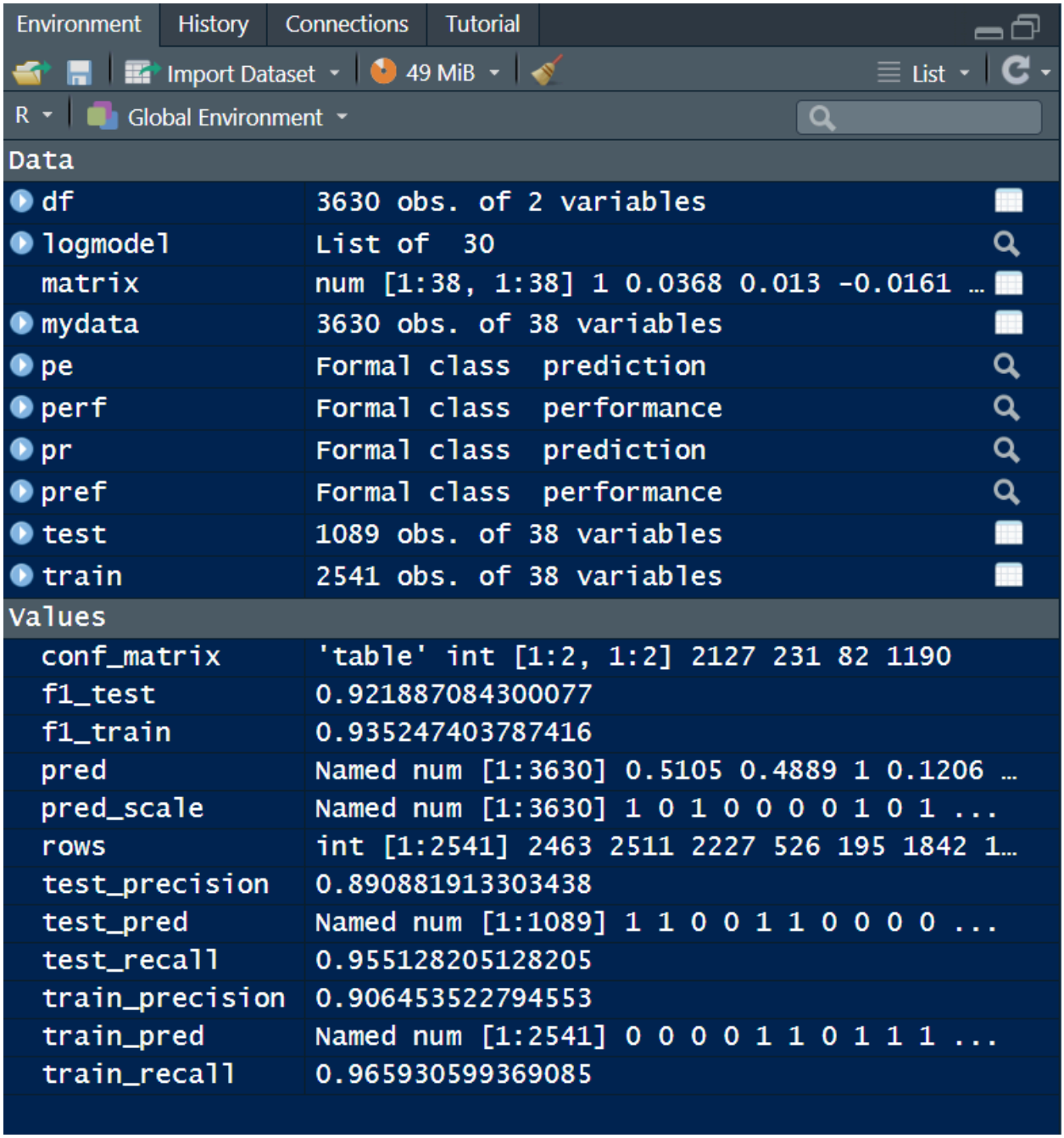
MACHINE LEARNING ANALYSIS USING RSTUDIO

Dataset: student.csv
Code: Rstudion Workspace
Output below:


Environment
History
Connections
Tutorial
Import Dataset
49 MiB
List
R
Global Environment
Data
df
3630 obs. of 2 variables
logmodel
List of 30
matrix
num [1:38, 1:38] 1 0.0368 0.013 -0.0161 …
mydata
3630 obs. of 38 variables
pe
Formal class prediction
perf
Formal class performance
pr
Formal class prediction
pref
Formal class performance
test
1089 obs. of 38 variables
train
2541 obs. of 38 variables
Values
conf_matrix
'table' int [1:2, 1:2] 2127 231 82 1190
f1_test
0.921887084300077
f1_train
0.935247403787416
pred
Named num [1:3630] 0.5105 0.4889 1 0.1206 …
pred_scale
Named num [1:3630] 1 0 1 0 0 0 0 1 0 1 ...
