Anotações:

Verificar se os dados estão balanceados, pois se trata de um problema de classificação

Os dados estão todos codificados, então não há porque fazer uma análise exploratória

Existem muitas variáveis, considerar usar redução de dimensionalidade

Testar os algoritmos com validação cruzada e construir boxplots com os resultados

Testar se feature selection melhora o modelo

1 = insatisfeito, 0 = satisfeito

In [28]:
```
import pandas as pd
import numpy as np
from sklearn.decomposition import PCA
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import cross_val_score, KFold
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
```

In [71]:
```
df = pd.read_csv('train.csv')
```

In [4]:
```
df.head()
```

Out[4]:

	ID	var3	var15	imp_ent_var16_ult1	imp_op_var39_comer_ult1	imp_op_var39_comer_ult3	imp_op_var40_comer_ult1	imp_op_var40_com
0	1	2	23	0.0	0.0	0.0	0.0	
1	3	2	34	0.0	0.0	0.0	0.0	
2	4	2	23	0.0	0.0	0.0	0.0	
3	8	2	37	0.0	195.0	195.0	0.0	
4	10	2	39	0.0	0.0	0.0	0.0	

5 rows × 371 columns

In [5]:
```
df.shape
```

Out[5]:
```
(76020, 371)
```

In [53]:
```
df.isnull().sum()
```

Out[53]:
```
var3                               0
var15                              0
imp_ent_var16_ult1                 0
imp_op_var39_comer_ult1            0
imp_op_var39_comer_ult3            0
imp_op_var40_comer_ult1            0
imp_op_var40_comer_ult3            0
imp_op_var40_efect_ult1            0
imp_op_var40_efect_ult3            0
imp_op_var40_ult1                  0
imp_op_var41_comer_ult1            0
imp_op_var41_comer_ult3            0
imp_op_var41_efect_ult1            0
imp_op_var41_efect_ult3            0
imp_op_var41_ult1                  0
imp_op_var39_efect_ult1            0
imp_op_var39_efect_ult3            0
imp_op_var39_ult1                  0
imp_sal_var16_ult1                 0
ind_var1_0                         0
ind_var1                           0
ind_var5_0                         0
ind_var5                           0
ind_var6_0                         0
ind_var6                           0
ind_var8_0                         0
ind_var8                           0
ind_var12_0                        0
ind_var12                          0
ind_var13_0                        0
ind_var13_corto_0                  0
ind_var13_corto                    0
ind_var13_largo_0                  0
ind_var13_largo                    0
ind_var13_medio_0                  0
ind_var13_medio                    0
ind_var13                          0
ind_var14_0                        0
ind_var14                          0
ind_var17_0                        0
ind_var17                          0
ind_var18_0                        0
ind_var18                          0
ind_var19                          0
ind_var20_0                        0
ind_var20                          0
ind_var24_0                        0
ind_var24                          0
ind_var25_cte                      0
ind_var26_0                        0
ind_var26_cte                      0
ind_var26                          0
ind_var25_0                        0
ind_var25                          0
ind_var29_0                        0
ind_var29                          0
ind_var30_0                        0
ind_var30                          0
ind_var31_0                        0
ind_var31                          0
ind_var32_cte                      0
ind_var32_0                        0
ind_var32                          0
ind_var33_0                        0
ind_var33                          0
ind_var34_0                        0
ind_var34                          0
ind_var37_cte                      0
ind_var37_0                        0
ind_var37                          0
ind_var39_0                        0
ind_var40_0                        0
ind_var40                          0
ind_var41_0                        0
ind_var39                          0
ind_var44_0                        0
ind_var44                          0
num_var1_0                         0
num_var1                           0
num_var4                           0
num_var5_0                         0
num_var5                           0
num_var6_0                         0
num_var6                           0
num_var8_0                         0
num_var8                           0
num_var12_0                        0
num_var12                          0
num_var13_0                        0
num_var13_corto_0                  0
num_var13_corto                    0
num_var13_largo_0                  0
num_var13_largo                    0
num_var13_medio_0                  0
num_var13_medio                    0
num_var13                          0
num_var14_0                        0
num_var14                          0
num_var17_0                        0
num_var17                          0
num_var18_0                        0
num_var18                          0
num_var20_0                        0
num_var20                          0
num_var24_0                        0
num_var24                          0
num_var26_0                        0
num_var26                          0
num_var25_0                        0
num_var25                          0
num_op_var40_hace2                 0
num_op_var40_hace3                 0
num_op_var40_ult1                  0
num_op_var40_ult3                  0
num_op_var41_hace2                 0
num_op_var41_hace3                 0
num_op_var41_ult1                  0
num_op_var41_ult3                  0
num_op_var39_hace2                 0
num_op_var39_hace3                 0
num_op_var39_ult1                  0
num_op_var39_ult3                  0
num_var29_0                        0
num_var29                          0
num_var30_0                        0
num_var30                          0
num_var31_0                        0
num_var31                          0
num_var32_0                        0
num_var32                          0
num_var33_0                        0
num_var33                          0
num_var34_0                        0
num_var34                          0
num_var35                          0
num_var37_med_ult2                 0
num_var37_0                        0
num_var37                          0
num_var39_0                        0
num_var40_0                        0
num_var40                          0
num_var41_0                        0
num_var39                          0
num_var42_0                        0
num_var42                          0
num_var44_0                        0
num_var44                          0
saldo_var1                         0
saldo_var5                         0
saldo_var6                         0
saldo_var8                         0
saldo_var12                        0
saldo_var13_corto                  0
