SARA GARCÍA DE LA MORA DÍAZ
DANIEL IGNACIO MATEOS
SILVIA BORREGO CAPITÁN
DANIEL GONZÁLEZ MARTÍNEZ
LOURDES PASCUAL CAMPOY
JAKE PATRICK TAYLOR-KING

PROPAGATION MODELS

ACCENTURE

VII UCM Modelling Week
Master in Mathematical Engineering UCM
10-14 june 2013

Index

INTRODUCTION

- The aim is to develop models of propagation of content on the social network - Twitter.
- The objective is to predict whether, given a hashtag, this will be used in the future. To do this, we wish to predict a binary variable.
- The speed at which a comment can spread may be harmful to the reputation of a company.

PROBLEM DESCRIPTION

- **General goal:** to develop models of content propagation within social networks. The general problem is rather complex and must have into account multiple factors.
- We will restrict ourselves to **Twitter** (wwww.twitter.com). The proposed problem is to identify whether a given content will be used by a user (hashtag) as a function of how this hashtag or other has been used before.
- To perform this task the following data will be available:
    - A table with the directed graph of (a part of) the twitter network, including all followers and friends connections.
    - A master table of users, with a user per record including her attributes, including all users within the previous table.
    - A table of hashtags. This table includes hashtags used by the users in the users table. The table includes details of the hashtag: message ID, user_id, datetime of the message, number of tweets, followers and friends of the user at the moment in which the message was sent.
    - Target table. This table contains records including user ID, hashtag ID, and the target variable: 1 if the user uses the hashtag, 0 if not.

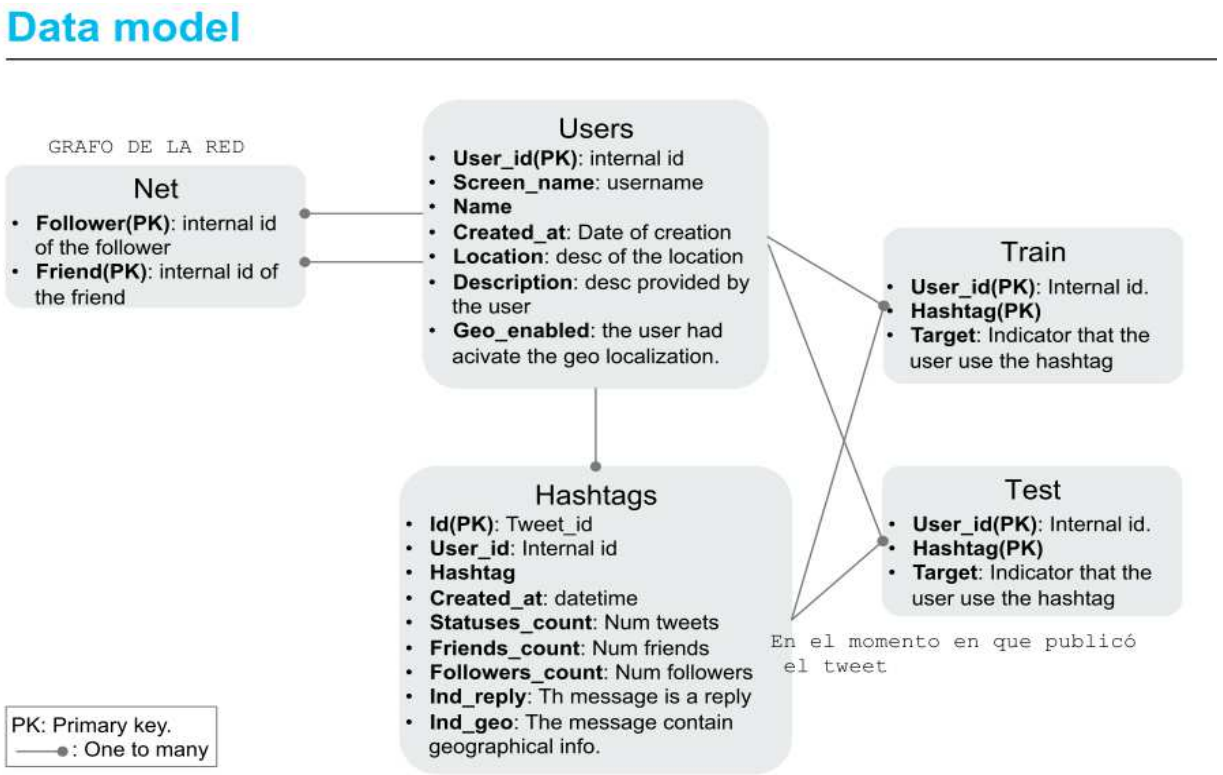
MATHEMATICAL TREATMENT. MODELLING

The desired model predicts a **binary target variable** (0 if the user DOES NOT use the hashtag / 1 if the user DO use the hashtag):

$$y_{ij} = \begin{cases} 1, \; if\ the\ user\ i\ uses\ the\ hashtag\ j \\ 0, \; if\ the\ user\ i\ doesn't\ use\ the\ hashtag\ j \end{cases}$$

To model the event starting with the proposed table, the models should allow capturing non-lineal effects as well as interactions between the explicative factors defined before.

We propose to use logistic models with cross effects, neural networks or decision trees.