rows
int [1:2541] 2463 2511 2227 526 195 1842 1…
test_precision
0.890881913303438
test_pred
Named num [1:1089] 1 1 0 0 1 1 0 0 0 0 ...
test_recall
0.955128205128205
train_precision
0.906453522794553
train_pred
Named num [1:2541] 0 0 0 0 1 1 0 1 1 1 ...
train_recall
0.965930599369085

```
+ INFERENCE ON TRAIN DATA
+ we observe that f1 score is 93% ,precision is 90% and recall is 96%  for the train dat
a
+ f1-score: F1 score above 90% indicates a good balance between precision and recall.
+ Precision: Precision above 90% means that the vast majority of the model's positive pr
edictions are correct.
+ The model is highly accurate and effecient at identifying positive instances while min
imizing false positives and false negatives.
+ Recall: A recall of 95% indicates that the model is correctly capturing 90%
+ of the actual positive instances present in the data
+
+ INFERENCE ON TEST DATA
+  we observe that precision in test data is 89% ,recall is 95% and f1 score is 92%.
+  f1-score:the metrics have decreased by 1% from the train data ,but we can conclude th
at A high F1 score (92% in this case) indicates that the model has achieved a good balan
ce between precision and recall.
+  It suggests that the model is accurate and effective at identifying positive instance
s while minimizing false positives and false negatives.
+  Recall:recall of 95% indicates that the model is correctly capturing 89% of the actua
l positive instances present in the data.
+  Precision: A precision of 89% indicates that out of all instances predicted as positi
ve by the model, 95% of them are actually positive. This suggests that the model is maki
ng positive predictions with high accuracy.##
```

```
+ CONCLUSIONS
+
+ Exploratory Data Analysis gives the variables in distict forms and the type of data is
analysed and
+ graphical representation is also depicted here. Categorical variables are described as
factors.