saldo_var13_largo                  0
saldo_var13_medio                  0
saldo_var13                        0
saldo_var14                        0
saldo_var17                        0
saldo_var18                        0
saldo_var20                        0
saldo_var24                        0
saldo_var26                        0
saldo_var25                        0
saldo_var29                        0
saldo_var30                        0
saldo_var31                        0
saldo_var32                        0
saldo_var33                        0
saldo_var34                        0
saldo_var37                        0
saldo_var40                        0
saldo_var42                        0
saldo_var44                        0
var36                              0
delta_imp_amort_var18_1y3          0
delta_imp_amort_var34_1y3          0
delta_imp_aport_var13_1y3          0
delta_imp_aport_var17_1y3          0
delta_imp_aport_var33_1y3          0
delta_imp_compra_var44_1y3         0
delta_imp_reemb_var13_1y3          0
delta_imp_reemb_var17_1y3          0
delta_imp_reemb_var33_1y3          0
delta_imp_trasp_var17_in_1y3       0
delta_imp_trasp_var17_out_1y3      0
delta_imp_trasp_var33_in_1y3       0
delta_imp_trasp_var33_out_1y3      0
delta_imp_venta_var44_1y3          0
delta_num_aport_var13_1y3          0
delta_num_aport_var17_1y3          0
delta_num_aport_var33_1y3          0
delta_num_compra_var44_1y3         0
delta_num_reemb_var13_1y3          0
delta_num_reemb_var17_1y3          0
delta_num_reemb_var33_1y3          0
delta_num_trasp_var17_in_1y3       0
delta_num_trasp_var17_out_1y3      0
delta_num_trasp_var33_in_1y3       0
delta_num_trasp_var33_out_1y3      0
delta_num_venta_var44_1y3          0
imp_amort_var18_ult1               0
imp_amort_var34_ult1               0
imp_aport_var13_hace3              0
imp_aport_var13_ult1               0
imp_aport_var17_hace3              0
imp_aport_var17_ult1               0
imp_aport_var33_hace3              0
imp_aport_var33_ult1               0
imp_var7_emit_ult1                 0
imp_var7_recib_ult1                0
imp_compra_var44_hace3             0
imp_compra_var44_ult1              0
imp_reemb_var13_ult1               0
imp_reemb_var17_hace3              0
imp_reemb_var17_ult1               0
imp_reemb_var33_ult1               0
imp_var43_emit_ult1                0
imp_trans_var37_ult1               0
imp_trasp_var17_in_hace3           0
imp_trasp_var17_in_ult1            0
imp_trasp_var17_out_ult1           0
imp_trasp_var33_in_hace3           0
imp_trasp_var33_in_ult1            0
imp_trasp_var33_out_ult1           0
imp_venta_var44_hace3              0
imp_venta_var44_ult1               0
ind_var7_emit_ult1                 0
ind_var7_recib_ult1                0
ind_var10_ult1                     0
ind_var10cte_ult1                  0
ind_var9_cte_ult1                  0
ind_var9_ult1                      0
ind_var43_emit_ult1                0
ind_var43_recib_ult1               0
var21                              0
num_aport_var13_hace3              0
num_aport_var13_ult1               0
num_aport_var17_hace3              0
num_aport_var17_ult1               0
num_aport_var33_hace3              0
num_aport_var33_ult1               0
num_var7_emit_ult1                 0
num_var7_recib_ult1                0
num_compra_var44_hace3             0
num_compra_var44_ult1              0
num_ent_var16_ult1                 0
num_var22_hace2                    0
num_var22_hace3                    0
num_var22_ult1                     0
num_var22_ult3                     0
num_med_var22_ult3                 0
num_med_var45_ult3                 0
num_meses_var5_ult3                0
num_meses_var8_ult3                0
num_meses_var12_ult3               0
num_meses_var13_corto_ult3         0
num_meses_var13_largo_ult3         0
num_meses_var13_medio_ult3         0
num_meses_var17_ult3               0
num_meses_var29_ult3               0
num_meses_var33_ult3               0
num_meses_var39_vig_ult3           0
num_meses_var44_ult3               0
num_op_var39_comer_ult1            0
num_op_var39_comer_ult3            0
num_op_var40_comer_ult1            0
num_op_var40_comer_ult3            0
num_op_var40_efect_ult1            0
num_op_var40_efect_ult3            0
num_op_var41_comer_ult1            0
num_op_var41_comer_ult3            0
num_op_var41_efect_ult1            0
num_op_var41_efect_ult3            0
num_op_var39_efect_ult1            0
num_op_var39_efect_ult3            0
num_reemb_var13_ult1               0
num_reemb_var17_hace3              0
num_reemb_var17_ult1               0
num_reemb_var33_ult1               0
num_sal_var16_ult1                 0
num_var43_emit_ult1                0
num_var43_recib_ult1               0
num_trasp_var11_ult1               0
num_trasp_var17_in_hace3           0
num_trasp_var17_in_ult1            0
num_trasp_var17_out_ult1           0
num_trasp_var33_in_hace3           0
num_trasp_var33_in_ult1            0
num_trasp_var33_out_ult1           0
num_venta_var44_hace3              0
num_venta_var44_ult1               0
num_var45_hace2                    0
num_var45_hace3                    0
num_var45_ult1                     0
num_var45_ult3                     0
saldo_medio_var5_hace2             0
saldo_medio_var5_hace3             0
saldo_medio_var5_ult1              0
saldo_medio_var5_ult3              0
saldo_medio_var8_hace2             0
saldo_medio_var8_hace3             0
saldo_medio_var8_ult1              0
saldo_medio_var8_ult3              0
saldo_medio_var12_hace2            0
saldo_medio_var12_hace3            0
saldo_medio_var12_ult1             0
saldo_medio_var12_ult3             0
saldo_medio_var13_corto_hace2      0
saldo_medio_var13_corto_hace3      0
saldo_medio_var13_corto_ult1       0
saldo_medio_var13_corto_ult3       0
saldo_medio_var13_largo_hace2      0
saldo_medio_var13_largo_hace3      0
saldo_medio_var13_largo_ult1       0
saldo_medio_var13_largo_ult3       0
saldo_medio_var13_medio_hace2      0
saldo_medio_var13_medio_ult1       0
saldo_medio_var13_medio_ult3       0
saldo_medio_var17_hace2            0
saldo_medio_var17_hace3            0
saldo_medio_var17_ult1             0
saldo_medio_var17_ult3             0
saldo_medio_var29_hace2            0
saldo_medio_var29_hace3            0
saldo_medio_var29_ult1             0
saldo_medio_var29_ult3             0
saldo_medio_var33_hace2            0
saldo_medio_var33_hace3            0
saldo_medio_var33_ult1             0
saldo_medio_var33_ult3             0
saldo_medio_var44_hace2            0
saldo_medio_var44_hace3            0
saldo_medio_var44_ult1             0
saldo_medio_var44_ult3             0
var38                              0
TARGET                             0
dtype: int64
```