Data model

GRAFO DE LA RED

Net
- **Follower(PK)**: internal id of the follower
- **Friend(PK)**: internal id of the friend

Users
- **User_id(PK)**: internal id
- **Screen_name**: username
- **Name**
- **Created_at**: Date of creation
- **Location**: desc of the location
- **Description**: desc provided by the user
- **Geo_enabled**: the user had acivate the geo localization.

Train
- **User_id(PK)**: Internal id.
- **Hashtag(PK)**
- **Target**: Indicator that the user use the hashtag

Hashtags
- **Id(PK)**: Tweet_id
- **User_id**: Internal id
- **Hashtag**
- **Created_at**: datetime
- **Statuses_count**: Num tweets
- **Friends_count**: Num friends
- **Followers_count**: Num followers
- **Ind_reply**: Th message is a reply
- **Ind_geo**: The message contain geographical info.

Test
- **User_id(PK)**: Internal id.
- **Hashtag(PK)**
- **Target**: Indicator that the user use the hashtag

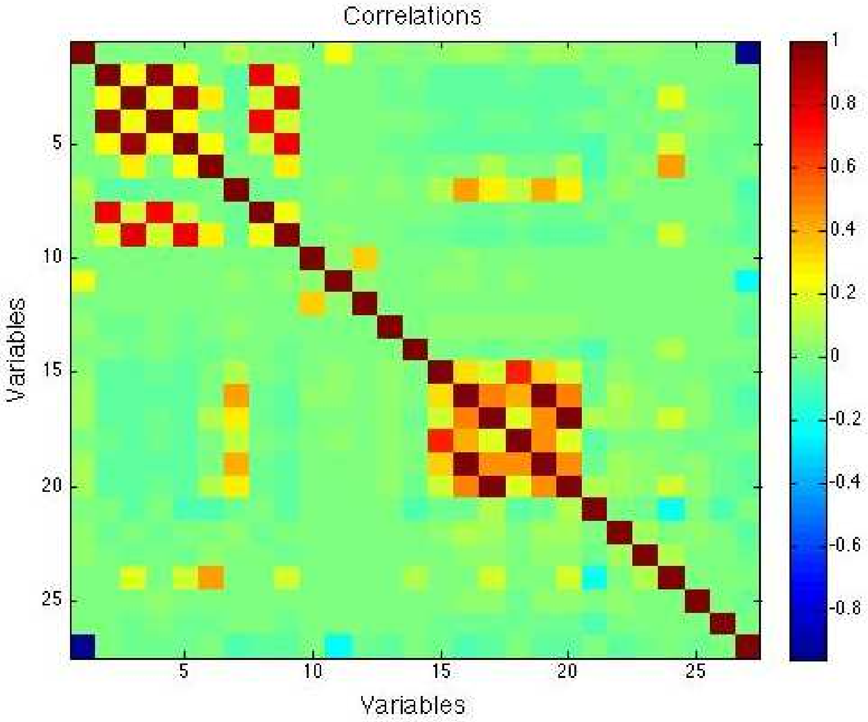
En el momento en que publicó el tweet

PK: Primary key.
——●: One to many

DATA TREATMENT

CORRELATIONS

The first thing we do is a study of the correlations between variables with a graphic made in Matlab. Those who are highly correlated with other don't provide information to the model so will not be taken into account.



In the following figure, the squares with red tones indicate a high correlation, while the green tones indicate no significant correlation. We can see that most of them are green; therefore, most of the variables provide information to the model. However, there are two variables which have the same values (highly correlated), so one of them has been removed.

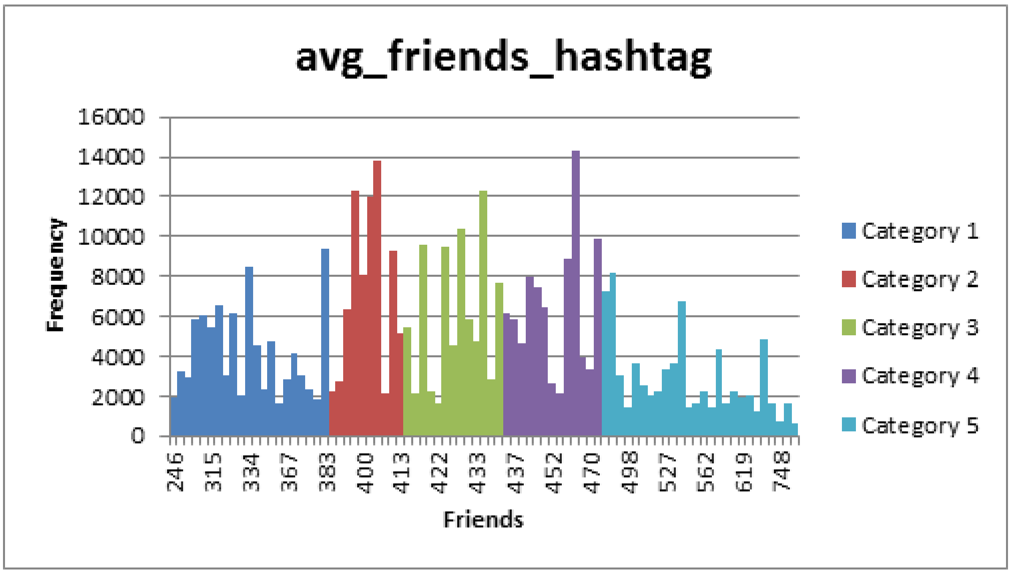
CATEGORIZATION

The next step is to categorize the variables (redefine) in order to obtain a better prediction of the model. To do this we look at the graphs of the frequency of each variable, and classify the data according to the range where they fall. Each range is defined so that each contain about the same amount of data.

