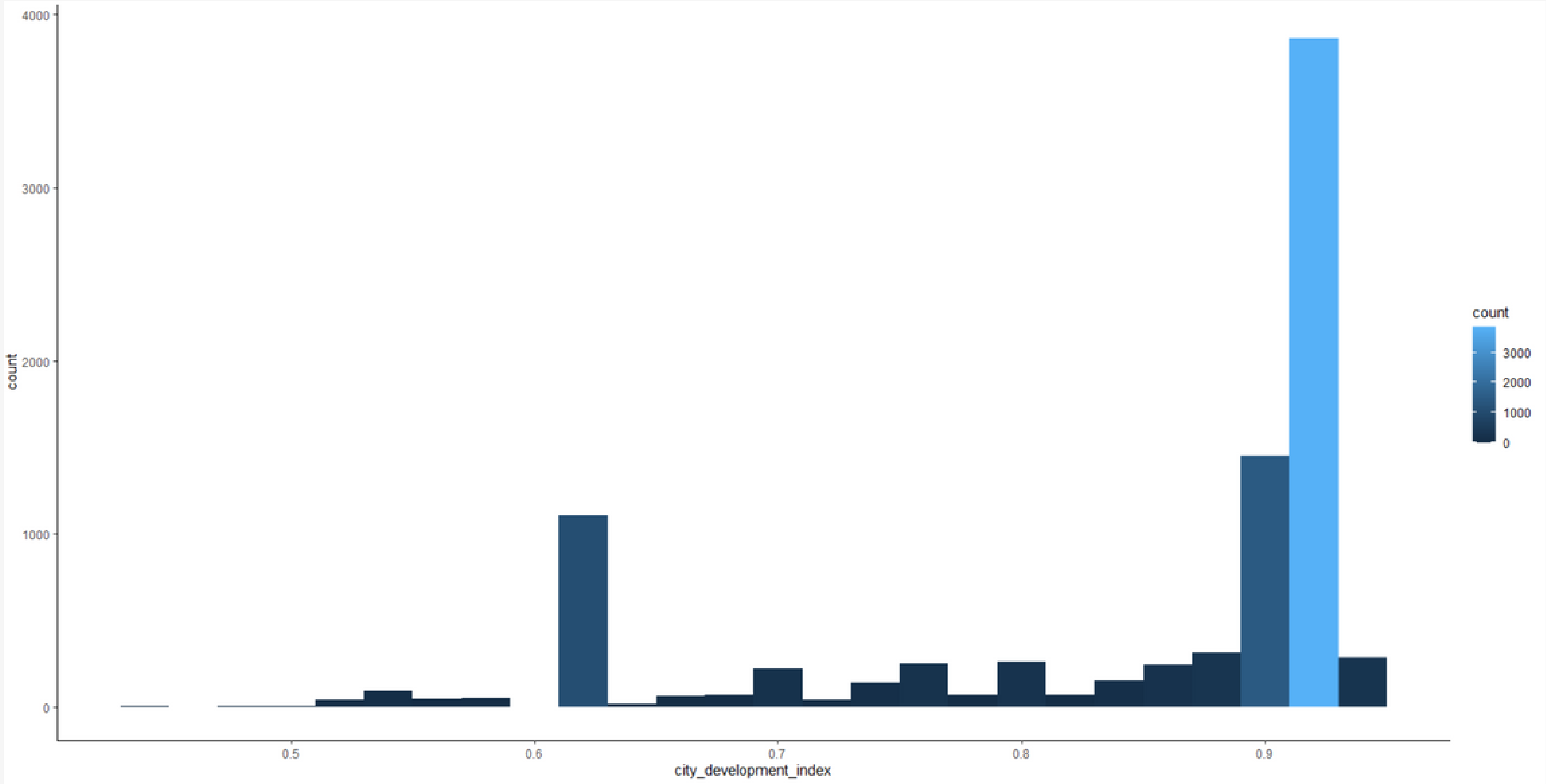


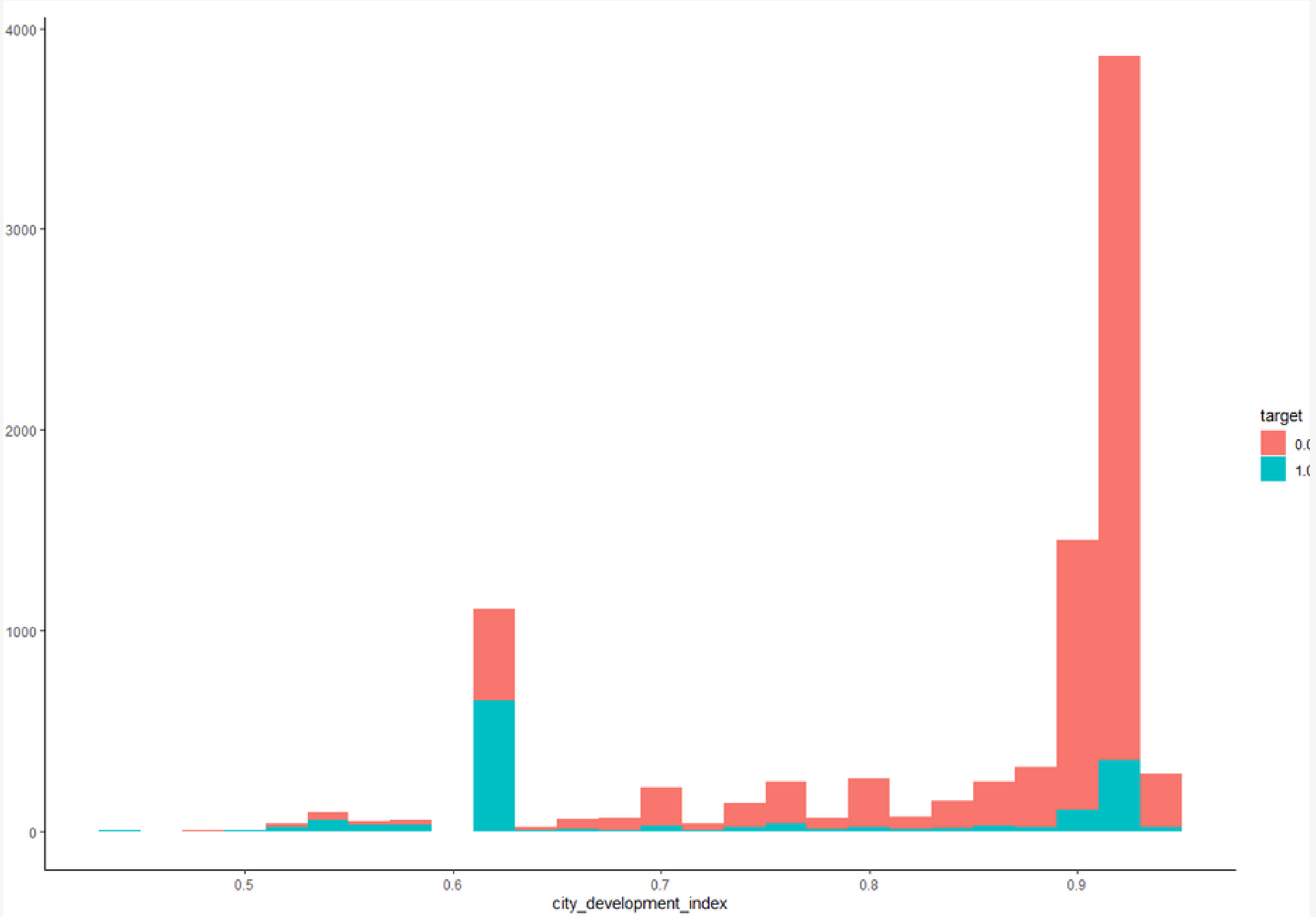


ANALISI ESPLORATIVA

CITY DEVELOPMENT INDEX

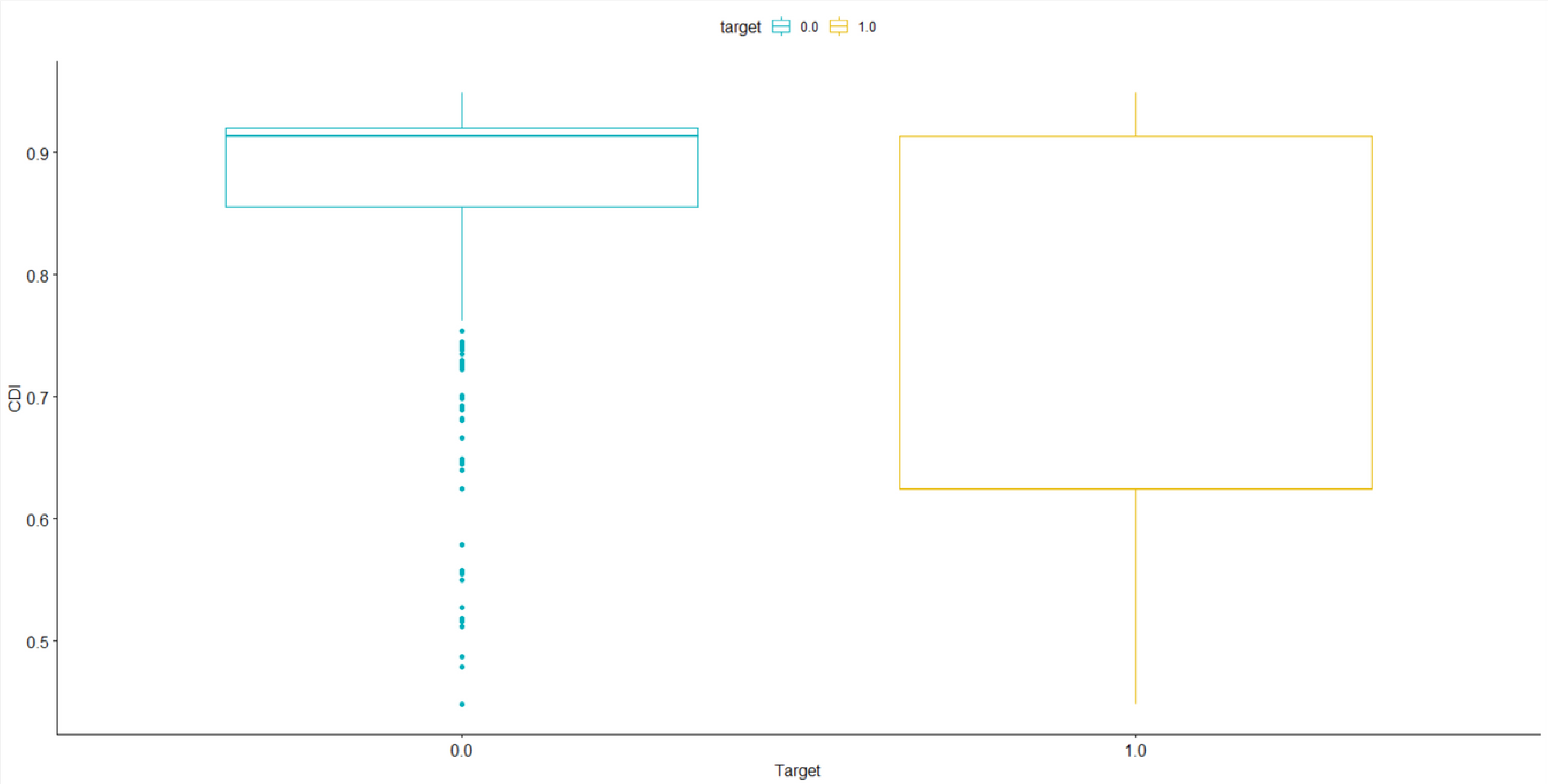
1





ANALISI ESPLORATIVA

CITY DEVELOPMENT INDEX

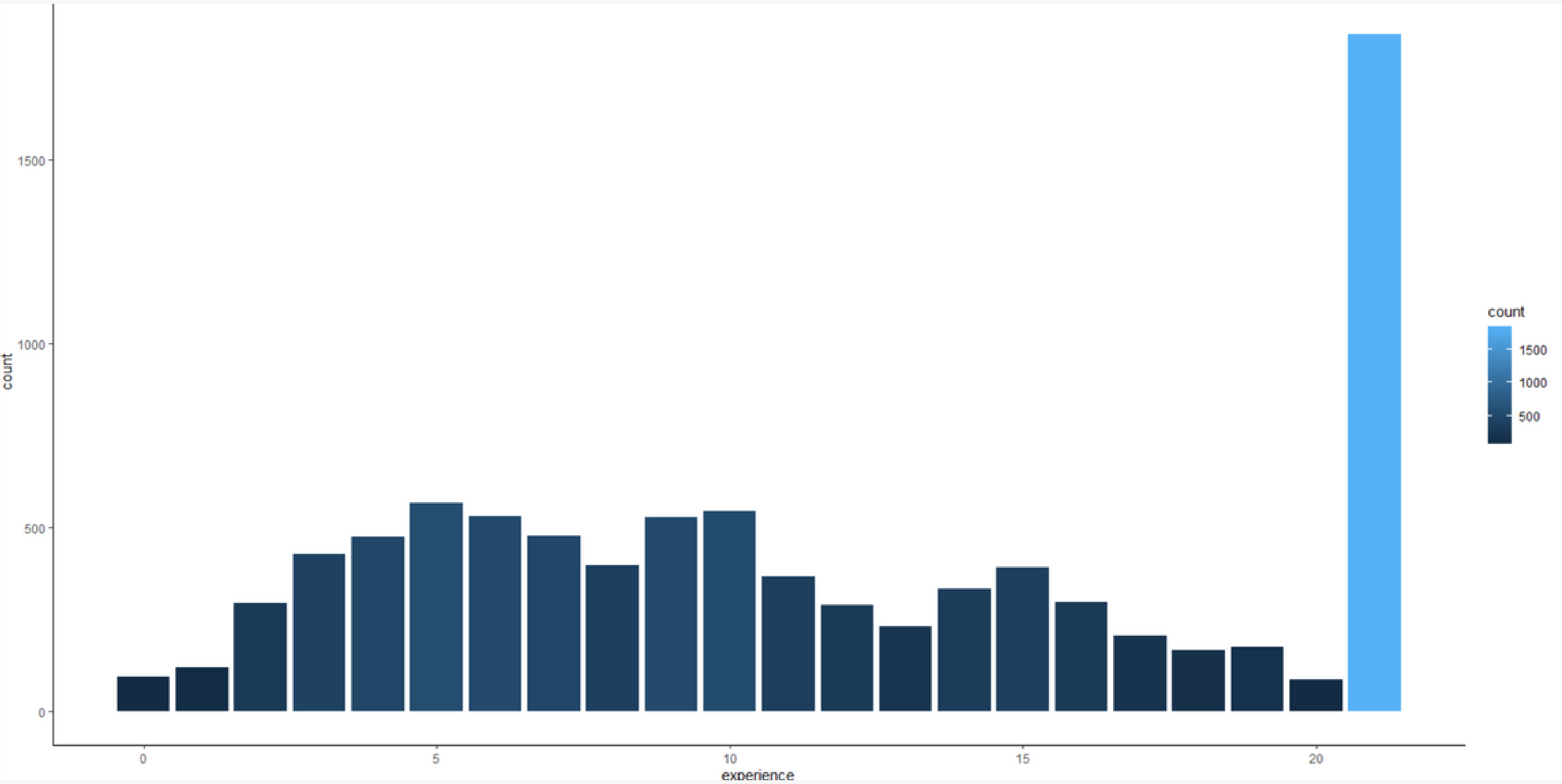


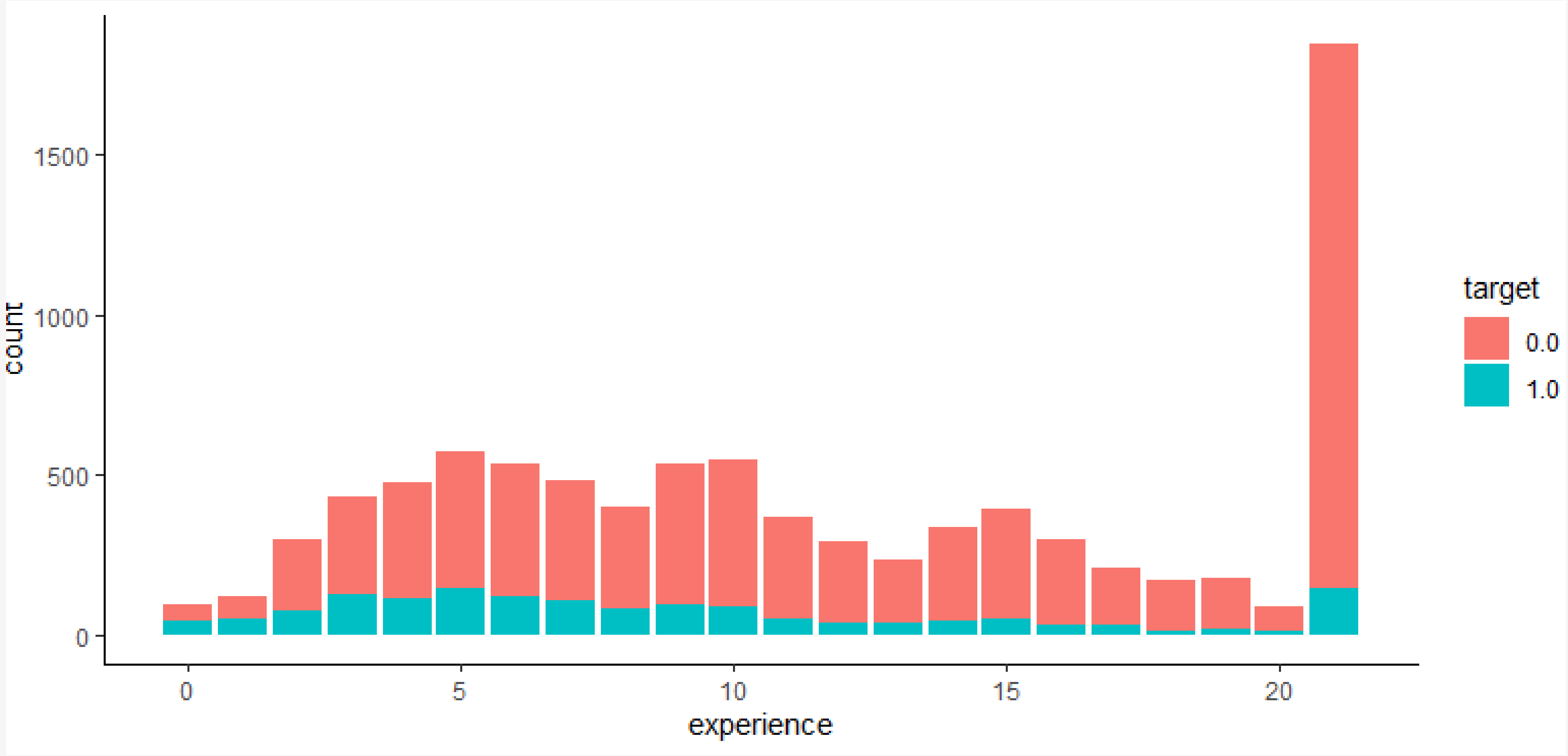

```
> t.test(city_development_index~target, data=data_set)

	Welch Two Sample t-test

data:  city_development_index by target
t = 33.867, df = 1740.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1244962 0.1398025
sample estimates:
mean in group 0.0 mean in group 1.0 
        0.8659320         0.7337827 
```