A coluna ID é inútil

In [72]:
```
df = df.iloc[:,1:]
```

In [67]:
```
df.dtypes
```

Out[67]:
```
imp_ent_var16_ult1               float64
imp_op_var39_comer_ult1          float64
imp_op_var39_comer_ult3          float64
imp_op_var40_comer_ult1          float64
imp_op_var40_comer_ult3          float64
imp_op_var40_efect_ult1          float64
imp_op_var40_efect_ult3          float64
imp_op_var40_ult1                float64
imp_op_var41_comer_ult1          float64
imp_op_var41_comer_ult3          float64
imp_op_var41_efect_ult1          float64
imp_op_var41_efect_ult3          float64
imp_op_var41_ult1                float64
imp_op_var39_efect_ult1          float64
imp_op_var39_efect_ult3          float64
imp_op_var39_ult1                float64
imp_sal_var16_ult1               float64
ind_var1_0                         int64
ind_var1                           int64
ind_var5_0                         int64
ind_var5                           int64
ind_var6_0                         int64
ind_var6                           int64
ind_var8_0                         int64
ind_var8                           int64
ind_var12_0                        int64
ind_var12                          int64
ind_var13_0                        int64
ind_var13_corto_0                  int64
ind_var13_corto                    int64
ind_var13_largo_0                  int64
ind_var13_largo                    int64
ind_var13_medio_0                  int64
ind_var13_medio                    int64
ind_var13                          int64
ind_var14_0                        int64
ind_var14                          int64
ind_var17_0                        int64
ind_var17                          int64
ind_var18_0                        int64
ind_var18                          int64
ind_var19                          int64
ind_var20_0                        int64
ind_var20                          int64
ind_var24_0                        int64
ind_var24                          int64
ind_var25_cte                      int64
ind_var26_0                        int64
ind_var26_cte                      int64
ind_var26                          int64
ind_var25_0                        int64
ind_var25                          int64
ind_var29_0                        int64
ind_var29                          int64
ind_var30_0                        int64
ind_var30                          int64
ind_var31_0                        int64
ind_var31                          int64
ind_var32_cte                      int64
ind_var32_0                        int64
ind_var32                          int64
ind_var33_0                        int64
ind_var33                          int64
ind_var34_0                        int64
ind_var34                          int64
ind_var37_cte                      int64
ind_var37_0                        int64
ind_var37                          int64
ind_var39_0                        int64
ind_var40_0                        int64
ind_var40                          int64
ind_var41_0                        int64
ind_var39                          int64
ind_var44_0                        int64
ind_var44                          int64
num_var1_0                         int64
num_var1                           int64
num_var4                           int64
num_var5_0                         int64
num_var5                           int64
num_var6_0                         int64
num_var6                           int64
num_var8_0                         int64
num_var8                           int64
num_var12_0                        int64
num_var12                          int64
num_var13_0                        int64
num_var13_corto_0                  int64
num_var13_corto                    int64
num_var13_largo_0                  int64
num_var13_largo                    int64
num_var13_medio_0                  int64
num_var13_medio                    int64
num_var13                          int64
num_var14_0                        int64
num_var14                          int64
num_var17_0                        int64
num_var17                          int64
num_var18_0                        int64
num_var18                          int64
num_var20_0                        int64
num_var20                          int64
num_var24_0                        int64
num_var24                          int64
num_var26_0                        int64
num_var26                          int64
num_var25_0                        int64
num_var25                          int64
num_op_var40_hace2                 int64
num_op_var40_hace3                 int64
num_op_var40_ult1                  int64
num_op_var40_ult3                  int64
num_op_var41_hace2                 int64
num_op_var41_hace3                 int64
num_op_var41_ult1                  int64
num_op_var41_ult3                  int64
num_op_var39_hace2                 int64
num_op_var39_hace3                 int64
num_op_var39_ult1                  int64
num_op_var39_ult3                  int64
num_var29_0                        int64
num_var29                          int64
num_var30_0                        int64
num_var30                          int64
num_var31_0                        int64
num_var31                          int64
num_var32_0                        int64
num_var32                          int64
num_var33_0                        int64
num_var33                          int64
num_var34_0                        int64
num_var34                          int64
num_var35                          int64
num_var37_med_ult2                 int64
num_var37_0                        int64
num_var37                          int64
num_var39_0                        int64
num_var40_0                        int64
num_var40                          int64
num_var41_0                        int64
num_var39                          int64
num_var42_0                        int64
num_var42                          int64
num_var44_0                        int64
num_var44                          int64
saldo_var1                       float64
saldo_var5                       float64
saldo_var6                       float64
saldo_var8                       float64
saldo_var12                      float64
saldo_var13_corto                float64
saldo_var13_largo                float64
saldo_var13_medio                  int64
saldo_var13                      float64
saldo_var14                      float64
saldo_var17                      float64
saldo_var18                        int64
saldo_var20                      float64
saldo_var24                      float64
saldo_var26                      float64
saldo_var25                      float64
saldo_var29                      float64
saldo_var30                      float64
saldo_var31                      float64
saldo_var32                      float64
saldo_var33                      float64