+
+ Significant variables:
+ Course                                        0.001860 **
+ Nacionality                                   0.000963 ***
+ Mother_s_qualification                        0.047555 *
+ Mother_s_occupation                           0.019027 *
+ Displaced                                    0.048245 *
+ Debtor1                                      0.000102 ***
+ Tuition_fees_up_to_date1                     5.98e-15 ***
+ Scholarship_holder1                          0.000367 ***
+ Age_at_enrollment                            0.024920 *
+ International                                 0.000428 ***
+ Curricular_units_1st_sem__approved_           2.36e-09 ***
+ Curricular_units_2nd_sem__enrolled_           1.05e-06 ***
+ Curricular_units_2nd_sem__approved_            < 2e-16 ***
+ Curricular_units_2nd_sem_grade                0.007086 **
+ Unemployment_rate                             0.014948 *
+
+ Other variables are not significant so they can be removed
```

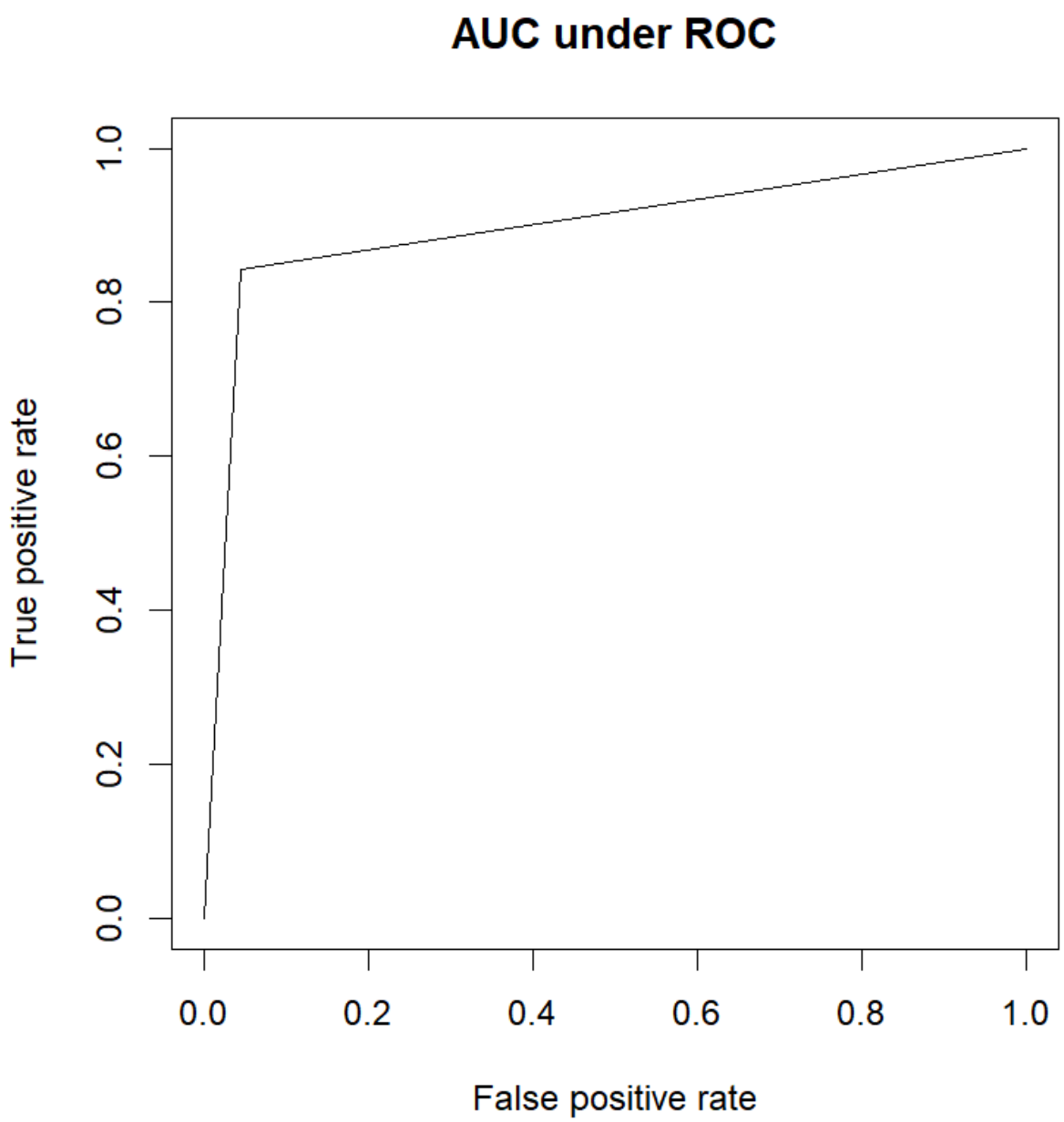
```
> library(Metrics)
> accuracy(mydata$Target,pred_scale)
[1] 0.9137741
> library(ROCR)
> pr<-prediction(pred_scale,mydata$Target)
> #creating a prediction class
> pref<- performance(pr,measure = "tpr",x.measure = "fpr")
> auc(mydata$Target,pred_scale)
[1] 0.9001588
> plot(pref,main="AUC under ROC")
> """
```

```
> table(test$Target, test_pred)
   test_pred
      0   1
  0 596  28
  1  73 392
> #precision = true ones/predicted ones
> test_precision = 596/(596+73)
> test_precision
[1] 0.8908819
> # recall = true ones/actual ones
> test_recall = 596/(596+28)
> test_recall
[1] 0.9551282
> #f1-score = 2pr/(p+r)
> f1_test = 2*test_precision*test_recall/(test_precision+test_recall)
> f1_test
[1] 0.9218871
> library(ROCR)
> pe=prediction(test_pred,test$Target)
> perf=performance(pe,measure = "tpr", x.measure = "fpr")
> plot(perf,main="AUC under ROC")
> #Additional analysis on entire data
> pred<-predict(logmodel,mydata,type = 'response')
> pred_scale<-ifelse(pred >0.5,1,0)
> df<-data.frame(pred,pred_scale)
> View(df)
> conf_matrix<-table(Actual=mydata$Target,Predicted= pred_scale > 0.5)