ANALISI ESPLORATIVA

EXPERIENCE

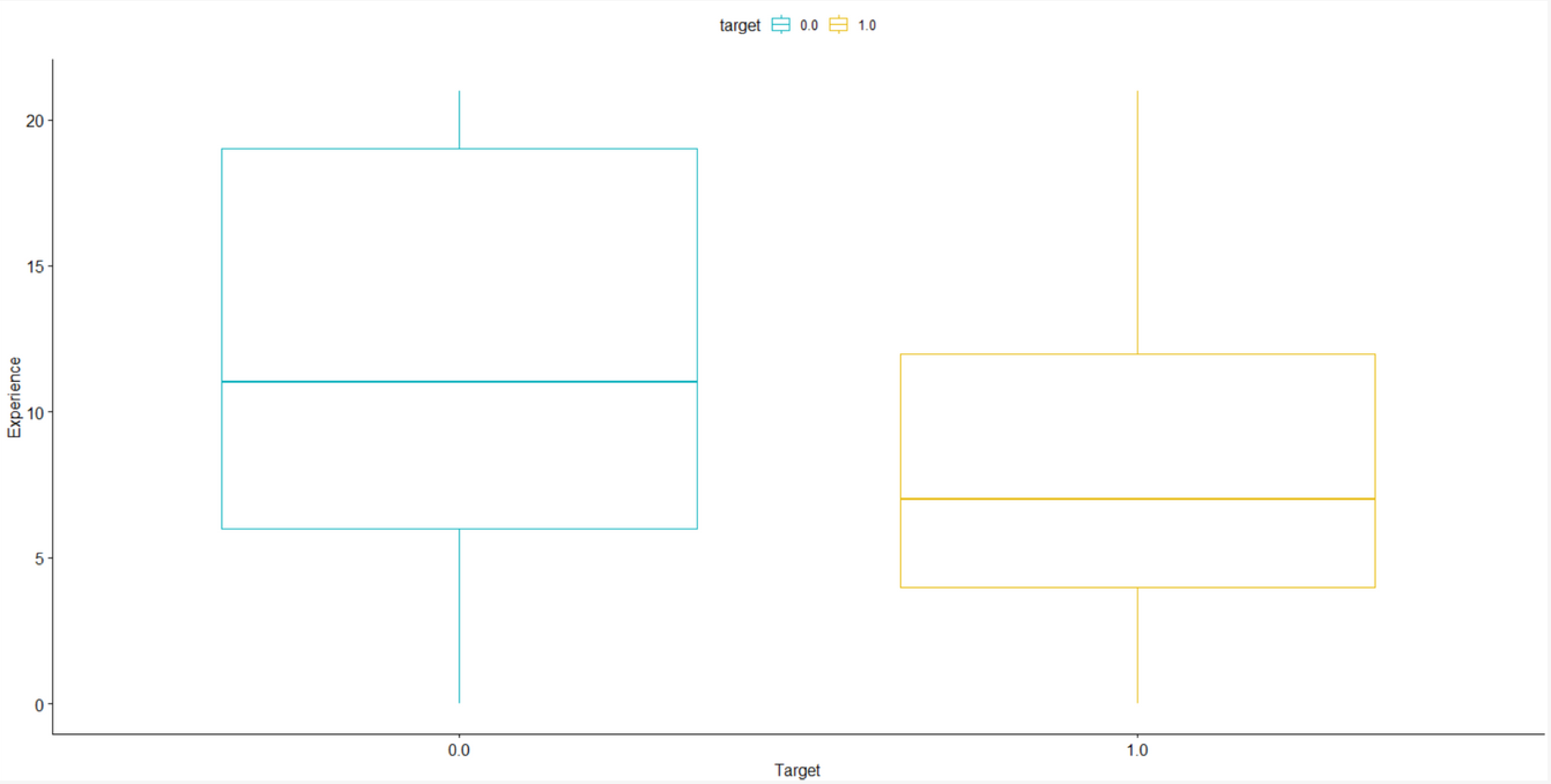
1





ANALISI ESPLORATIVA

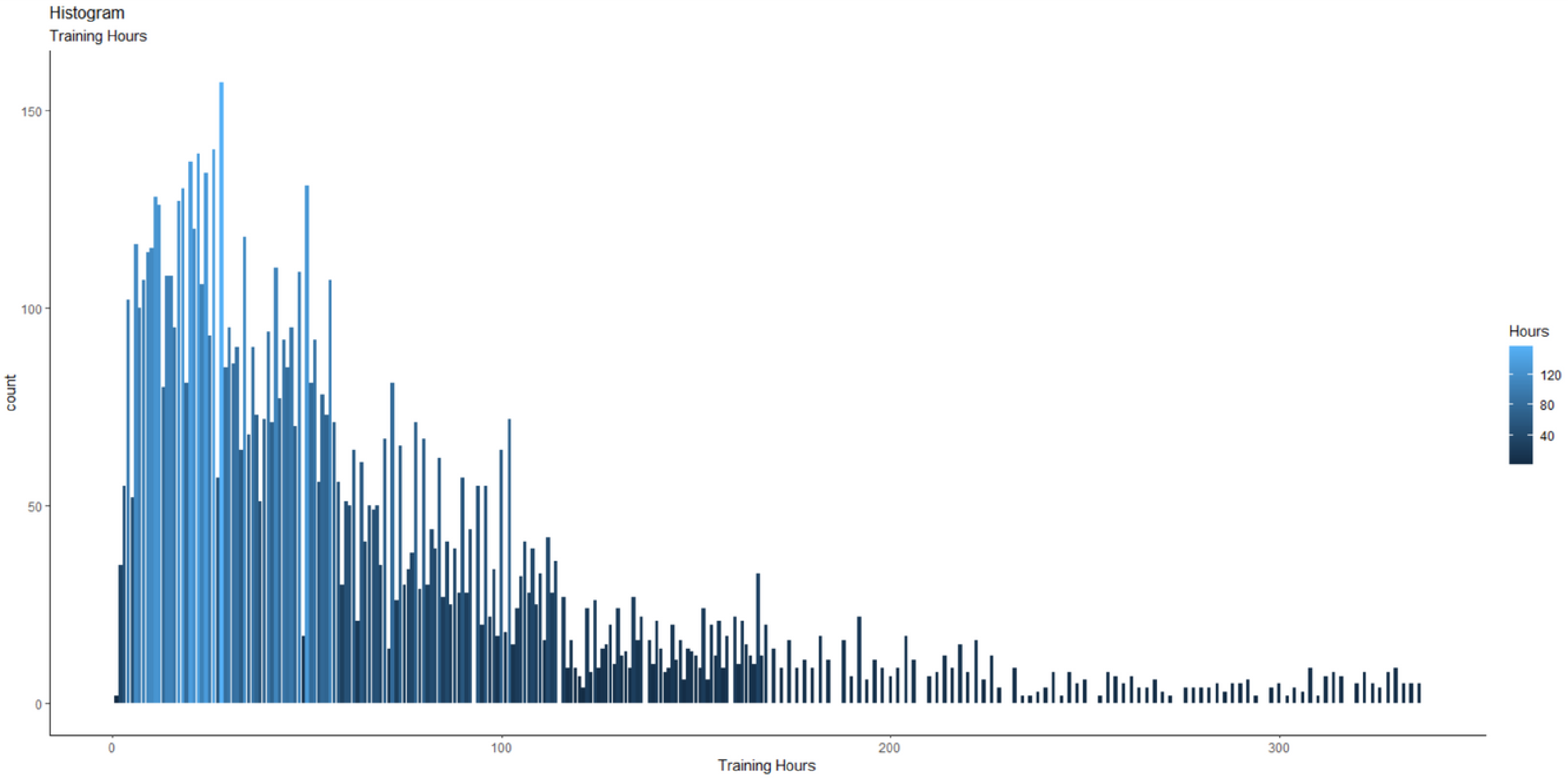
EXPERIENCE vs TARGET

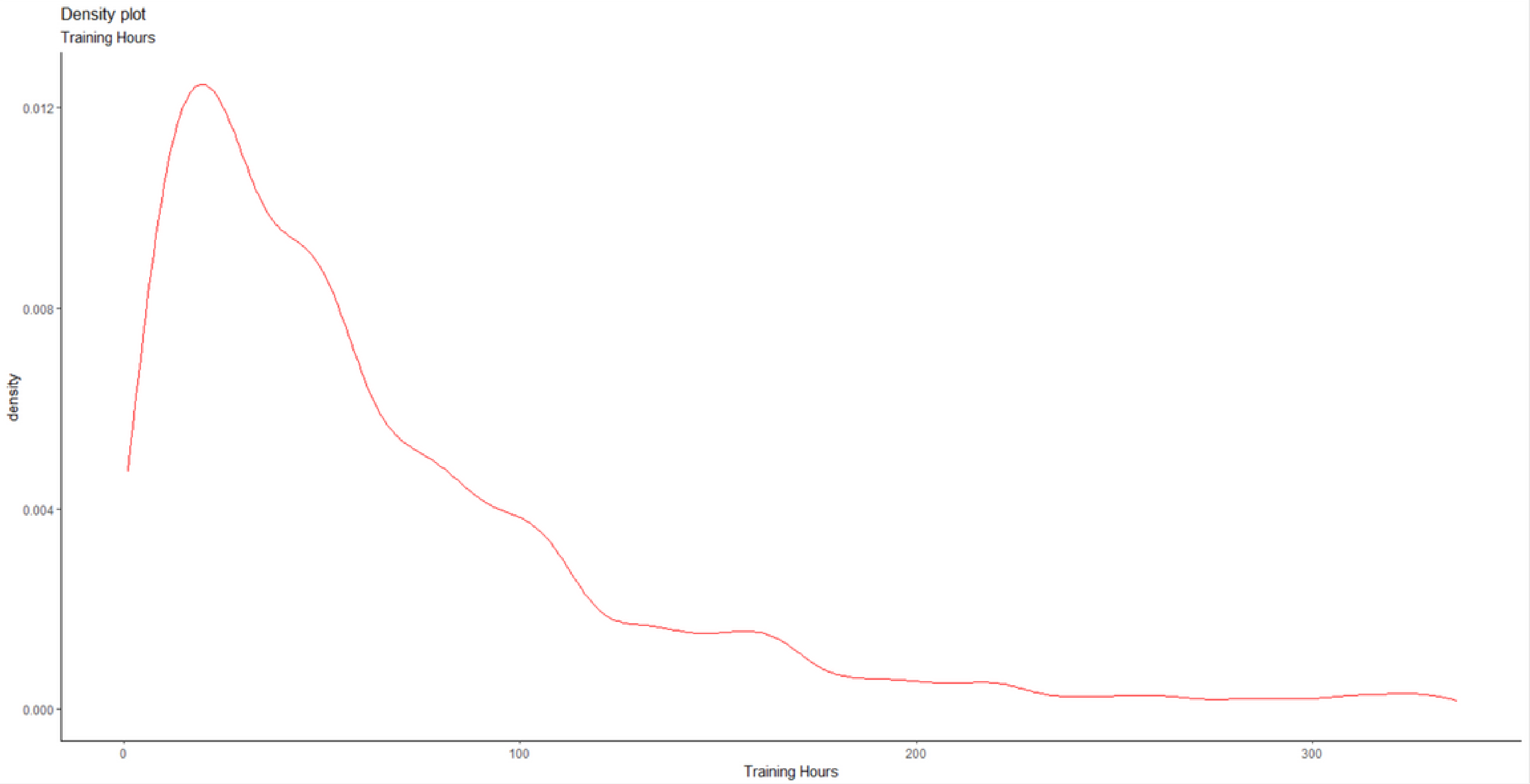



```
> t.test(experience~target, data=data_set)

	Welch Two Sample t-test

data:  experience by target
t = 20.113, df = 2221, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.135590 3.813108
sample estimates:
mean in group 0.0 mean in group 1.0 
         12.19029           8.71594 
```