saldo_var34                        int64
saldo_var37                      float64
saldo_var40                      float64
saldo_var42                      float64
saldo_var44                      float64
var36                              int64
delta_imp_amort_var18_1y3          int64
delta_imp_amort_var34_1y3          int64
delta_imp_aport_var13_1y3        float64
delta_imp_aport_var17_1y3        float64
delta_imp_aport_var33_1y3        float64
delta_imp_compra_var44_1y3       float64
delta_imp_reemb_var13_1y3          int64
delta_imp_reemb_var17_1y3          int64
delta_imp_reemb_var33_1y3          int64
delta_imp_trasp_var17_in_1y3       int64
delta_imp_trasp_var17_out_1y3      int64
delta_imp_trasp_var33_in_1y3       int64
delta_imp_trasp_var33_out_1y3      int64
delta_imp_venta_var44_1y3        float64
delta_num_aport_var13_1y3        float64
delta_num_aport_var17_1y3        float64
delta_num_aport_var33_1y3        float64
delta_num_compra_var44_1y3       float64
delta_num_reemb_var13_1y3          int64
delta_num_reemb_var17_1y3          int64
delta_num_reemb_var33_1y3          int64
delta_num_trasp_var17_in_1y3       int64
delta_num_trasp_var17_out_1y3      int64
delta_num_trasp_var33_in_1y3       int64
delta_num_trasp_var33_out_1y3      int64
delta_num_venta_var44_1y3        float64
imp_amort_var18_ult1             float64
imp_amort_var34_ult1             float64
imp_aport_var13_hace3            float64
imp_aport_var13_ult1             float64
imp_aport_var17_hace3            float64
imp_aport_var17_ult1             float64
imp_aport_var33_hace3              int64
imp_aport_var33_ult1               int64
imp_var7_emit_ult1               float64
imp_var7_recib_ult1              float64
imp_compra_var44_hace3           float64
imp_compra_var44_ult1            float64
imp_reemb_var13_ult1             float64
imp_reemb_var17_hace3            float64
imp_reemb_var17_ult1             float64
imp_reemb_var33_ult1               int64
imp_var43_emit_ult1              float64
imp_trans_var37_ult1             float64
imp_trasp_var17_in_hace3         float64
imp_trasp_var17_in_ult1          float64
imp_trasp_var17_out_ult1         float64
imp_trasp_var33_in_hace3         float64
imp_trasp_var33_in_ult1          float64
imp_trasp_var33_out_ult1           int64
imp_venta_var44_hace3            float64
imp_venta_var44_ult1             float64
ind_var7_emit_ult1                 int64
ind_var7_recib_ult1                int64
ind_var10_ult1                     int64
ind_var10cte_ult1                  int64
ind_var9_cte_ult1                  int64
ind_var9_ult1                      int64
ind_var43_emit_ult1                int64
ind_var43_recib_ult1               int64
var21                              int64
num_aport_var13_hace3              int64
num_aport_var13_ult1               int64
num_aport_var17_hace3              int64
num_aport_var17_ult1               int64
num_aport_var33_hace3              int64
num_aport_var33_ult1               int64
num_var7_emit_ult1                 int64
num_var7_recib_ult1                int64
num_compra_var44_hace3             int64
num_compra_var44_ult1              int64
num_ent_var16_ult1                 int64
num_var22_hace2                    int64
num_var22_hace3                    int64
num_var22_ult1                     int64
num_var22_ult3                     int64
num_med_var22_ult3                 int64
num_med_var45_ult3                 int64
num_meses_var5_ult3                int64
num_meses_var8_ult3                int64
num_meses_var12_ult3               int64
num_meses_var13_corto_ult3         int64
num_meses_var13_largo_ult3         int64
num_meses_var13_medio_ult3         int64
num_meses_var17_ult3               int64
num_meses_var29_ult3               int64
num_meses_var33_ult3               int64
num_meses_var39_vig_ult3           int64
num_meses_var44_ult3               int64
num_op_var39_comer_ult1            int64
num_op_var39_comer_ult3            int64
num_op_var40_comer_ult1            int64
num_op_var40_comer_ult3            int64
num_op_var40_efect_ult1            int64
num_op_var40_efect_ult3            int64
num_op_var41_comer_ult1            int64
num_op_var41_comer_ult3            int64
num_op_var41_efect_ult1            int64
num_op_var41_efect_ult3            int64
num_op_var39_efect_ult1            int64
num_op_var39_efect_ult3            int64
num_reemb_var13_ult1               int64
num_reemb_var17_hace3              int64
num_reemb_var17_ult1               int64
num_reemb_var33_ult1               int64
num_sal_var16_ult1                 int64
num_var43_emit_ult1                int64
num_var43_recib_ult1               int64
num_trasp_var11_ult1               int64
num_trasp_var17_in_hace3           int64
num_trasp_var17_in_ult1            int64
num_trasp_var17_out_ult1           int64
num_trasp_var33_in_hace3           int64
num_trasp_var33_in_ult1            int64
num_trasp_var33_out_ult1           int64
num_venta_var44_hace3              int64
num_venta_var44_ult1               int64
num_var45_hace2                    int64
num_var45_hace3                    int64
num_var45_ult1                     int64
num_var45_ult3                     int64
saldo_medio_var5_hace2           float64
saldo_medio_var5_hace3           float64
saldo_medio_var5_ult1            float64
saldo_medio_var5_ult3            float64
saldo_medio_var8_hace2           float64
saldo_medio_var8_hace3           float64
saldo_medio_var8_ult1            float64
saldo_medio_var8_ult3            float64
saldo_medio_var12_hace2          float64
saldo_medio_var12_hace3          float64
saldo_medio_var12_ult1           float64
saldo_medio_var12_ult3           float64
saldo_medio_var13_corto_hace2    float64
saldo_medio_var13_corto_hace3    float64
saldo_medio_var13_corto_ult1     float64
saldo_medio_var13_corto_ult3     float64
saldo_medio_var13_largo_hace2    float64
saldo_medio_var13_largo_hace3    float64
saldo_medio_var13_largo_ult1     float64
saldo_medio_var13_largo_ult3     float64
saldo_medio_var13_medio_hace2    float64
saldo_medio_var13_medio_ult1       int64
saldo_medio_var13_medio_ult3     float64
saldo_medio_var17_hace2          float64
saldo_medio_var17_hace3          float64
saldo_medio_var17_ult1           float64
saldo_medio_var17_ult3           float64
saldo_medio_var29_hace2          float64
saldo_medio_var29_hace3          float64
saldo_medio_var29_ult1           float64
saldo_medio_var29_ult3           float64
saldo_medio_var33_hace2          float64
saldo_medio_var33_hace3          float64
saldo_medio_var33_ult1           float64
saldo_medio_var33_ult3           float64
saldo_medio_var44_hace2          float64
saldo_medio_var44_hace3          float64
saldo_medio_var44_ult1           float64
saldo_medio_var44_ult3           float64
var38                            float64
TARGET                             int64
dtype: object
```