> print(conf_matrix)
      Predicted
Actual FALSE TRUE
     0  2127   82
     1   231 1190
```

Filter

	pred	pred_scale
1	0.510540886	1
2	0.488874510	0
3	0.999984194	1
4	0.120552234	0
5	0.054094779	0
6	0.314985574	0
7	0.008063308	0
8	0.999975503	1
9	0.016700675	0
10	0.992204441	1
11	0.123352543	0
12	0.024635459	0
13	0.999658415	1
14	0.041661547	0
15	0.056888517	0
16	0.997364163	1
17	0.026346102	0
18	0.189600274	0
19	0.164118809	0
20	0.078434321	0
21	0.016710457	0
22	0.057870674	0

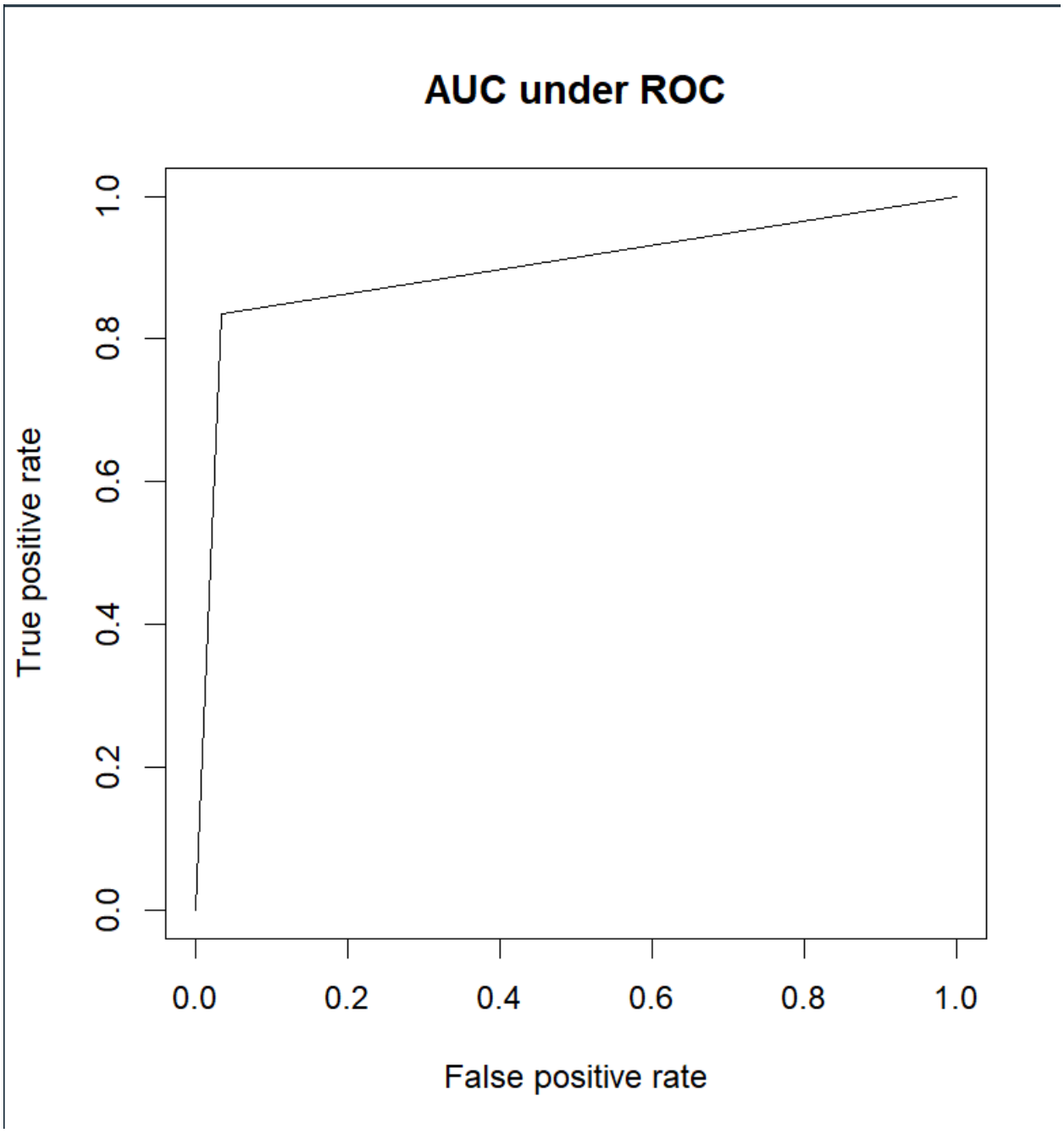
Showing 1 to 22 of 3,630 entries, 2 total columns


AUC under ROC
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

```
> #precision, recall, f1-score
> table(test$Target, test_pred)
   test_pred
      0   1
  0 596  28
  1  73 392
> #precision = true ones/predicted ones
> test_precision = 596/(596+73)
> test_precision
[1] 0.8908819
> # recall = true ones/actual ones
> test_recall = 596/(596+28)
> test_recall
[1] 0.9551282
> #f1-score = 2pr/(p+r)
> f1_test = 2*test_precision*test_recall/(test_precision+test_recall)
> f1_test
[1] 0.9218871
> library(ROCR)
> pe=prediction(test_pred,test$Target)
> perf=performance(pe,measure = "tpr", x.measure = "fpr")
> plot(perf,main="AUC under ROC")
> #Additional analysis on entire data
> pred<-predict(logmodel,mydata,type = 'response')
> pred_scale<-ifelse(pred >0.5,1,0)
> df<-data.frame(pred,pred_scale)
> View(df)
```

```
> install.packages("ROCR")
WARNING: Rtools is required to build R packages but is not currently installed. Please d
ownload and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Warning in install.packages :
  package 'ROCR' is in use and will not be installed
> library(ROCR)
> pe=prediction(train_pred,train$Target)
> perf=performance(pe,measure = "tpr", x.measure = "fpr")
> plot(perf,main="AUC under ROC")
> #test data
> test_pred = predict(logmodel, test, type = "response")
> test_pred = ifelse(test_pred >=0.5,1,0)
> table(test_pred)
test_pred
  0   1
669 420
> head(test$Target, 10)
 [1] 1 1 0 0 1 1 0 1 0 0
Levels: 0 1
> test_pred[1:10]
2542 2543 2544 2545 2546 2547 2548 2549 2550 2551