We realize that most of the variables have missing values. We go to use categorization to solve this problem.

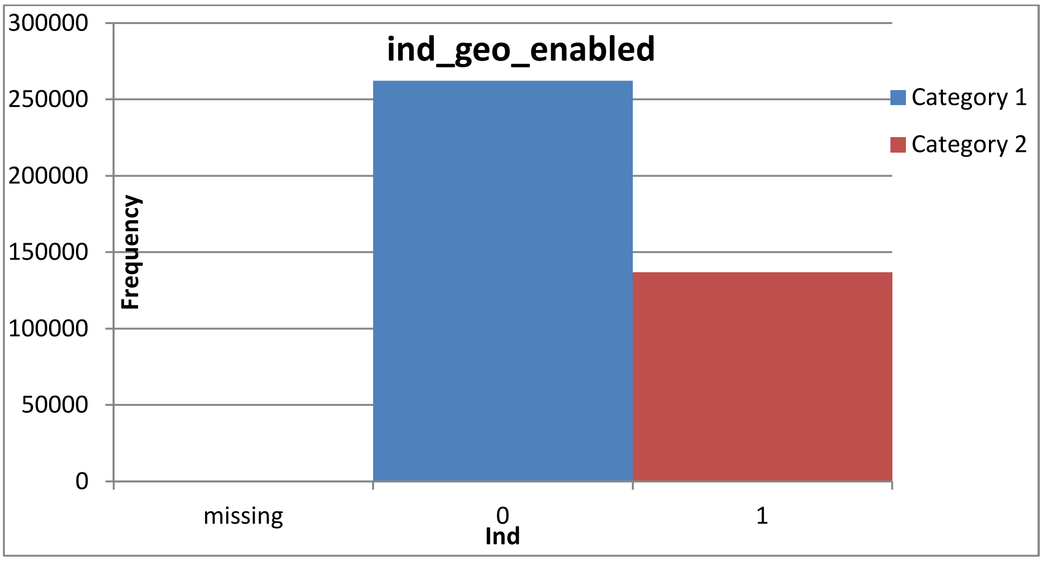
Let's look in more detail this classification:

1. The variables that don't contain missing, take nominal values, starting with the value 1 to the number of categories in which data have been distributed. Each category contains approximately the same number of data.