In [68]:
```
pd.set_option('max_row',None)
```

In [73]:
```
unicos = pd.DataFrame(df.nunique())
```

In [74]:
```
unicos.loc[unicos[0]==1].index
```

Out[74]:
```
Index(['ind_var2_0', 'ind_var2', 'ind_var27_0', 'ind_var28_0', 'ind_var28',
       'ind_var27', 'ind_var41', 'ind_var46_0', 'ind_var46', 'num_var27_0',
       'num_var28_0', 'num_var28', 'num_var27', 'num_var41', 'num_var46_0',
       'num_var46', 'saldo_var28', 'saldo_var27', 'saldo_var41', 'saldo_var46',
       'imp_amort_var18_hace3', 'imp_amort_var34_hace3',
       'imp_reemb_var13_hace3', 'imp_reemb_var33_hace3',
       'imp_trasp_var17_out_hace3', 'imp_trasp_var33_out_hace3',
       'num_var2_0_ult1', 'num_var2_ult1', 'num_reemb_var13_hace3',
       'num_reemb_var33_hace3', 'num_trasp_var17_out_hace3',
       'num_trasp_var33_out_hace3', 'saldo_var2_ult1',
       'saldo_medio_var13_medio_hace3'],
      dtype='object')
```

Algumas colunas só tem um valor único, e por isso podem ser descartadas

In [75]:
```
df.drop(list(unicos.loc[unicos[0]==1].index),axis=1,inplace=True)
```

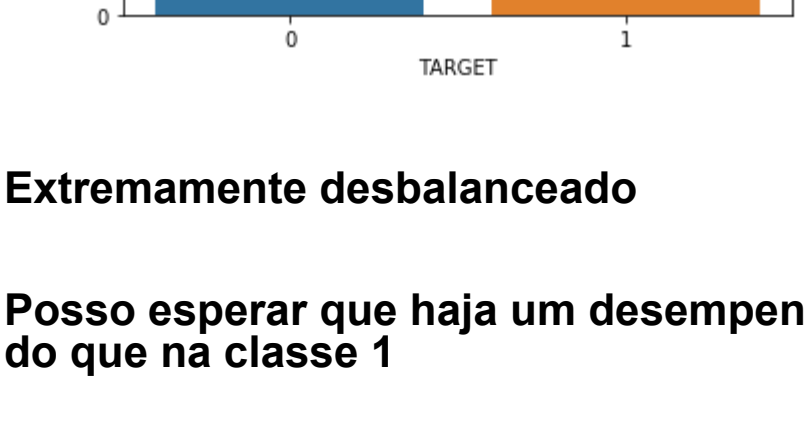
In [76]:
```
unicos.sort_values(0,ascending=False)
```
Out[76]:

	0
var38	57736
saldo_medio_var5_ult3	17330
saldo_var30	16940
saldo_var42	15730
saldo_medio_var5_ult1	14778
saldo_medio_var5_hace2	14486
saldo_var5	11642
imp_op_var39_comer_ult3	9099
imp_op_var41_comer_ult3	8961
imp_op_var39_ult1	8149
imp_op_var41_ult1	8032
saldo_medio_var5_hace3	7787
imp_op_var39_comer_ult1	7551
imp_op_var41_comer_ult1	7421
saldo_var37	4041
imp_trans_var37_ult1	3831
saldo_medio_var12_ult3	3447
saldo_medio_var12_ult1	3405
saldo_var12	3059
saldo_medio_var12_hace2	2834
saldo_var24	2614
saldo_medio_var13_corto_ult3	2576
imp_var43_emit_ult1	2342
saldo_medio_var8_ult3	2173
saldo_medio_var8_ult1	2116
saldo_var8	1989
saldo_medio_var13_corto_hace2	1628
saldo_var26	1592
saldo_var25	1524
saldo_medio_var8_hace2	1325
saldo_medio_var12_hace3	1152
saldo_medio_var13_corto_hace3	968
saldo_medio_var13_corto_ult1	943
saldo_var13	859
saldo_var13_corto	730
imp_ent_var16_ult1	596
saldo_medio_var13_largo_ult3	511
imp_op_var39_efect_ult3	462
imp_op_var41_efect_ult3	454
saldo_medio_var8_hace3	439
imp_aport_var13_hace3	425
imp_op_var40_comer_ult3	346
saldo_var14	345
imp_op_var39_efect_ult1	336
imp_op_var41_efect_ult1	331
saldo_medio_var13_largo_hace2	295
imp_op_var40_comer_ult1	293
saldo_var1	285
saldo_var40	282
saldo_var31	278
saldo_medio_var13_largo_hace3	264
saldo_medio_var13_largo_ult1	233
saldo_var13_largo	229
imp_op_var40_ult1	224
var3	208
imp_var7_recib_ult1	184
imp_aport_var13_ult1	182
saldo_var20	176
num_var45_ult3	172
saldo_medio_var44_ult1	141
saldo_medio_var44_ult3	141
saldo_var44	129
saldo_medio_var17_ult3	119
saldo_medio_var17_ult1	119
saldo_var17	111
var15	100
saldo_medio_var44_hace2	99
num_op_var39_ult3	99
num_op_var41_ult3	96
num_var45_ult1	94
num_op_var39_comer_ult3	92
saldo_medio_var17_hace2	88
num_op_var41_comer_ult3	88
num_var45_hace2	85
imp_compra_var44_ult1	85
saldo_var32	75
num_op_var39_ult1	71
num_med_var45_ult3	71
num_op_var41_ult1	68
imp_sal_var16_ult1	66
num_var45_hace3	66
num_op_var39_comer_ult1	63
num_op_var41_comer_ult1	60
num_op_var41_hace2	51
num_op_var39_hace2	50
saldo_medio_var33_ult3	48
saldo_medio_var33_ult1	48
saldo_var33	48
num_op_var40_comer_ult3	47
imp_venta_var44_ult1	45
saldo_medio_var33_hace2	43
imp_aport_var17_ult1	42
num_op_var41_efect_ult3	40
num_op_var39_efect_ult3	40
num_var43_recib_ult1	38
num_op_var40_ult3	35
num_op_var40_comer_ult1	34
num_var22_ult3	33
imp_compra_var44_hace3	33
imp_reemb_var13_ult1	33
saldo_medio_var44_hace3	33
num_op_var40_ult1	30
imp_op_var40_efect_ult3	29
delta_imp_aport_var13_1y3	27
num_var43_emit_ult1	27
num_op_var39_efect_ult1	26
num_op_var41_efect_ult1	25
var21	24
saldo_medio_var33_hace3	24
imp_op_var40_efect_ult1	23
num_var37_0	22
num_var37	22
num_op_var39_hace3	22
num_var22_hace2	22
num_op_var41_hace3	22
num_var37_med_ult2	21
imp_aport_var17_hace3	20
imp_reemb_var17_ult1	20
num_var22_hace3	19
num_op_var40_hace2	19
num_var22_ult1	18
num_trasp_var11_ult1	18
saldo_medio_var17_hace3	18
delta_imp_compra_var44_1y3	17
imp_aport_var33_hace3	16
num_med_var22_ult3	15
num_var35	13
num_ent_var16_ult1	13
num_var17_0	11
num_var31_0	11
num_var30_0	11
num_compra_var44_ult1	10
num_venta_var44_ult1	9
delta_imp_aport_var33_1y3	9
num_var31	9
num_var30	9
delta_num_compra_var44_1y3	9
num_var25	9
num_var39_0	9
num_var41_0	9
num_var17	9
num_op_var40_efect_ult3	9
num_var26_0	9
num_var26	9
num_var25_0	9
num_var42_0	8
num_var4	8
num_aport_var13_hace3	8
num_aport_var17_ult1	8
imp_trasp_var33_in_hace3	7
num_aport_var13_ult1	7
num_op_var40_efect_ult1	7
num_var13	7
num_var13_largo	7
num_var42	7
num_var13_largo_0	7
num_var13_0	7
delta_imp_aport_var17_1y3	7
imp_trasp_var33_in_ult1	6
num_var7_recib_ult1	6
num_sal_var16_ult1	6
delta_num_aport_var17_1y3	6
delta_num_aport_var13_1y3	6
imp_aport_var33_ult1	6
num_var12_0	6
num_reemb_var17_ult1	5
num_var32_0	5