   1    1    0    0    1    1    0    0    0    0
> #precision, recall, f1-score
> table(test$Target, test_pred)
   test_pred
      0   1
  0 596  28
  1  73 392
```


AUC under ROC
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

```
> train_recall
[1] 0.9659306
> #f1-score = 2pr/(p+r)
> f1_train = 2*train_precision*train_recall/(train_precision+train_recall)
> f1_train
[1] 0.9352474
> install.packages("ROCR")
Error in install.packages : Updating loaded packages
> install.packages("ROCR")
WARNING: Rtools is required to build R packages but is not currently installed. Please d
ownload and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Warning in install.packages :
  package 'ROCR' is in use and will not be installed
> library(ROCR)
> pe=prediction(train_pred,train$Target)
> perf=performance(pe,measure = "tpr", x.measure = "fpr")
> plot(perf,main="AUC under ROC")
```

```
Curricular_units_2nd_sem_grade                 15.146875  1        3.891899
Curricular_units_2nd_sem__without_evaluations_  1.209564  1        1.099802
Unemployment_rate                               1.387676  1        1.177997
Inflation_rate                                  1.094271  1        1.046074
GDP                                             1.256348  1        1.120869
> #train data
> train_pred = predict(logmodel, train, type = "response")
> train_pred = ifelse(train_pred >=0.5,1,0)
> table(train_pred)
train_pred
   0    1
1689  852
> head(train$Target, 10)
 [1] 0 0 0 1 1 1 0 1 0 1
Levels: 0 1
> train_pred[1:10]
2463 2511 2227  526  195 1842 1142 1253 1268 1038
   0    0    0    0    1    1    0    1    1    1
> #precision, recall, f1-score
> table(train$Target, train_pred)
   train_pred
      0    1
  0 1531   54
  1  158  798
> #precision = true ones/predicted ones
> train_precision = 1531/(1531+158)
> train_precision
[1] 0.9064535
> # recall = true ones/actual ones
> train_recall = 1531/(1531+54)
> train_recall
[1] 0.9659306
```

```
Number of Fisher Scoring iterations: 8

> library(car)
> car::vif(logmodel)
                                        GVIF Df GVIF^(1/(2*Df))