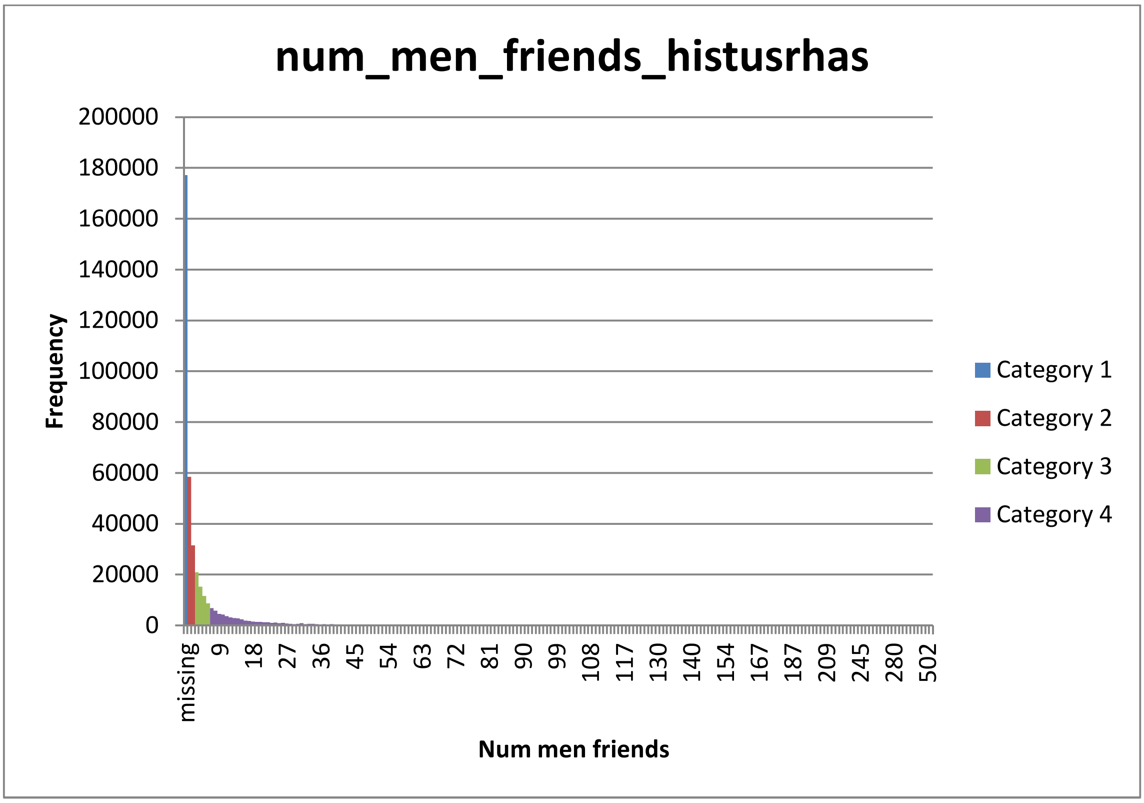


2. Variables that have the missing values categorized as follows:

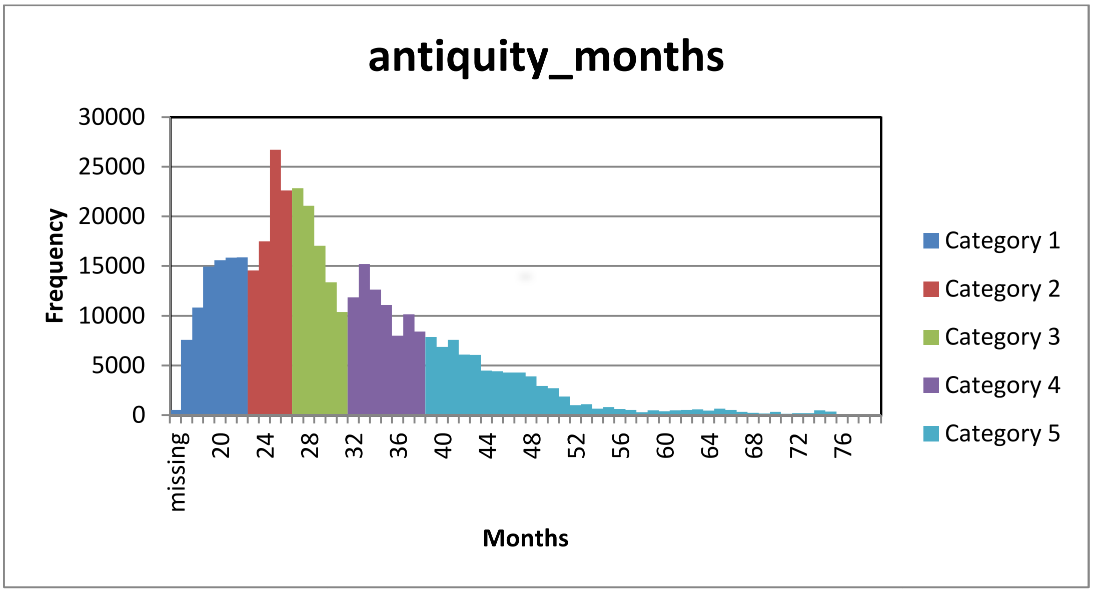
    a. When the variables are binary (in this problem are indicators that take the value 0 if there aren't information on the question referred and 1 otherwise) we assign the value 0 to the values missing.



b. In the other variables, the missing value is assigned the value 0, identifying as one category, and the other values will be classified as in case 1.



The exception is the variable "antiquity_months" whose percentage of missing is low and included in the same category as those with a shorter length giving the value 1.



NEW VARIABLES

As one of the objectives of the problem is the introduction of new variables, we have defined some variables that we believe may improving the predictive capabilities of the model, which also have been categorized too:

- **Hashtag content**: The meanings of hashtags have been grouped in six categories: *youth-topics, politics*, *sports*, *nonsense*, *media* and *others*. The variable take values 1 to 6 respectively.

- **User's activity time**: Interval day when a user is connected more often. If the range is from 00:00 to 10:00 then the variable will take the value 1. If you are from 10:00 to 14:00 will take the value 2. If you are from 14:00 to 19:00 will take the value 3. The remaining range will take the value 4.

- Other variables have been proposed, as **geographic location**, but for reasons of time have not been introduced in the model.

APPLICATION OF OVERSAMPLING TECHNIQUES

The success rate that has our training sample is 5.51%, because the training table contains a total of 399.403 records, of which 21.988 have the value TARGET = 1. As this percentage is very low, we will do an oversampling that we know is a sampling technique that is usually used when we have a low proportion of positive cases binomial classifications. If not, our model will reject some positive cases.

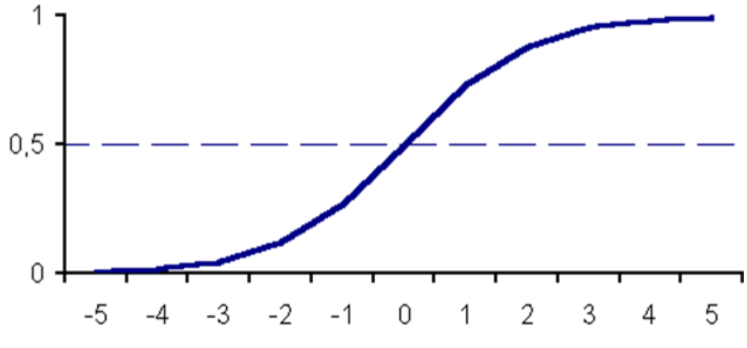
At first, we will sort the training table by target indicating method (SRS) and the number of observations per stratum. So our table will contain two stratums: one containing 21.988 records for TARGET = 0 and the other with the same observations and TARGET = 1.

Therefore, we have the true filtered training sample and the desired volumes, so we are in conditions to develop the different models: logistic model, decision tree model and neuronal network model, once it has been set for the training sample and validation sample.

LOGISTIC REGRESSION

Logistic regression models are regression models that allow us to study binomial variables dependent on other variables.

The logistic function allows us a transformation in the range 0 to 1, the natural range of a probability.



The model equation is

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$$

Independent variables are represented by x and p_i is the probability for hashtag's propagation.

The main objective of logistic regression is to model the influence of different factors on the probability of an event to occur.

To apply logistic regression to the training sample oversampled we will develop three models:

- Inputting initials variables.
- Inputting initials variables and new variables.
- Inputting initials variables, new variables and iterations.