delta_imp_venta_var44_1y3	5
num_var32	5
saldo_medio_var29_hace2	5
imp_trasp_var17_out_ult1	5
imp_trasp_var17_in_ult1	5
num_aport_var17_hace3	5
num_var14_0	5
num_var5_0	5
var36	5
num_op_var40_hace3	5
delta_num_venta_var44_1y3	5
num_var5	5
num_var33_0	4
num_var12	4
num_meses_var8_ult3	4
imp_var7_emit_ult1	4
num_aport_var33_hace3	4
imp_venta_var44_hace3	4
num_var14	4
num_var24_0	4
num_meses_var44_ult3	4
num_meses_var39_vig_ult3	4
num_meses_var33_ult3	4
num_compra_var44_hace3	4
num_meses_var17_ult3	4
num_meses_var13_largo_ult3	4
num_meses_var13_corto_ult3	4
num_meses_var12_ult3	4
delta_num_aport_var33_1y3	4
num_meses_var5_ult3	4
saldo_medio_var29_ult3	4
saldo_medio_var29_ult1	4
num_var13_corto	3
num_trasp_var33_in_ult1	3
saldo_var18	3
imp_amort_var18_ult1	3
num_venta_var44_hace3	3
num_var40_0	3
num_var24	3
num_var33	3
delta_num_trasp_var33_in_1y3	3
num_trasp_var17_in_hace3	3
num_var13_corto_0	3
imp_trasp_var17_in_hace3	3
saldo_medio_var13_medio_hace2	3
saldo_var34	3
saldo_medio_var13_medio_ult1	3
saldo_medio_var13_medio_ult3	3
saldo_var6	3
imp_amort_var34_ult1	3
num_aport_var33_ult1	3
num_var44_0	3
delta_imp_reemb_var17_1y3	3
num_var1	3
delta_imp_trasp_var17_in_1y3	3
num_var1_0	3
delta_imp_trasp_var33_in_1y3	3
delta_num_trasp_var17_in_1y3	3
saldo_var13_medio	3
saldo_var29	3
num_var8_0	3
delta_num_reemb_var17_1y3	3
num_meses_var29_ult3	3
ind_var1	2
num_var6_0	2
ind_var1_0	2
num_var8	2
num_reemb_var13_ult1	2
ind_var5	2
ind_var5_0	2
num_var18_0	2
num_meses_var13_medio_ult3	2
num_var18	2
num_reemb_var17_hace3	2
num_var6	2
ind_var25_cte	2
ind_var26	2
num_var20_0	2
ind_var26_cte	2
num_var20	2
saldo_medio_var29_hace3	2
ind_var26_0	2
ind_var13_largo_0	2
ind_var6_0	2
ind_var13_0	2
ind_var13_medio_0	2
ind_var13_medio	2
ind_var13	2
ind_var14_0	2
ind_var14	2
ind_var13_corto	2
ind_var17_0	2
ind_var13_corto_0	2
ind_var17	2
ind_var18_0	2
ind_var18	2
ind_var19	2
ind_var20_0	2
ind_var20	2
num_reemb_var33_ult1	2
ind_var12	2
ind_var24	2
ind_var12_0	2
num_trasp_var33_out_ult1	2
ind_var8	2
ind_var8_0	2
num_trasp_var33_in_hace3	2
num_trasp_var17_out_ult1	2
num_trasp_var17_in_ult1	2
ind_var6	2
ind_var13_largo	2
num_var13_medio_0	2
num_var13_medio	2
ind_var24_0	2
TARGET	2
ind_var25_0	2
delta_imp_reemb_var33_1y3	2
ind_var29	2
delta_imp_amort_var18_1y3	2
delta_imp_amort_var34_1y3	2
ind_var39	2
num_var44	2
delta_imp_reemb_var13_1y3	2
delta_imp_trasp_var17_out_1y3	2
ind_var30	2
delta_imp_trasp_var33_out_1y3	2
num_var39	2
delta_num_reemb_var13_1y3	2
delta_num_reemb_var33_1y3	2
delta_num_trasp_var17_out_1y3	2
ind_var29_0	2
ind_var30_0	2
ind_var31_0	2
num_var34	2
ind_var37_cte	2
ind_var41_0	2
ind_var40	2
ind_var40_0	2
ind_var39_0	2
ind_var37	2
ind_var37_0	2
ind_var34	2
ind_var31	2
ind_var34_0	2
ind_var33	2
ind_var33_0	2
ind_var32	2
ind_var32_0	2
ind_var32_cte	2
num_var40	2
delta_num_trasp_var33_out_1y3	2
ind_var7_emit_ult1	2
imp_reemb_var33_ult1	2
ind_var7_recib_ult1	2
ind_var9_cte_ult1	2
ind_var9_ult1	2
ind_var43_emit_ult1	2
num_var29	2
imp_trasp_var33_out_ult1	2
ind_var43_recib_ult1	2
num_var29_0	2
ind_var25	2
ind_var10cte_ult1	2
ind_var10_ult1	2
num_var7_emit_ult1	2
imp_reemb_var17_hace3	2
num_var34_0	2
ind_var44	2
ind_var44_0	2
num_var27	1
saldo_medio_var13_medio_hace3	1
ind_var46_0	1
num_var27_0	1
ind_var2_0	1
num_var28	1
saldo_var27	1
saldo_var28	1
ind_var2	1
num_var2_0_ult1	1
num_var28_0	1
num_var2_ult1	1
ind_var28	1
saldo_var41	1
imp_reemb_var33_hace3	1
ind_var28_0	1
ind_var27	1
num_trasp_var17_out_hace3	1
imp_amort_var34_hace3	1
imp_amort_var18_hace3	1
ind_var46	1
saldo_var2_ult1	1
imp_reemb_var13_hace3	1
num_var41	1
num_trasp_var33_out_hace3	1
ind_var27_0	1
imp_trasp_var17_out_hace3	1
num_reemb_var33_hace3	1
imp_trasp_var33_out_hace3	1
num_var46_0	1
num_var46	1
num_reemb_var13_hace3	1
saldo_var46	1
ind_var41	1