ANALISI ESPLORATIVA

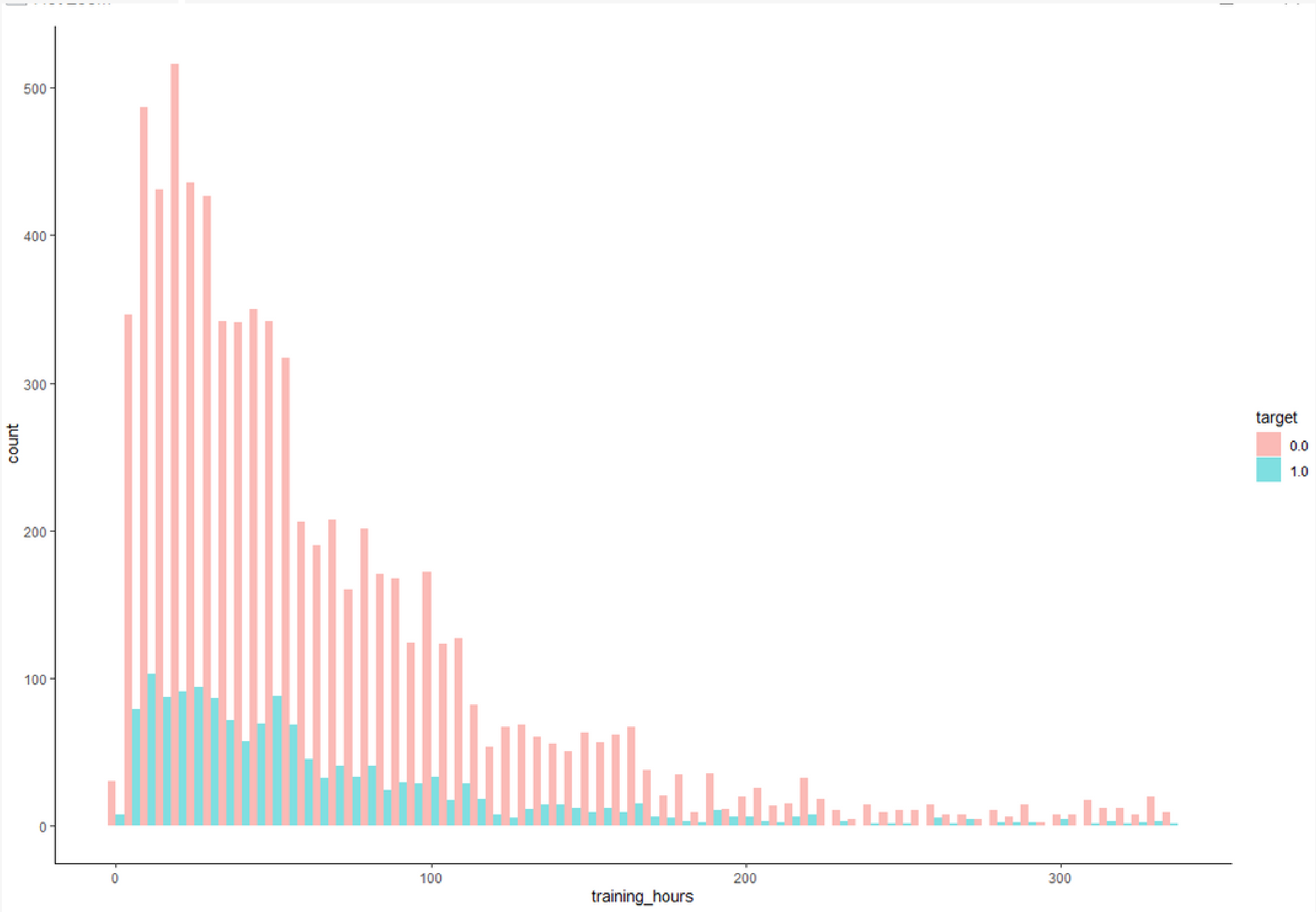
TRAINING HOURS

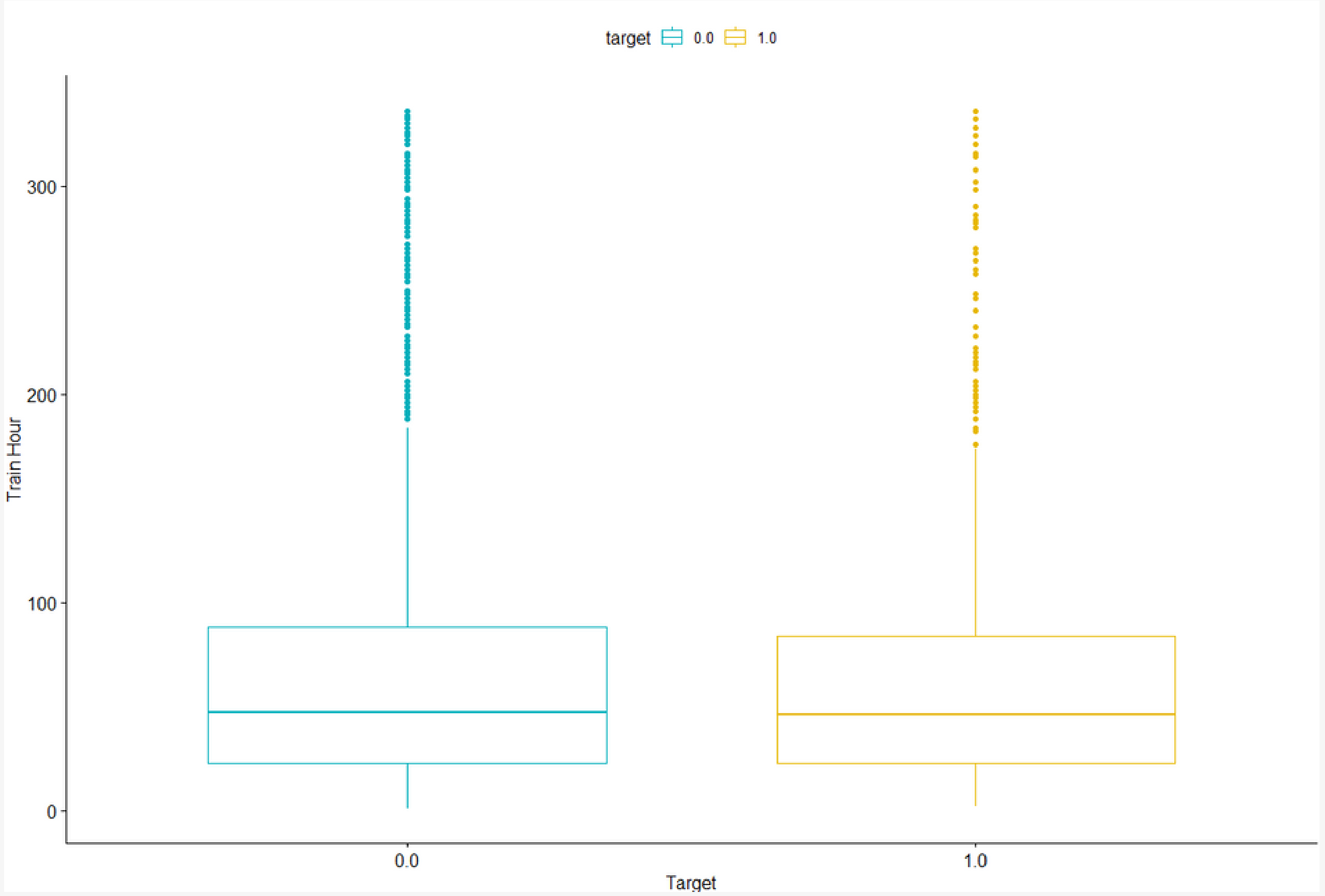




ANALISI ESPLORATIVA

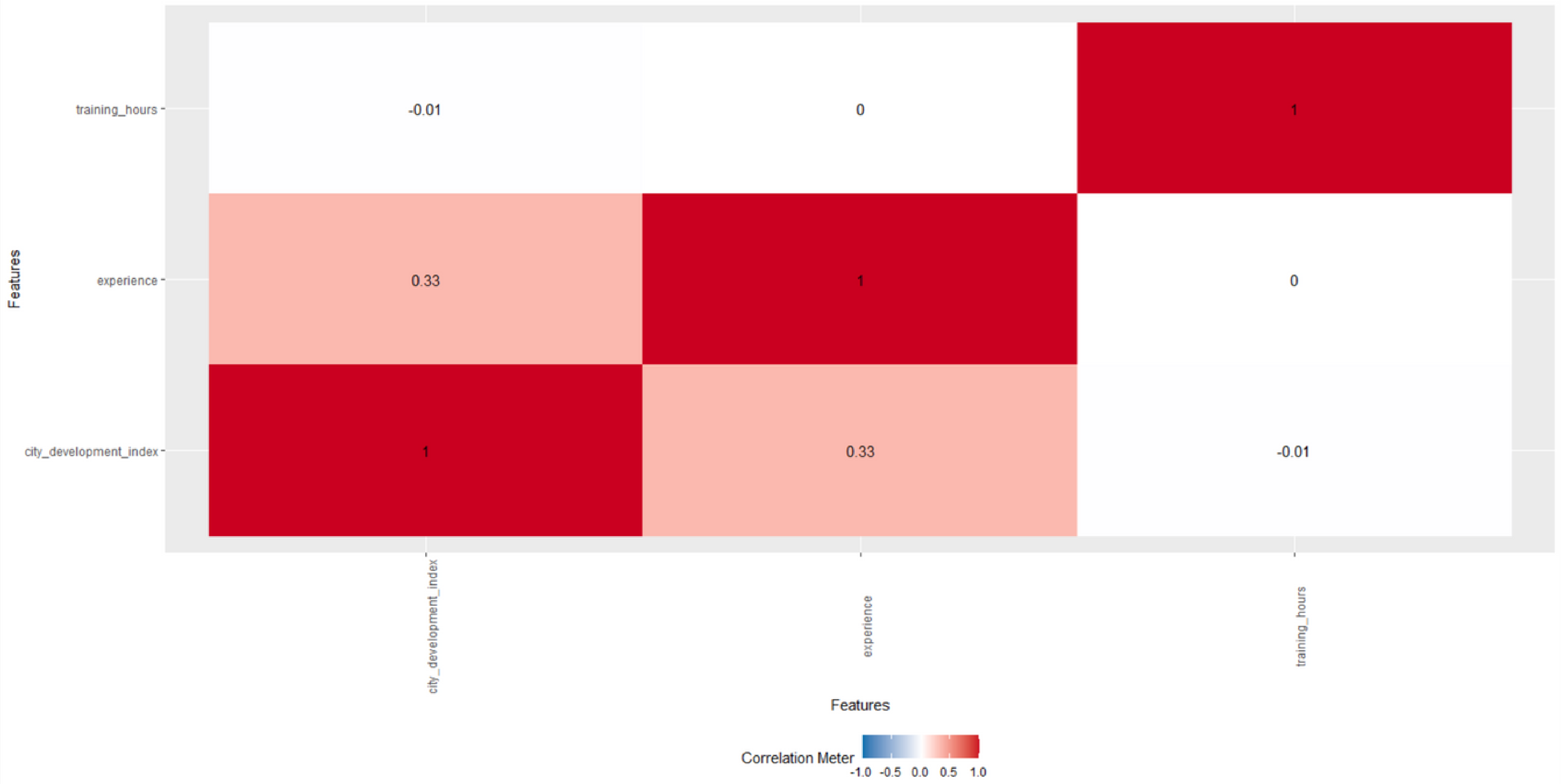
TRAINING HOURS VS TARGET

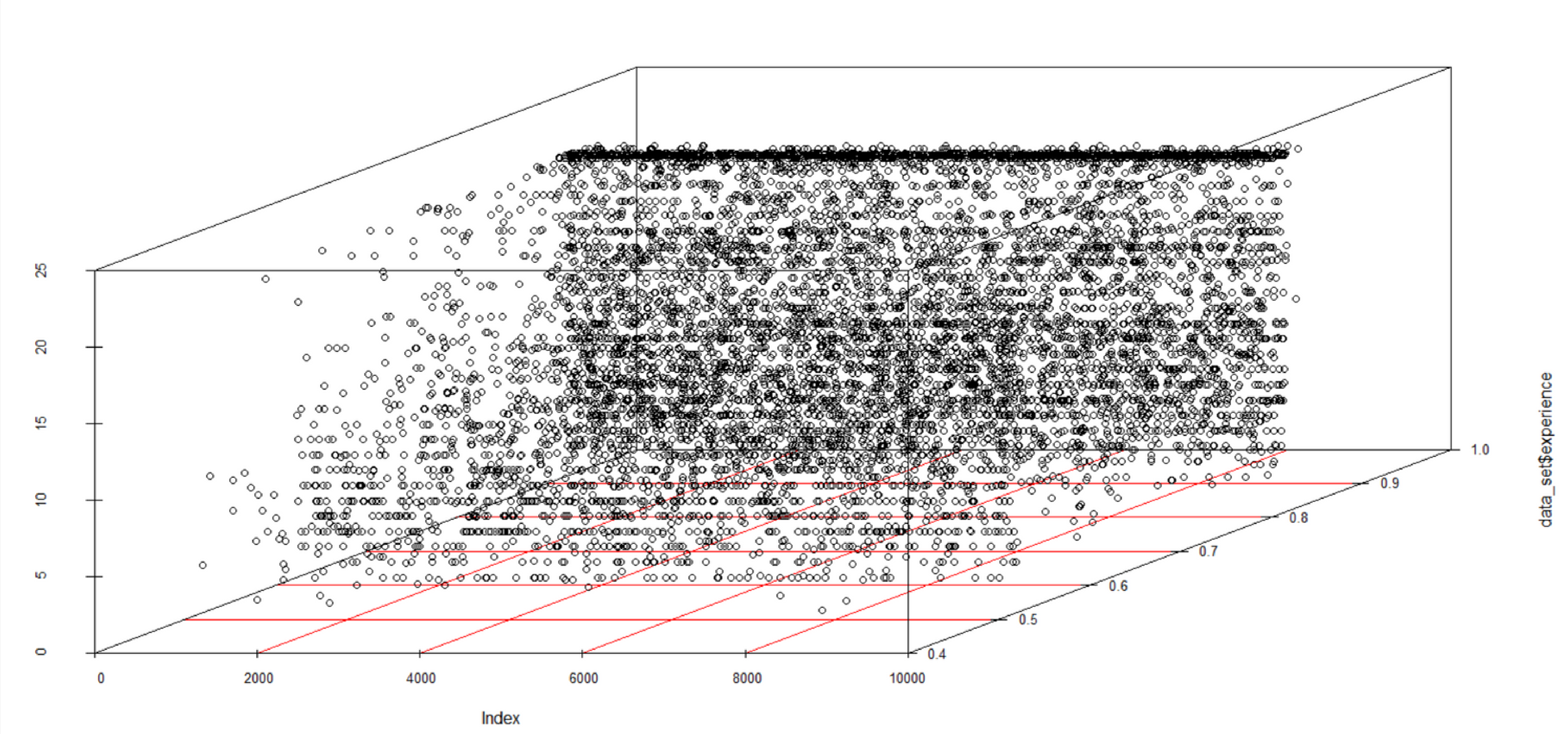





```
> t.test(training_hours~ target, data=data_set)

	Welch Two Sample t-test

data:  training_hours by target
t = 1.1746, df = 2121.5, p-value = 0.2403
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.337343  5.332254
sample estimates:
mean in group 0.0 mean in group 1.0 
         65.59623          63.59877 
```

CORRELATION PLOT





ANALISI DELLA CORRELAZIONE TRA CITY DEVELOPMENT INDEX E EXPERIENCE

```
> cor.test(data_set$city_development_index, data_set$experience, method=c("pearson"))

	Pearson's product-moment correlation

data:  data_set$city_development_index and data_set$experience
t = 33.203, df = 8839, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3143448 0.3514140
sample estimates:
      cor
0.3330081
```

MULTICOLLINEARITY CHECK

```
checkmulticol<- lm(city_development_index~., data = train_df[,-12] )
summary(checkmulticol)
vif(checkmulticol)
```

```
> vif(checkmulticol)
                        GVIF Df GVIF^(1/(2*Df))
gender              1.029681  1        1.014732
relevent_experience 1.109429  1        1.053294
enrolled_university 1.108719  1        1.052957
education_level     1.089362  2        1.021629
major_discipline    1.038403  3        1.006300
experience          1.131933  1        1.063923
company_size        1.100510  5        1.009623
company_type        1.179167  3        1.027849
last_new_job        1.010763  1        1.005367
training_hours      1.004351  1        1.002173
```

MODEL SELECTION WITH STEPWISE METHOD

1

```
Start:  AIC=5262.03
target ~ city_development_index + gender + relevent_experience +
    enrolled_university + education_level + major_discipline +
    experience + company_size + company_type + last_new_job +
    training_hours

                         Df Deviance    AIC
- major_discipline        3   5220.1 5258.1
- company_type            3   5220.4 5258.4
- gender                  1   5218.5 5260.5
- last_new_job            2   5220.6 5260.6
- enrolled_university     1   5218.9 5260.9
- relevent_experience     1   5219.8 5261.8
<none>                        5218.0 5262.0
- education_level         2   5222.1 5262.1
- training_hours          1   5220.6 5262.6
- company_size            5   5229.3 5263.3
- experience              1   5249.9 5291.9
- city_development_index  1   6019.8 6061.8
```