As logistic regression is a multivariate method, it is interesting to select the best independent variable to input into the model. Therefore we use the stepwise method.

STEPWISE METHOD

This technique is an algorithm that establishes criteria input and output of covariates.

As mentioned above, we will include a lot of variables and stepwise technique will discriminate those that are not significant. Also, develop a logistic model.

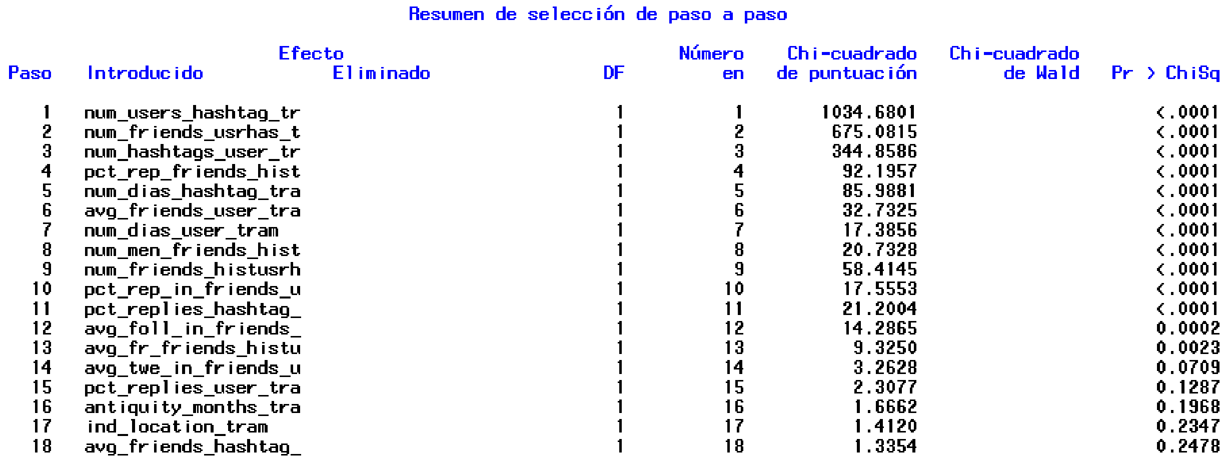
On the other hand we will use the ROC curve which gives us information about the predictive power of our model.

It's important to say that it has modeled the probability of 1, i.e., the probability that a person follow a hashtag.

On the other hand we introduce in the code a lot of variables, violating the principle of parsimony says that models should be as simple as possible, but we have decided to introduce all possible variables with all possible combinations two by two, and the stepwise technique is responsible for discriminating the variables that should enter according to their p-value. We have eliminated some variables because they don't have contributed in a significant way in our model.

INPUTTING INITIALS VARIABLES.

After running the code we can see a summary of the method:

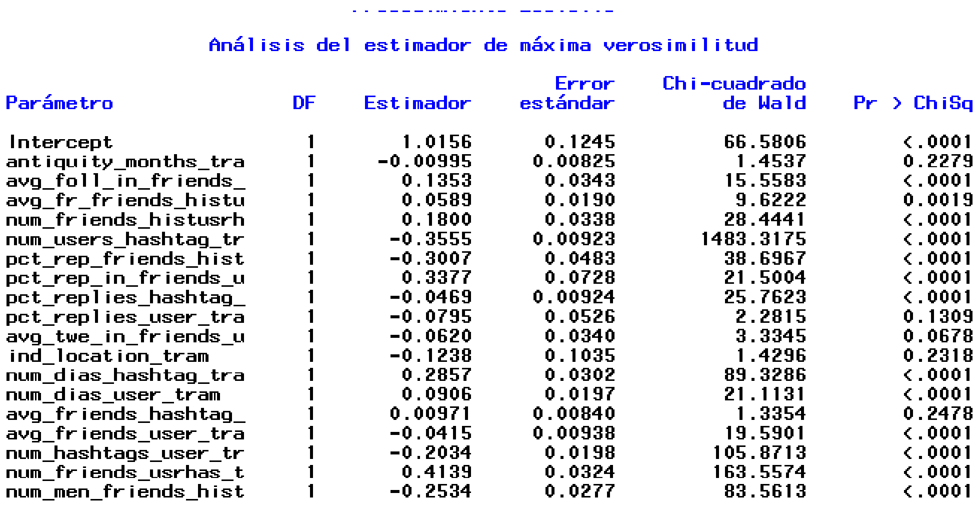


```
                          Resumen de selección de paso a paso

                    Efecto                                 Número   Chi-cuadrado   Chi-cuadrado
Paso    Introducido        Eliminado              DF        en     de puntuación       de Wald    Pr > ChiSq

   1    num_users_hashtag_tr                        1         1       1034.6801                    <.0001
   2    num_friends_usrhas_t                        1         2        675.0815                    <.0001
   3    num_hashtags_user_tr                        1         3        344.8586                    <.0001
   4    pct_rep_friends_hist                        1         4         92.1957                    <.0001
   5    num_dias_hashtag_tra                        1         5         85.9881                    <.0001
   6    avg_friends_user_tra                        1         6         32.7325                    <.0001
   7    num_dias_user_tram                          1         7         17.3856                    <.0001
   8    num_men_friends_hist                        1         8         20.7328                    <.0001
   9    num_friends_histusrh                        1         9         58.4145                    <.0001
  10    pct_rep_in_friends_u                        1        10         17.5553                    <.0001
  11    pct_replies_hashtag_                        1        11         21.2004                    <.0001
  12    avg_foll_in_friends_                        1        12         14.2865                    0.0002
  13    avg_fr_friends_histu                        1        13          9.3250                    0.0023
  14    avg_twe_in_friends_u                        1        14          3.2628                    0.0709
  15    pct_replies_user_tra                        1        15          2.3077                    0.1287
  16    antiquity_months_tra                        1        16          1.6662                    0.1968
  17    ind_location_tram                           1        17          1.4120                    0.2347
  18    avg_friends_hashtag_                        1        18          1.3354                    0.2478
```

We can observe the variables that have been introduced into the model, and the p-values associated with each one.