A coluna var38 tem 57736 valores únicos, sendo que há 76020 linhas. Avaliar se preciso apagar ela

In [77]:
```
df.duplicated().sum()
```
Out[77]: 4807

Removendo duplicatas

In [78]:
```
df.drop_duplicates(inplace=True)
```
In [13]:
```
sns.countplot(data=df,x='TARGET');
```


Extremamente desbalanceado

Posso esperar que haja um desempenho muito melhor na identificação da classe 0 do que na classe 1

Posso tentar um balanceamento por undersampling (Tomek links)

É preciso normalizar os dados antes de aplicar PCA

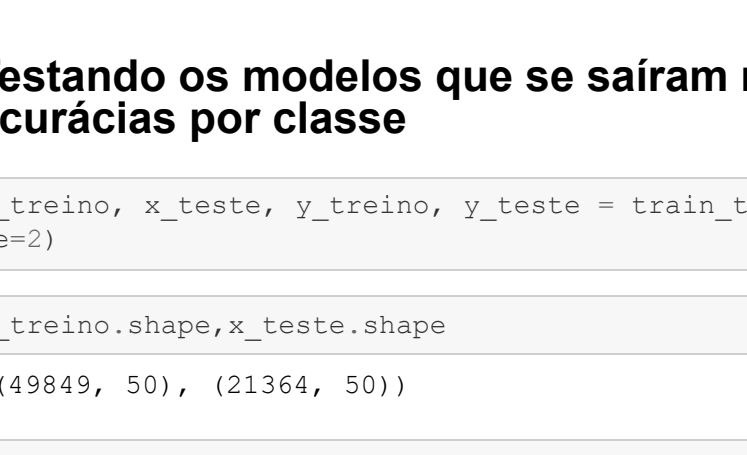
In [79]:
```
x_values = df.iloc[:,0:df.shape[1]].values
y_values = df.iloc[:,df.shape[1]-1].values
```
In [80]:
```
min_max = MinMaxScaler()
min_max.fit(x_values)
x_values_s = min_max.transform(x_values)
```
In [81]:
```
x_values_s
```
Out[81]:
```
array([[0.99976406, 0.18      , 0.        , ..., 0.        , 0.00154526,
        0.        ],
       [0.99976406, 0.29      , 0.        , ..., 0.        , 0.0020025 ,
        0.        ],
       [0.99976406, 0.18      , 0.        , ..., 0.        , 0.00282212,
        0.        ],
       ...,
       [0.99976406, 0.34      , 0.        , ..., 0.        , 0.00515084,
        0.        ],
       [0.99976406, 0.18      , 0.        , ..., 0.        , 0.003126  ,
        0.        ],
       [0.99976406, 0.2       , 0.        , ..., 0.        , 0.00359128,
        0.        ]])
```
In [82]:
```
pca = PCA(n_components=50)
pca.fit(x_values_s)
x_componentes = pca_fit.transform(x_values_s)
```
In [83]:
```
pca_fit.explained_variance_ratio_.sum()
```
Out[83]: 0.9921693943006986

In [84]:
```
x_componentes.shape
```
Out[84]: (71213, 50)

50 componentes explicam praticamente toda a informação (99,2%)

Fazendo testes sem balanceamento com validação cruzada

RandomForest, Regressão logística, SVM, NaiveBayes, LDA, KNN, XGBClassifier

In [89]:
```
kf = KFold(n_splits=10, shuffle=True)
```
In [123]:
```
modelos = [('RandomForest',RandomForestClassifier()),('KNN',KNeighborsClassifier()),('Reg. L.',Logistic
Regression()),\
          ('SVM',SVC()),('NB',GaussianNB()),('LDA',LinearDiscriminantAnalysis()),('XGBoost',XGBClassifi
er())]
```
In [ ]:
```
resultados = []
for _,modelo in modelos:
    resultados.append(cross_val_score(estimator=modelo,X=x_componentes,y=y_values,cv=kf,scoring='balanc
ed_accuracy'))
```
In [125]:
```
plt.boxplot(x=resultados,labels=[x[0] for x in modelos]);
```


Os modelos RandomForest, Regressão Logística, LDA e XGBoost preveram perfeitamente.

Testando os modelos que se saíram melhor individualmente para coletar as acurácias por classe

In [85]:
```
x_treino, x_teste, y_treino, y_teste = train_test_split(x_componentes,y_values,test_size=0.3,random_sta
te=2)
```
In [86]:
```
x_treino.shape,x_teste.shape
```
Out[86]: ((49849, 50), (21364, 50))

In [87]:
```
randf = RandomForestClassifier()
randf.fit(x_treino,y_treino)
```
Out[87]:
```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```
In [88]:
```
prev_randf = randf.predict(x_teste)
```
In [89]:
```
print(classification_report(y_teste,prev_randf))
```
```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     20536
           1       1.00      1.00      1.00       828

    accuracy                           1.00     21364
   macro avg       1.00      1.00      1.00     21364
weighted avg       1.00      1.00      1.00     21364
```

O Random Forest já prevê perfeitamente, não há porque ir em frente