X                                   1.056890  1        1.028052
Marital_status                      2.458166  5        1.094110
Application_mode                    1.869895  1        1.367441
Application_order                   1.250562  1        1.118285
Course                             11.297178  1        3.361128
Daytime_evening_attendance          1.418575  1        1.191039
Previous_qualification              1.432761  1        1.196980
Previous_qualification_grade        1.669962  1        1.292270
Nacionality                         3.313247  1        1.820233
Mother_s_qualification              1.613535  1        1.270250
Father_s_qualification              1.552553  1        1.246015
Mother_s_occupation                 3.760490  1        1.939198
Father_s_occupation                 3.753696  1        1.937446
Admission_grade                     1.733380  1        1.316579
Displaced                           1.334347  1        1.155139
Educational_special_needs           1.031110  1        1.015436
Debtor                              1.197039  1        1.094093
Tuition_fees_up_to_date             1.190829  1        1.091251
Gender                              1.146546  1        1.070769
Scholarship_holder                  1.093837  1        1.045866
Age_at_enrollment                   2.833084  1        1.683177
International                       3.439883  1        1.854692
Curricular_units_1st_sem__credited_ 15.539656  1        3.942037
Curricular_units_1st_sem__enrolled_ 50.602411  1        7.113537
Curricular_units_1st_sem__evaluations_ 7.394280  1        2.719243
Curricular_units_1st_sem__approved  12.473122  1        3.531731
```

```
Father_s_occupation                             0.124839
Admission_grade                                 0.968908
Displaced                                       0.048245 *
Educational_special_needs                       0.309336
Debtor1                                         0.000102 ***
Tuition_fees_up_to_date1                        5.98e-15 ***
Gender1                                         0.063489 .