All variables are significant because their p-values are small.

Now we can observe a table of estimates associated with each variable, i.e. the weight that a variable has on the probability of success.

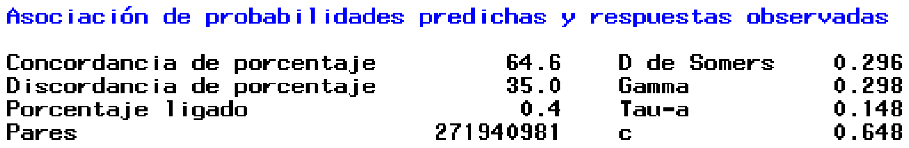


```
              Análisis del estimador de máxima verosimilitud

                                                Error    Chi-cuadrado
Parámetro                DF     Estimador     estándar         de Wald    Pr > ChiSq

Intercept                 1        1.0156       0.1245         66.5806        <.0001
antiquity_months_tra      1      -0.00995      0.00825          1.4537        0.2279
avg_foll_in_friends_      1        0.1353       0.0343         15.5583        <.0001
avg_fr_friends_histu      1        0.0589       0.0190          9.6222        0.0019
num_friends_histusrh      1        0.1800       0.0338         28.4441        <.0001
num_users_hashtag_tr      1       -0.3555      0.00923       1483.3175        <.0001
pct_rep_friends_hist      1       -0.3007       0.0483         38.6967        <.0001
pct_rep_in_friends_u      1        0.3377       0.0728         21.5004        <.0001
pct_replies_hashtag_      1       -0.0469      0.00924         25.7623        <.0001
pct_replies_user_tra      1       -0.0795       0.0526          2.2815        0.1309
avg_twe_in_friends_u      1       -0.0620       0.0340          3.3345        0.0678
ind_location_tram         1       -0.1238       0.1035          1.4296        0.2318
num_dias_hashtag_tra      1        0.2857       0.0302         89.3286        <.0001
num_dias_user_tram        1        0.0906       0.0197         21.1131        <.0001
avg_friends_hashtag_      1       0.00971      0.00840          1.3354        0.2478
avg_friends_user_tra      1       -0.0415      0.00938         19.5901        <.0001
num_hashtags_user_tr      1       -0.2034       0.0198        105.8713        <.0001
num_friends_usrhas_t      1        0.4139       0.0324        163.5574        <.0001
num_men_friends_hist      1       -0.2534       0.0277         83.5613        <.0001
```

We can also get an idea of how each explicative variable affects in the likelihood that a person follows a hasthag.

The process of association of predicted probabilities and observed responses is to consider all pairs of values and sort them into two concordant and discordant pair.

Then we get four measures that indicate the predictive ability of the model.



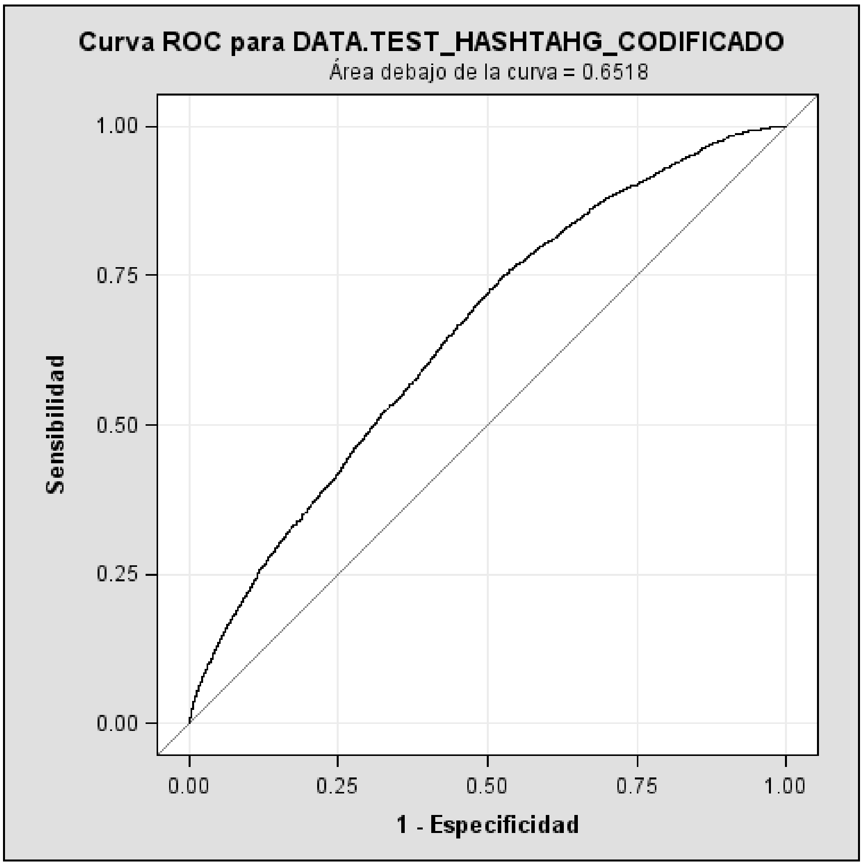
```
Asociación de probabilidades predichas y respuestas observadas

Concordancia de porcentaje          64.6    D de Somers    0.296
Discordancia de porcentaje          35.0    Gamma          0.298
Porcentaje ligado                    0.4    Tau-a          0.148
Pares                          271940981    c              0.648
```