2

```
Step:  AIC=5258.06
target ~ city_development_index + gender + relevent_experience +
    enrolled_university + education_level + experience + company_size +
    company_type + last_new_job + training_hours

                         Df Deviance    AIC
- company_type            3   5222.4 5254.4
- gender                  1   5220.4 5256.4
- last_new_job            2   5222.6 5256.6
- enrolled_university     1   5221.0 5257.0
- relevent_experience     1   5221.6 5257.6
<none>                        5220.1 5258.1
- education_level         2   5224.2 5258.2
- training_hours          1   5222.5 5258.5
- company_size            5   5231.6 5259.6
- experience              1   5251.2 5287.2
- city_development_index  1   6052.5 6088.5
```

3

```
Step:  AIC=5254.38
target ~ city_development_index + gender + relevent_experience +
    enrolled_university + education_level + experience + company_size +
    last_new_job + training_hours

                         Df Deviance    AIC
- gender                  1   5222.8 5252.8
- last_new_job            2   5225.3 5253.3
- enrolled_university     1   5223.6 5253.6
<none>                        5222.4 5254.4
- relevent_experience     1   5224.6 5254.6
- training_hours          1   5225.0 5255.0
- education_level         2   5227.0 5255.0
- company_size            5   5235.0 5257.0
- experience              1   5253.4 5283.4
- city_development_index  1   6053.4 6083.4
```

4

```
Step:  AIC=5252.77
target ~ city_development_index + relevent_experience + enrolled_university +
    education_level + experience + company_size + last_new_job +
    training_hours

                         Df Deviance    AIC
- last_new_job            2   5225.6 5251.6
- enrolled_university     1   5223.9 5251.9
<none>                        5222.8 5252.8
- relevent_experience     1   5225.2 5253.2
- training_hours          1   5225.4 5253.4
- education_level         2   5227.4 5253.4
- company_size            5   5235.3 5255.3
- experience              1   5254.7 5282.7
- city_development_index  1   6055.3 6083.3
```

5

```
Step:  AIC=5251.65
target ~ city_development_index + relevent_experience + enrolled_university +
    education_level + experience + company_size + training_hours

                         Df Deviance    AIC
- enrolled_university     1   5226.7 5250.7
<none>                        5225.6 5251.6
- relevent_experience     1   5228.2 5252.2
- education_level         2   5230.3 5252.3
- training_hours          1   5228.4 5252.4
- company_size            5   5238.6 5254.6
- experience              1   5254.9 5278.9
- city_development_index  1   6062.6 6086.6
```

6

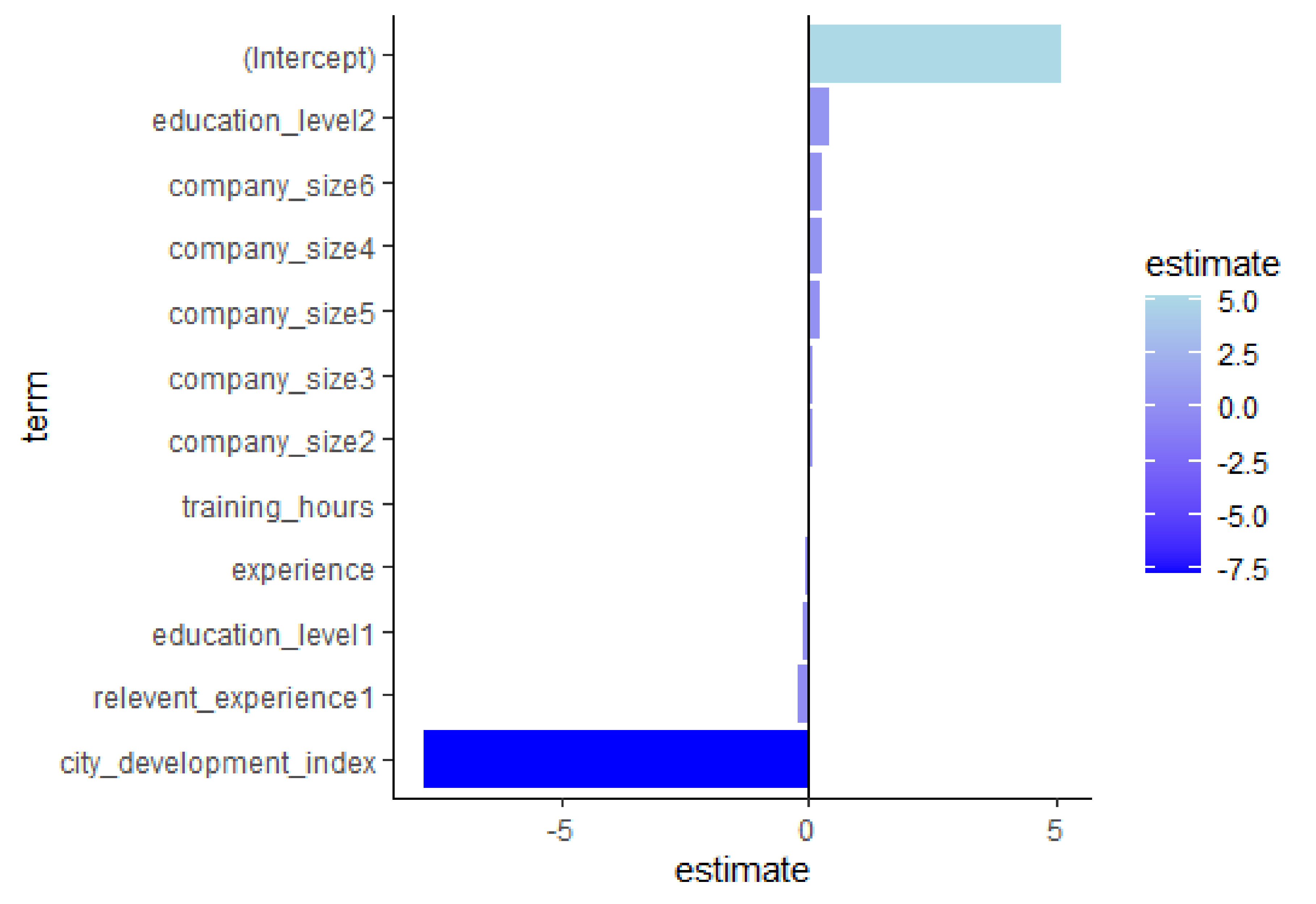
```
Step:  AIC=5250.67
target ~ city_development_index + relevent_experience + education_level +
    experience + company_size + training_hours

                         Df Deviance    AIC
<none>                        5226.7 5250.7
- education_level         2   5231.3 5251.3
- training_hours          1   5229.3 5251.3
- relevent_experience     1   5229.7 5251.7
- company_size            5   5239.7 5253.7
- experience              1   5259.3 5281.3
- city_development_index  1   6066.8 6088.8
```

ANALISI DEI COEFFICIENTI

```
Call:
glm(formula = target ~ experience + city_development_index +
    education_level + relevent_experience + company_size + training_hours,
    family = "binomial", data = train_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7922  -0.4809  -0.3968  -0.3284   2.5461

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            5.1065890  0.2379578  21.460  < 2e-16 ***
experience            -0.0358975  0.0063359  -5.666 1.46e-08 ***
city_development_index -7.7529803  0.2800297 -27.686  < 2e-16 ***
education_level1      -0.0794270  0.0824577  -0.963 0.335424
education_level2       0.4225684  0.2241889   1.885 0.059447 .
relevent_experience1  -0.1817405  0.1041000  -1.746 0.080841 .
company_size2          0.0895243  0.1020074   0.878 0.380147
company_size3          0.1131415  0.1551393   0.729 0.465824
company_size4          0.2685095  0.1335352   2.011 0.044349 *
company_size5          0.2353879  0.1819669   1.294 0.195812
company_size6          0.3065133  0.0919523   3.333 0.000858 ***
training_hours        -0.0009686  0.0005993  -1.616 0.106065
---
```

EXP OF THE COEFFICIENTS

(Intercept)	1.651062e+02
experience	9.647392e-01
city_development_index	4.294607e-04
education_level1	9.236454e-01
education_level2	1.525876e+00
relevent_experience1	8.338177e-01
company_size2	1.093654e+00
company_size3	1.119790e+00
company_size4	1.308013e+00
company_size5	1.265400e+00
company_size6	1.358680e+00
training_hours	9.990319e-01



CONFUSION MATRIX

CUT-OFF 0.5

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1417  202
         1   57   91

               Accuracy : 0.8534
                 95% CI : (0.8361, 0.8696)
    No Information Rate : 0.8342
    P-Value [Acc > NIR] : 0.015

                  Kappa : 0.3391

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.31058
            Specificity : 0.96133
         Pos Pred Value : 0.61486
         Neg Pred Value : 0.87523
             Prevalence : 0.16582
         Detection Rate : 0.05150
   Detection Prevalence : 0.08376
      Balanced Accuracy : 0.63595