Scholarship_holder1                             0.000367 ***
Age_at_enrollment                               0.024920 *
International                                   0.000428 ***
Curricular_units_1st_sem__credited_             0.089426 .
Curricular_units_1st_sem__enrolled_             0.369898
Curricular_units_1st_sem__evaluations_          0.682282
Curricular_units_1st_sem__approved_             2.36e-09 ***
Curricular_units_1st_sem_grade                  0.458894
Curricular_units_1st_sem__without_evaluations_ 0.989561
Curricular_units_2nd_sem__credited_             0.337160
Curricular_units_2nd_sem__enrolled_             1.05e-06 ***
Curricular_units_2nd_sem__evaluations_          0.462268
Curricular_units_2nd_sem__approved_              < 2e-16 ***
Curricular_units_2nd_sem_grade                  0.007086 **
Curricular_units_2nd_sem__without_evaluations_ 0.465254
Unemployment_rate                               0.014948 *
Inflation_rate                                  0.950473
GDP                                             0.790142
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3365.2  on 2540  degrees of freedom
Residual deviance: 1126.7  on 2499  degrees of freedom
AIC: 1210.7
```

```
(Intercept)                                 0.836732
X                                           0.437080
Marital_status2                             0.121968
Marital_status3                             0.665489
Marital_status4                             0.061540 .
Marital_status5                             0.694231
Marital_status6                             0.907518
Application_mode                            0.390161
Application_order                           0.559834
Course                                      0.001860 **
Daytime_evening_attendance1                 0.904759
Previous_qualification                      0.108765
Previous_qualification_grade                0.839724
Nacionality                                 0.000963 ***
Mother_s_qualification                      0.047555 *
Father_s_qualification                      0.063781 .
Mother_s_occupation                         0.019027 *
Father_s_occupation                         0.124839
Admission_grade                             0.968908
Displaced                                   0.048245 *
Educational_special_needs                   0.309336
Debtor1                                     0.000102 ***
Tuition_fees_up_to_date1                    5.98e-15 ***
Gender1                                     0.063489 .
Scholarship_holder1                         0.000367 ***
Age_at_enrollment                           0.024920 *
International                               0.000428 ***
Curricular_units_1st_sem__credited_         0.089426 .
Curricular_units_1st_sem__enrolled          0.369898
```

```
> summary(logmodel)

Call:
glm(formula = train$Target ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2241  -0.3552  -0.1922   0.0293   2.9197

Coefficients:
                                          Estimate Std. Error z value
(Intercept)                             -2.520e-01  1.223e+00  -0.206
X                                        6.910e-05  8.891e-05   0.777
Marital_status2                         -6.423e-01  4.153e-01  -1.547
Marital_status3                         -2.667e+00  6.169e+00  -0.432
Marital_status4                         -1.342e+00  7.175e-01  -1.870
Marital_status5                         -4.048e-01  1.030e+00  -0.393
Marital_status6                         -3.530e-01  3.038e+00  -0.116
Application_mode                         5.267e-03  6.129e-03   0.859
Application_order                        3.828e-02  6.565e-02   0.583
Course                                   2.582e-04  8.297e-05   3.112
Daytime_evening_attendance1              3.771e-02  3.152e-01   0.120
Previous_qualification                  -1.460e-02  9.105e-03  -1.604
Previous_qualification_grade             1.495e-03  7.390e-03   0.202
Nacionality                              5.742e-02  1.739e-02   3.301
Mother_s_qualification                   1.258e-02  6.348e-03   1.981
Father_s_qualification                  -1.168e-02  6.301e-03  -1.854
Mother_s_occupation                     -1.928e-02  8.224e-03  -2.345
Father_s_occupation                      1.226e-02  7.988e-03   1.535
Admission_grade                         -2.646e-04  6.789e-03  -0.039
Displaced                                3.652e-01  1.849e-01   1.975
Educational_special_needs                5.930e-01  5.833e-01   1.017
Debtor1                                  1.148e+00  2.954e-01   3.886
```