The four measures mentioned above indicate the predictive ability of the model so that if they are near zero indicate that the predictive power of the model is not very good, and the bigger they are, the greater the predictive power of our model.

We note that such measures are not very high so that our model does not predict very well.

CURVA ROC

A parameter to evaluate the goodness of the test is the area under the ROC curve that will take values between 1 (perfect test) and 0.5 (useless test).

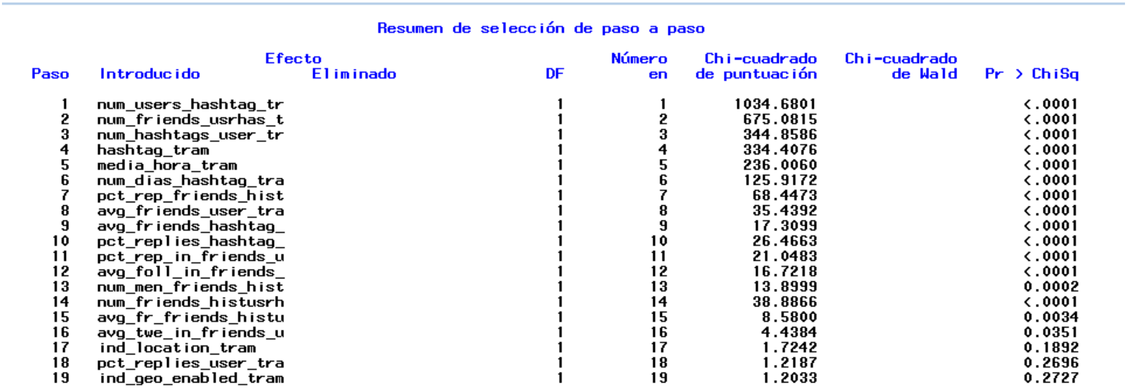


INPUTTING INITIALS VARIABLES AND NEW VARIABLES

Now we will develop the same procedure but considering the two new variables defined:

- ✓ HASTAGH_TRAM.
- ✓ MEDIA_HORA_TRAM

First we have the summary of the variables that have been included in the model.

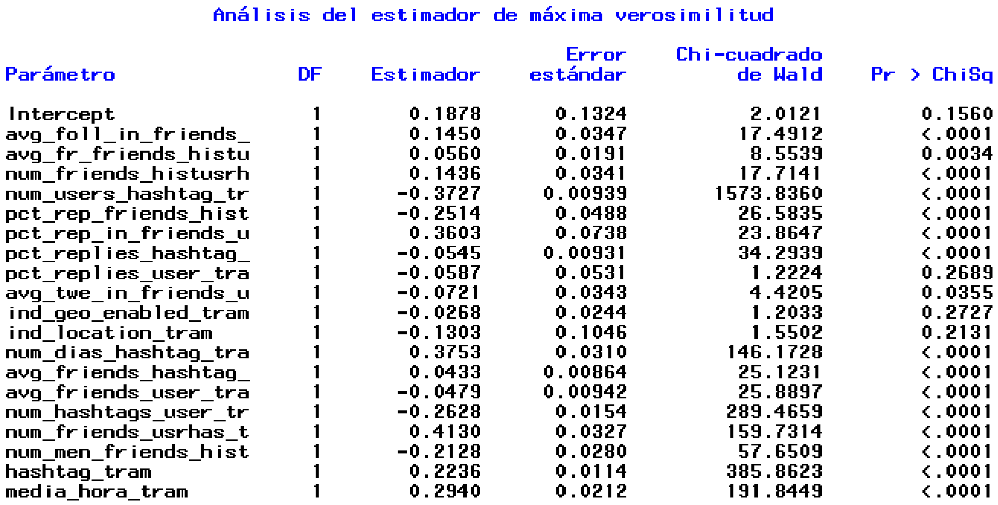


```
                              Resumen de selección de paso a paso

                      Efecto                                Número    Chi-cuadrado    Chi-cuadrado
 Paso    Introducido        Eliminado          DF            en    de puntuación        de Wald    Pr > ChiSq

    1    num_users_hashtag_tr                   1             1        1034.6801                     <.0001
    2    num_friends_usrhas_t                   1             2         675.0815                     <.0001
    3    num_hashtags_user_tr                   1             3         344.8586                     <.0001
    4    hashtag_tram                           1             4         334.4076                     <.0001
    5    media_hora_tram                        1             5         236.0060                     <.0001
    6    num_dias_hashtag_tra                   1             6         125.9172                     <.0001
    7    pct_rep_friends_hist                   1             7          68.4473                     <.0001
    8    avg_friends_user_tra                   1             8          35.4392                     <.0001
    9    avg_friends_hashtag_                   1             9          17.3099                     <.0001
   10    pct_replies_hashtag_                   1            10          26.4663                     <.0001
   11    pct_rep_in_friends_u                   1            11          21.0483                     <.0001
   12    avg_foll_in_friends_                   1            12          16.7218                     <.0001
   13    num_men_friends_hist                   1            13          13.8999                     0.0002
   14    num_friends_histusrh                   1            14          38.8866                     <.0001
   15    avg_fr_friends_histu                   1            15           8.5800                     0.0034
   16    avg_twe_in_friends_u                   1            16           4.4384                     0.0351
   17    ind_location_tram                      1            17           1.7242                     0.1892
   18    pct_replies_user_tra                   1            18           1.2187                     0.2696
   19    ind_geo_enabled_tram                   1            19           1.2033                     0.2727
```