       'Positive' Class : 1
```

TEST ERROR RATE: 0.1465761

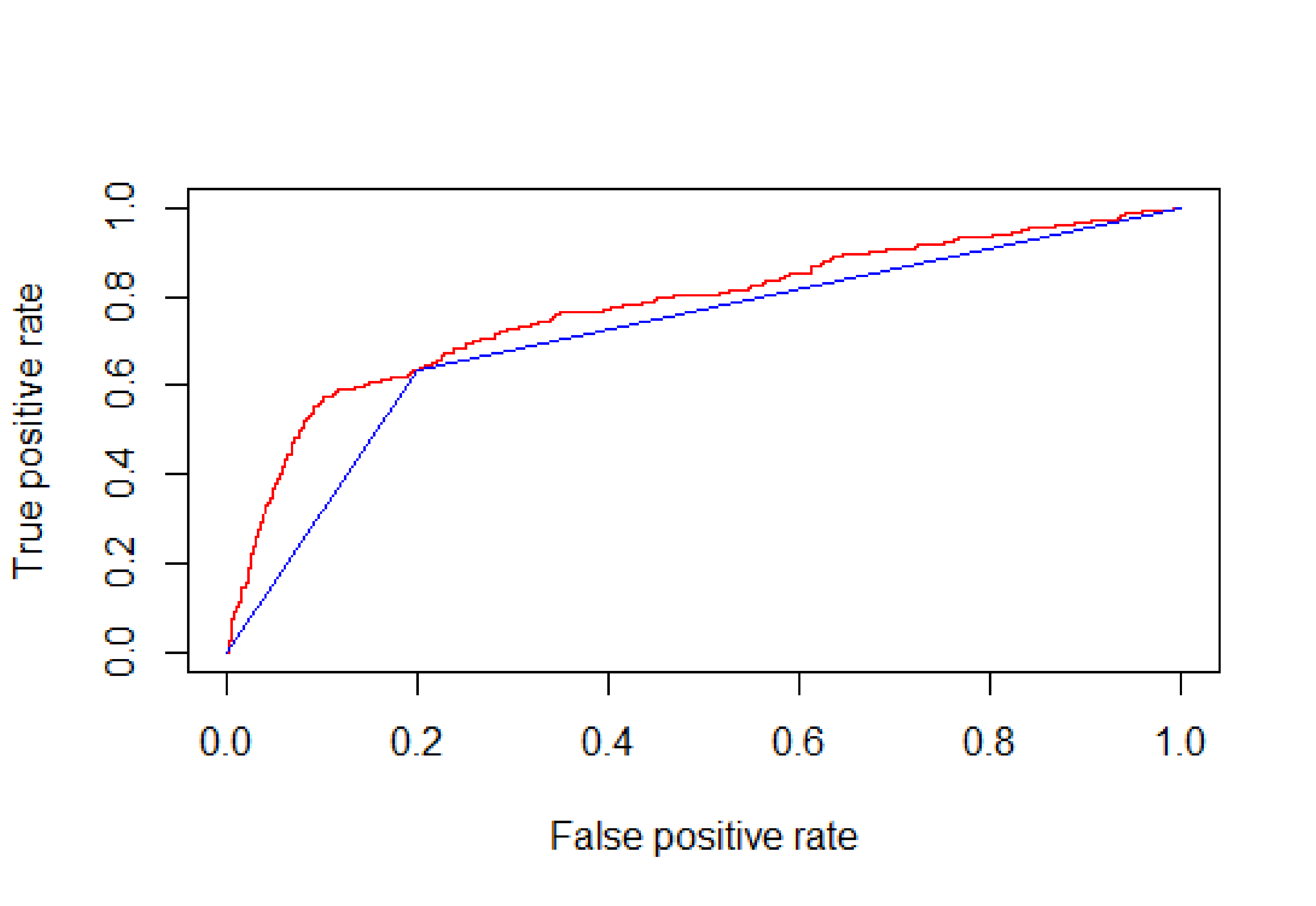
CUT-OFF 0.16

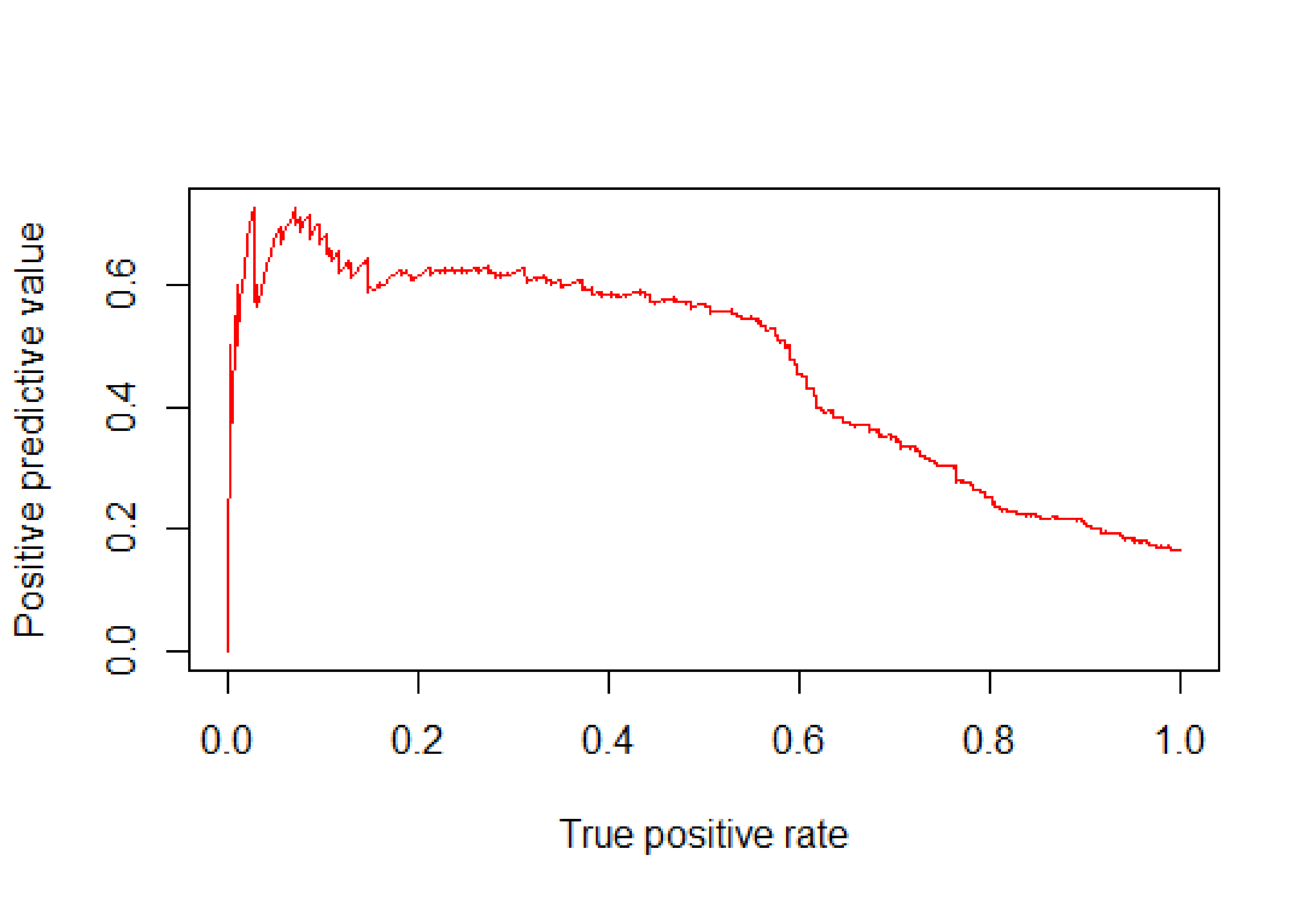
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1181  107
         1  293  186

               Accuracy : 0.7736
                 95% CI : (0.7534, 0.793)
    No Information Rate : 0.8342
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3476

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6348
            Specificity : 0.8012
         Pos Pred Value : 0.3883
         Neg Pred Value : 0.9169
             Prevalence : 0.1658
         Detection Rate : 0.1053
   Detection Prevalence : 0.2711
      Balanced Accuracy : 0.7180

       'Positive' Class : 1
```

TEST ERROR RATE: 0.2263724

ROC CURVE

PRECISION-RECALL CURVE





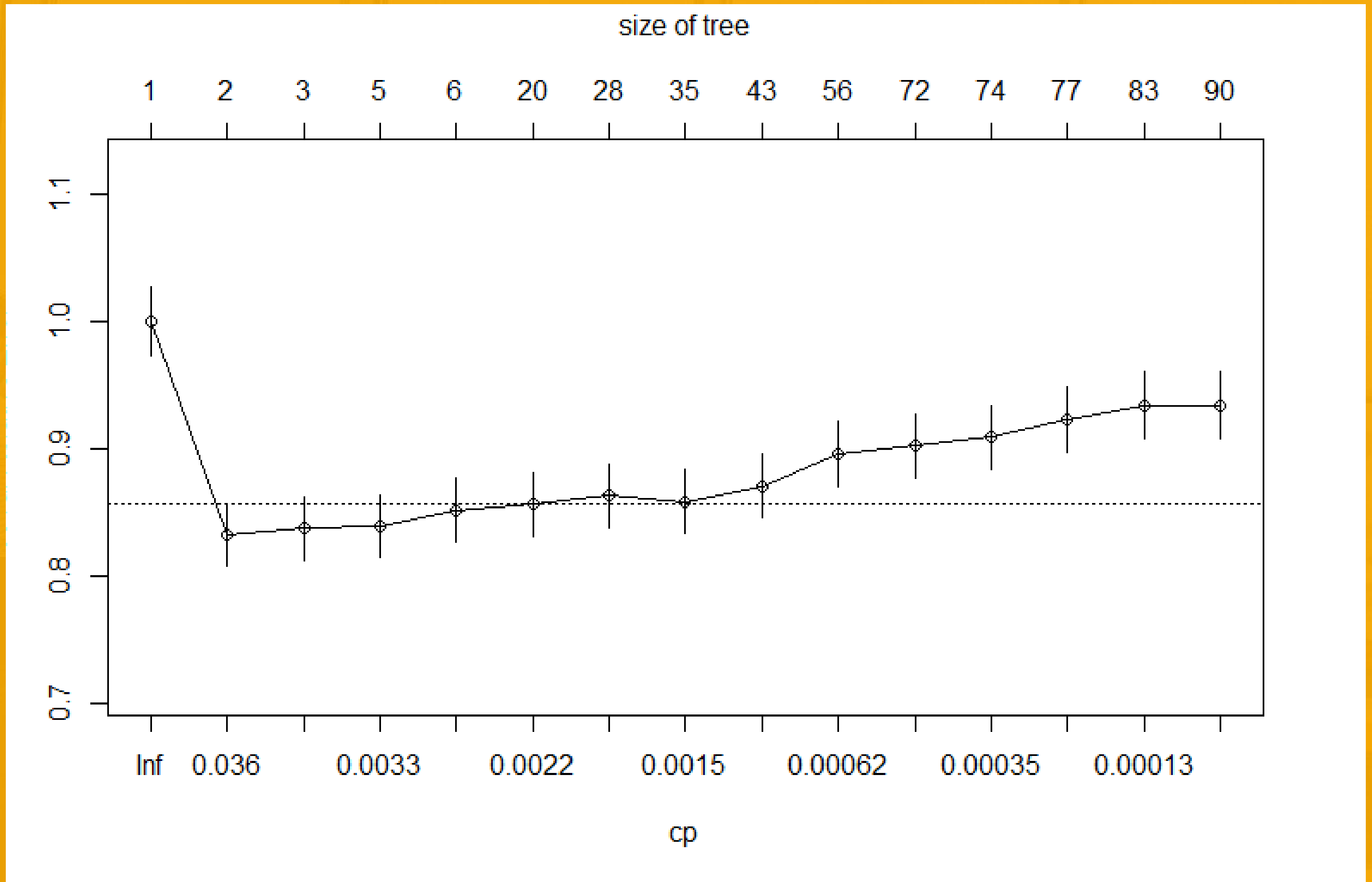
AUC

```
       [,1]      [,2]
[1,] 0.7743203 0.7180167
```

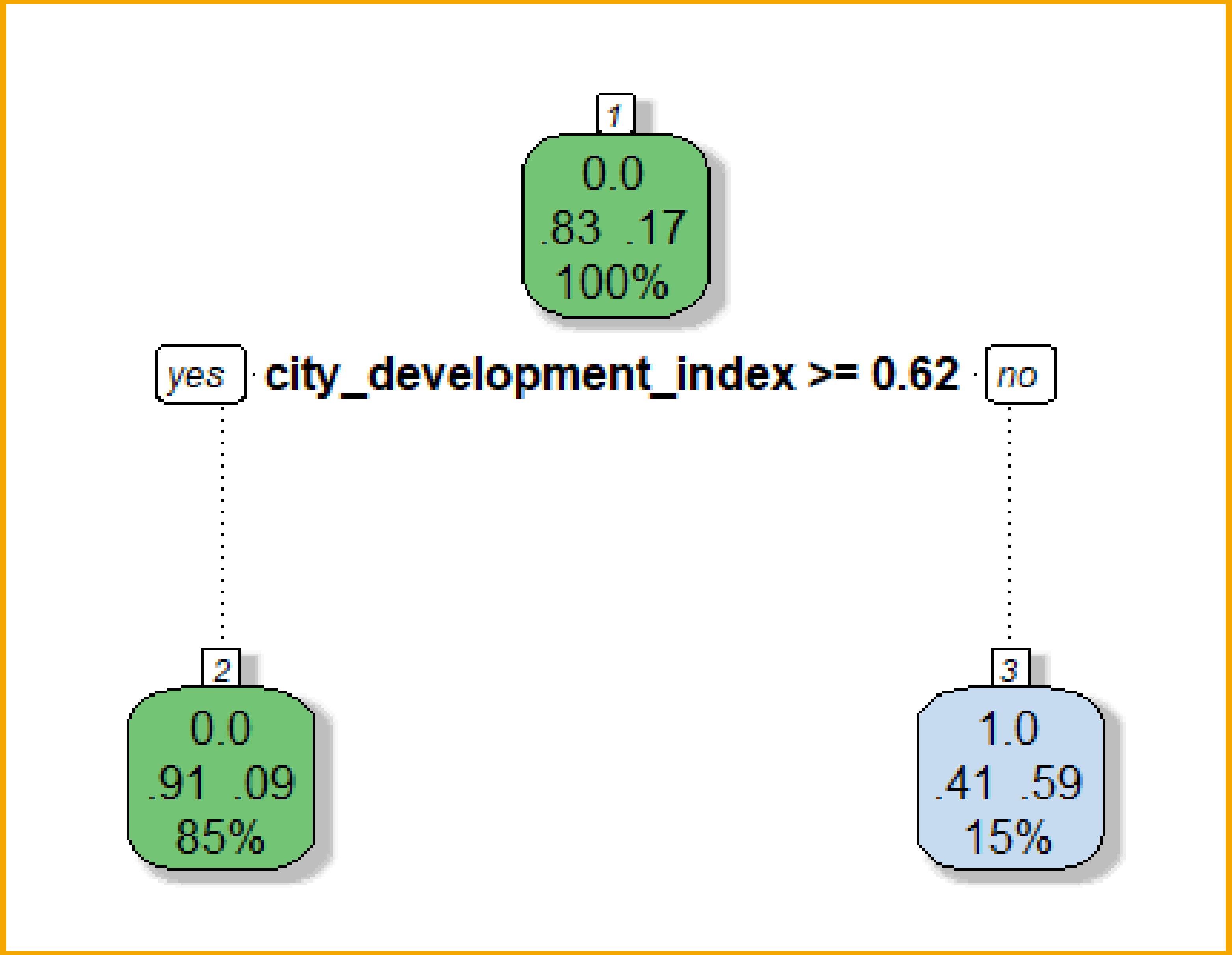
DECISION TREES

1

Choose the tree that CP minimizes the X error rate



	CP	nsplit	rel error	xerror	xstd
1	0.16851064	0	1.00000	1.00000	0.026640
2	0.00765957	1	0.83149	0.83234	0.024707
3	0.00425532	2	0.82383	0.83745	0.024771
4	0.00255319	4	0.81532	0.83915	0.024792
5	0.00226950	5	0.81277	0.85191	0.024949
6	0.00212766	19	0.76936	0.85617	0.025001
7	0.00170213	27	0.75149	0.86298	0.025083
8	0.00127660	34	0.73957	0.85872	0.025032
9	0.00085106	42	0.72936	0.87064	0.025176
10	0.00045827	55	0.71660	0.89617	0.025479
11	0.00042553	71	0.70894	0.90213	0.025548
12	0.00028369	73	0.70809	0.90894	0.025628
13	0.00014184	76	0.70723	0.92340	0.025794
14	0.00012158	82	0.70638	0.93447	0.025920
15	0.00010000	89	0.70553	0.93447	0.025920



TRAIN ERROR RATE: 0.1381114

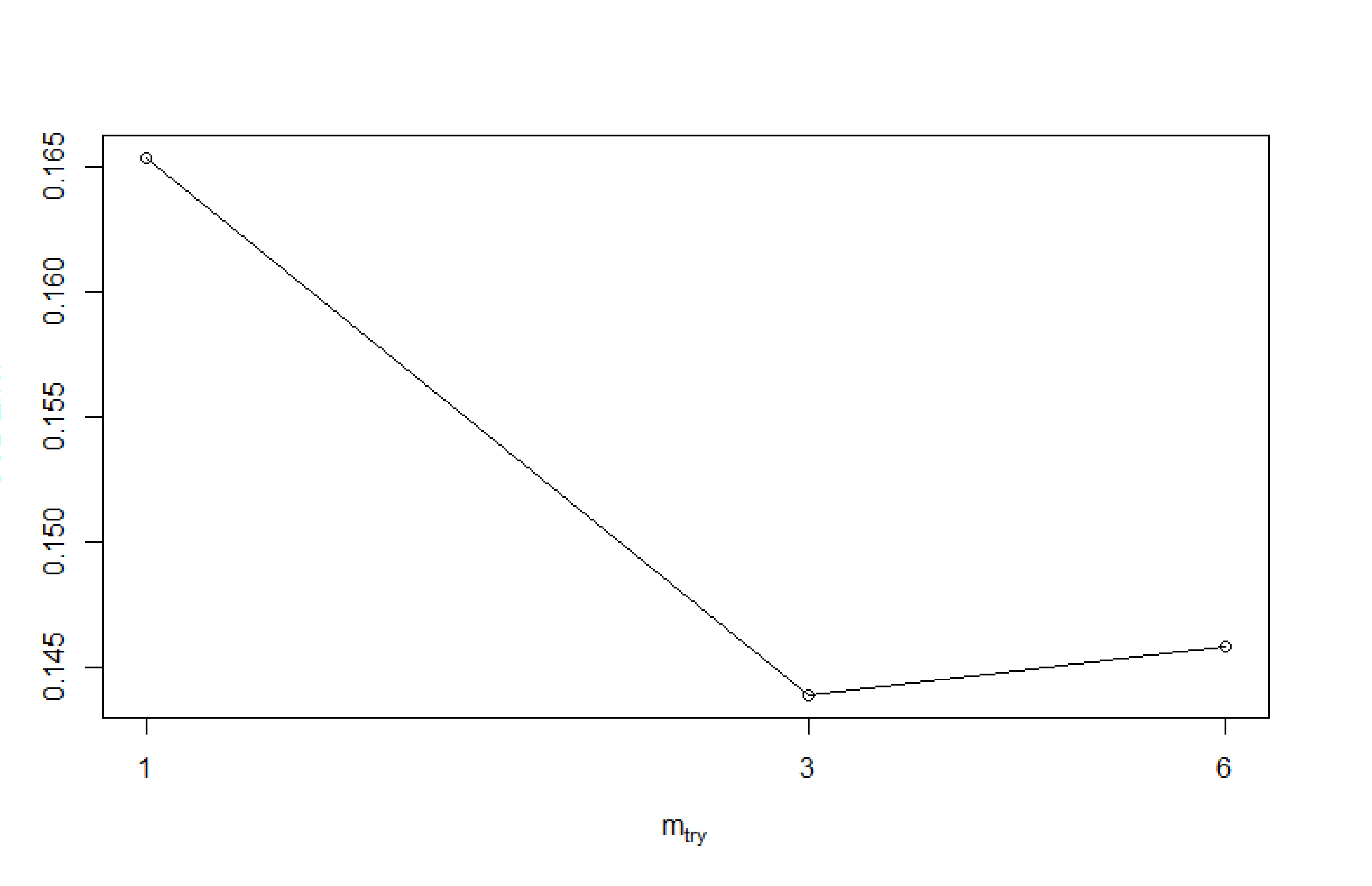
TEST ERROR RATE: 0.1437465

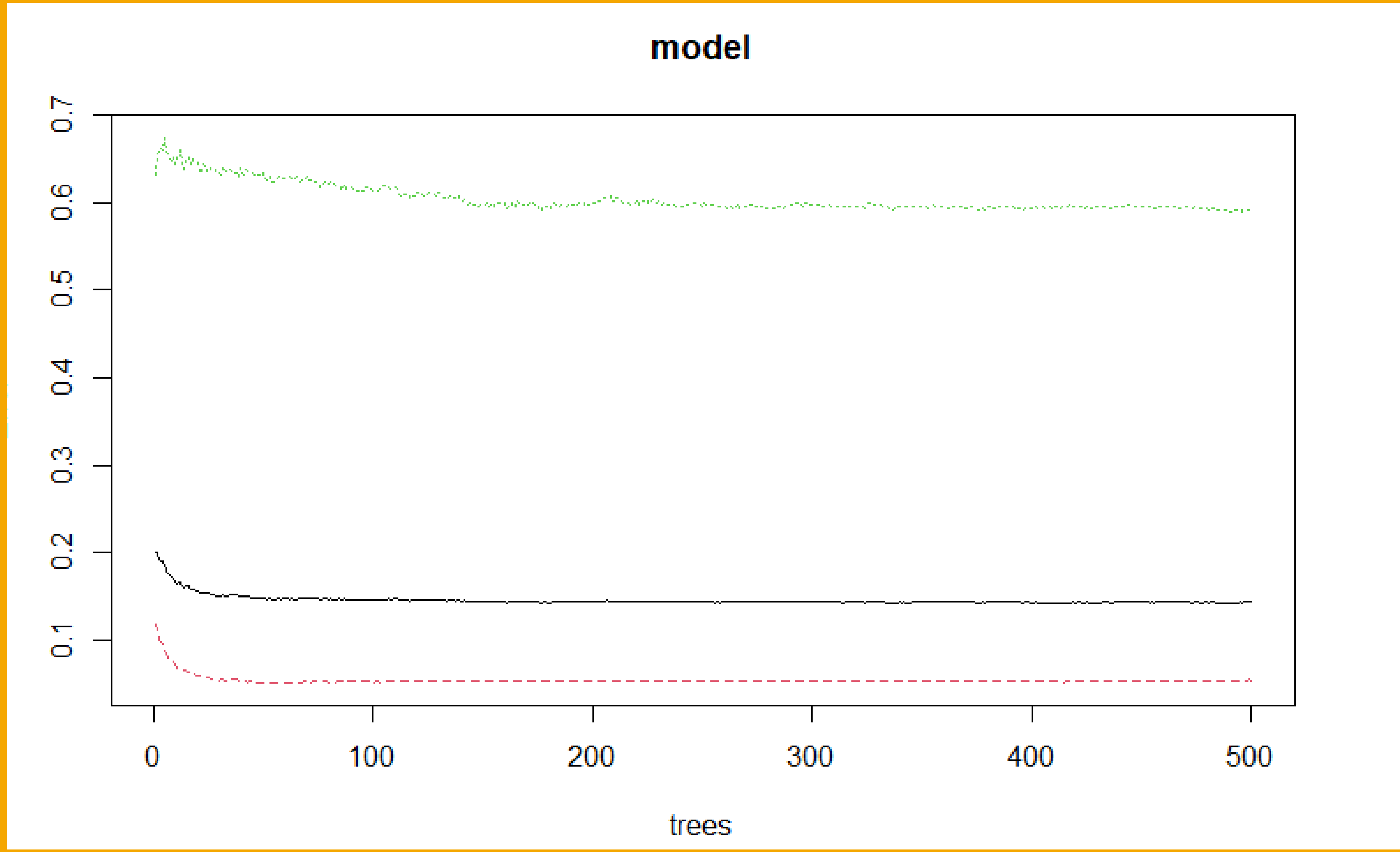
Dai risultati si nota un leggero overfitting.

RANDOM FOREST

```r
####TUNING PARAMETER MTRY
metric <- "Accuracy"
customRF <- list(type = "Classification", library = "randomForest", loop = NULL)
control <- trainControl(method="repeatedcv", number=10, repeats=3)
tunegrid <- expand.grid(mtry=c(1:11))
set.seed(157)
tune <- train(target~., data=train_df, method = "rf",
          metric=metric, tuneGrid=tunegrid, trControl=control)
```

```
> head(model$err.rate)
           OOB        0.0       1.0
[1,] 0.1993101 0.11834320 0.6310680
[2,] 0.2010806 0.11057828 0.6516854
[3,] 0.1915418 0.09838673 0.6616972
[4,] 0.1898265 0.09510320 0.6603015
[5,] 0.1864567 0.08765619 0.6750936
[6,] 0.1808012 0.08526851 0.6582049
```





RANDOM FOREST

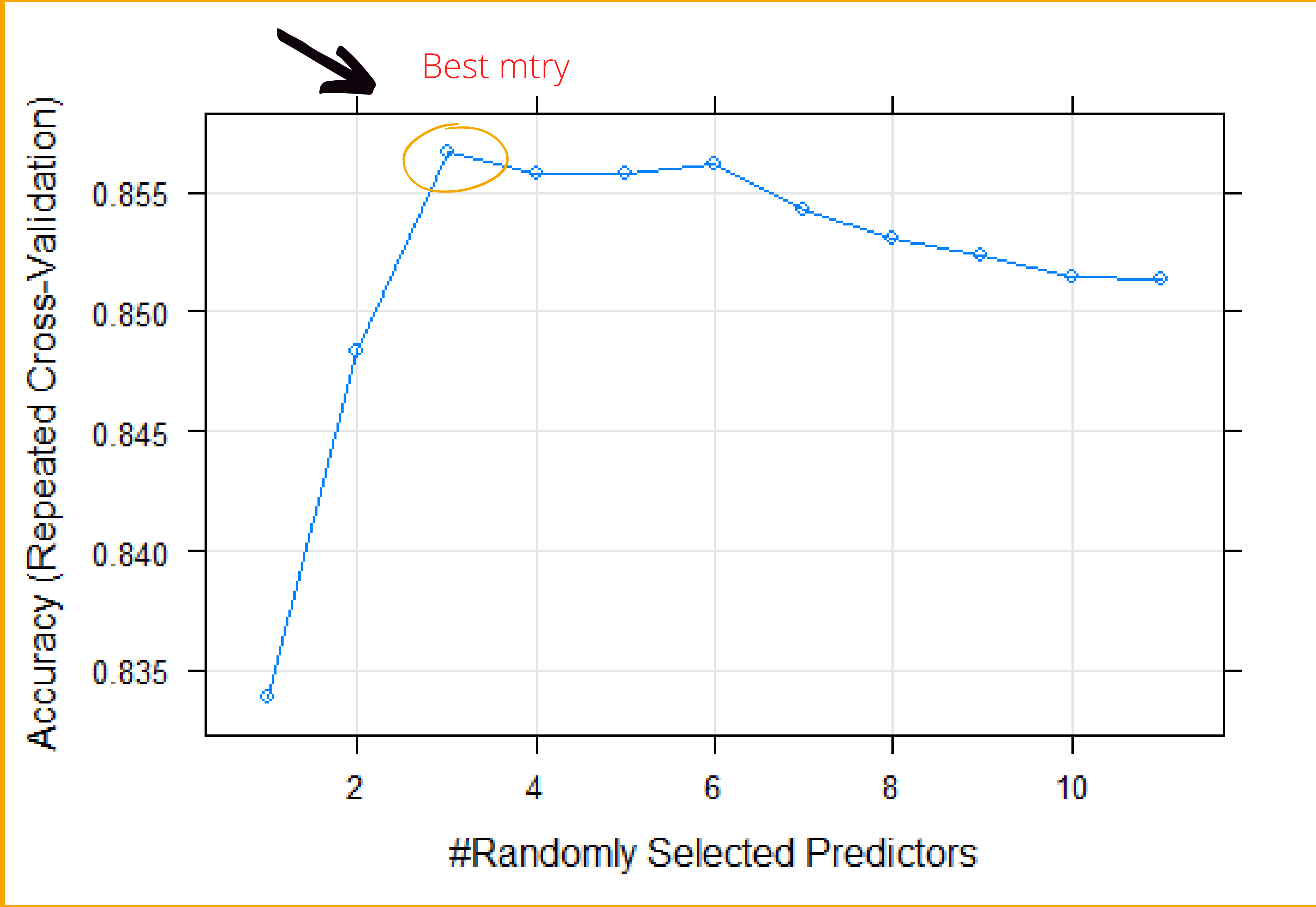
```
Random Forest

7074 samples
  11 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 6366, 6366, 6367, 6366, 6367, 6367, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   1    0.8338991  0.0000000
   2    0.8483665  0.2665144
   3    0.8566582  0.4025714
   4    0.8557632  0.4041924
   5    0.8557162  0.4020299
   6    0.8561888  0.3983245
   7    0.8542561  0.3858359
   8    0.8530780  0.3815330
   9    0.8523708  0.3766517
  10    0.8514286  0.3709386
  11    0.8513340  0.3722354

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 3.
```

ERROR OOB: 0.1450

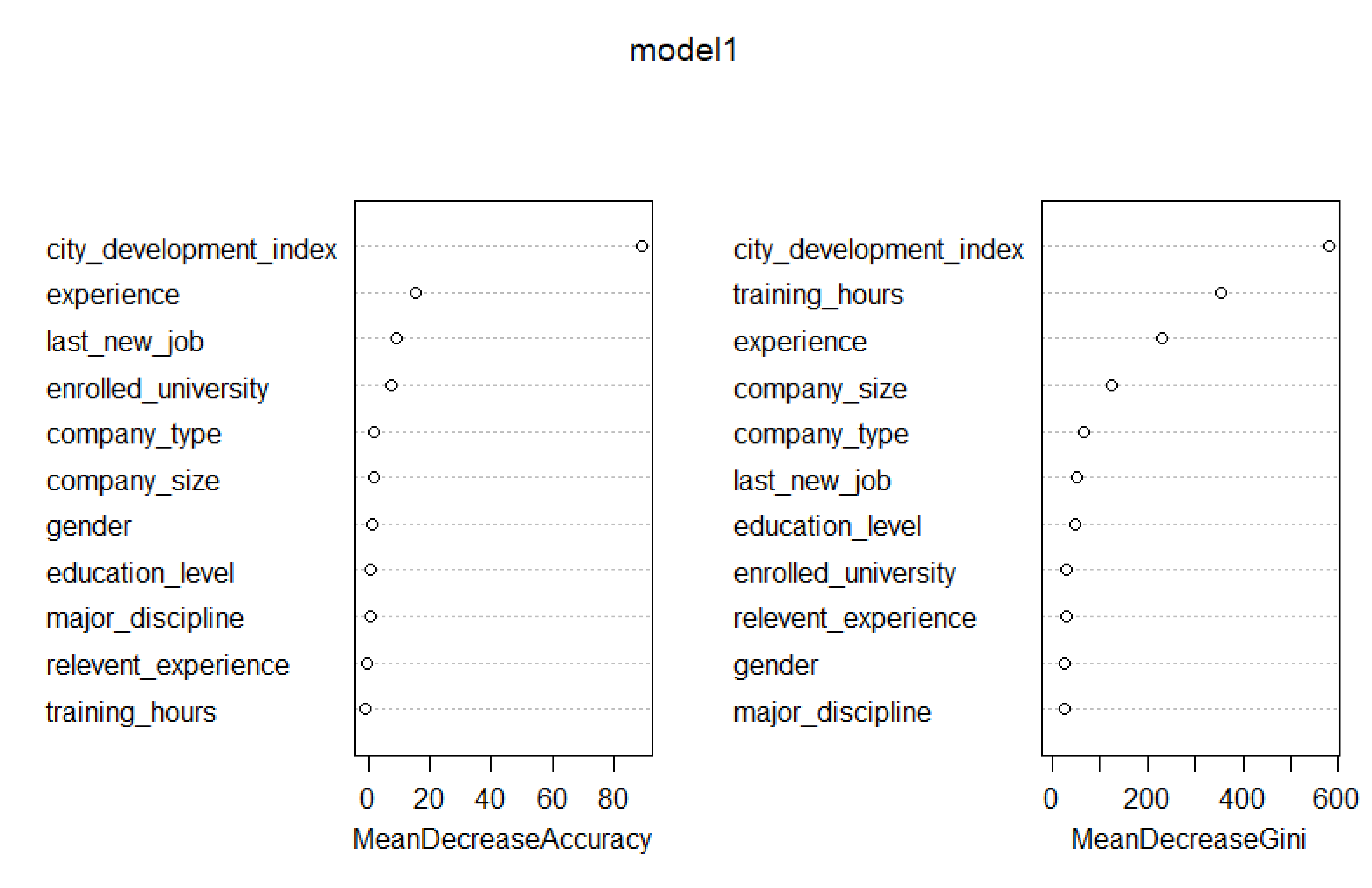
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1382  176
         1   92  117

               Accuracy : 0.8483
                 95% CI : (0.8307, 0.8647)
    No Information Rate : 0.8342
    P-Value [Acc > NIR] : 0.05728

                  Kappa : 0.3806

 Mcnemar's Test P-Value : 3.977e-07

            Sensitivity : 0.39932
            Specificity : 0.93758
         Pos Pred Value : 0.55981
         Neg Pred Value : 0.88703
             Prevalence : 0.16582
         Detection Rate : 0.06621
   Detection Prevalence : 0.11828
      Balanced Accuracy : 0.66845

       'Positive' Class : 1
```

TEST ERROR RATE: 0.152

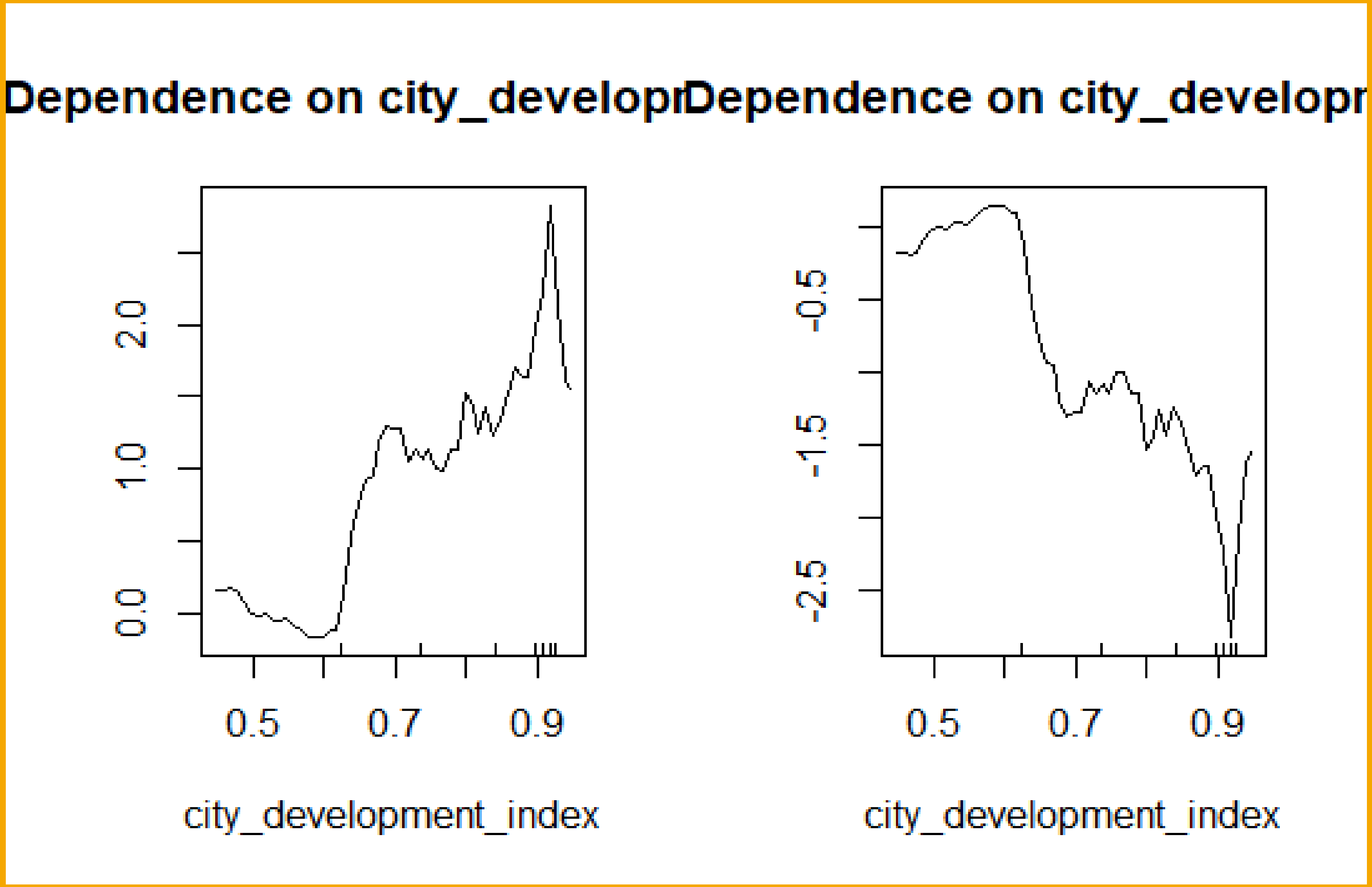
Optimal mtry



VARIABLE IMPORTANCE PLOT



PARTIAL DEPENDENCE PLOT



```
> cbind(err.logit, err.tree, err.random)
     err.logit  err.tree err.random
[1,] 0.1465761 0.1437465  0.1522354
```

CONFRONTO DEGLI ERRORI DI PREVISIONE IN TEST DEI TRE MODELLI

CONCLUSIONI

- **Dataset sbilanciato**
- **Maggior parte degli individui provenienti da città con tasso di sviluppo elevato (e.g. caratteristiche simili)**
- **Presenza di eccessivi NA (circa 10.000)**
- Modelli capaci di predire bene la categoria 0, ma non la 1.
- City development index è una variabile contenente a sua volta altre variabili e, per questo, molto esplicativa.
- L'inserimento di variabili soggettive e/o psicologiche (e.g. reddito, numero di figli ecc.) potrebbe migliorare la previsione.

Thanks for the attention!