We can observe that the two new defined variables have been entered the model so both variables are significant, providing new information to our model.

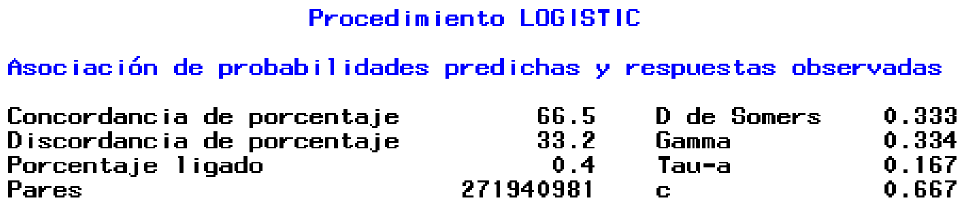
The p-values are not very large so that the variables are significant.

Associated estimators can be seen in the next image:



```
                 Análisis del estimador de máxima verosimilitud

                                                 Error    Chi-cuadrado
Parámetro                DF    Estimador      estándar         de Wald    Pr > ChiSq

Intercept                 1       0.1878        0.1324          2.0121        0.1560
avg_foll_in_friends_      1       0.1450        0.0347         17.4912        <.0001
avg_fr_friends_histu      1       0.0560        0.0191          8.5539        0.0034
num_friends_histusrh      1       0.1436        0.0341         17.7141        <.0001
num_users_hashtag_tr      1      -0.3727       0.00939       1573.8360        <.0001
pct_rep_friends_hist      1      -0.2514        0.0488         26.5835        <.0001
pct_rep_in_friends_u      1       0.3603        0.0738         23.8647        <.0001
pct_replies_hashtag_      1      -0.0545       0.00931         34.2939        <.0001
pct_replies_user_tra      1      -0.0587        0.0531          1.2224        0.2689
avg_twe_in_friends_u      1      -0.0721        0.0343          4.4205        0.0355
ind_geo_enabled_tram      1      -0.0268        0.0244          1.2033        0.2727
ind_location_tram         1      -0.1303        0.1046          1.5502        0.2131
num_dias_hashtag_tra      1       0.3753        0.0310        146.1728        <.0001
avg_friends_hashtag_      1       0.0433       0.00864         25.1231        <.0001
avg_friends_user_tra      1      -0.0479       0.00942         25.8897        <.0001
num_hashtags_user_tr      1      -0.2628        0.0154        289.4659        <.0001
num_friends_usrhas_t      1       0.4130        0.0327        159.7314        <.0001
num_men_friends_hist      1      -0.2128        0.0280         57.6509        <.0001
hashtag_tram              1       0.2236        0.0114        385.8623        <.0001
media_hora_tram           1       0.2940        0.0212        191.8449        <.0001
```

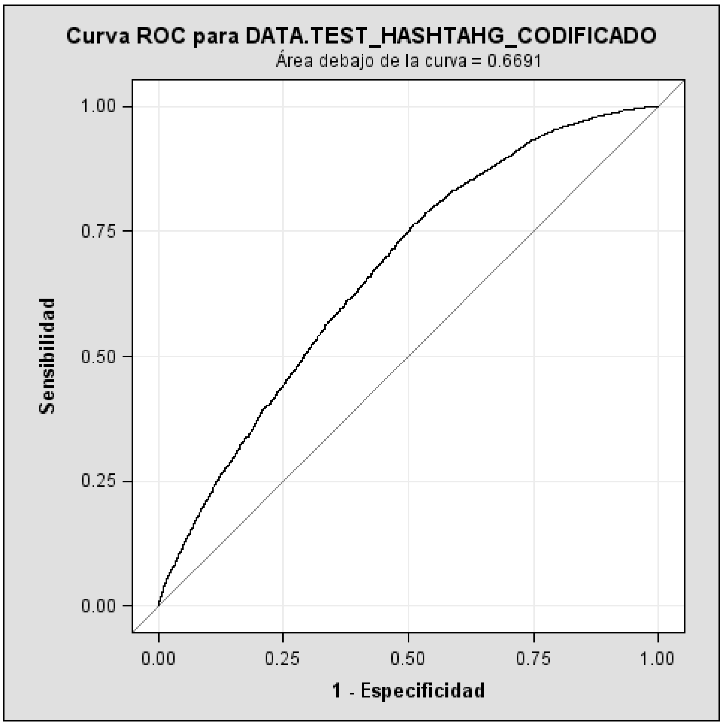
Finally, the four measures which indicate the association of predicted probabilities and observed responses are given by:



```
                     Procedimiento LOGISTIC

 Asociación de probabilidades predichas y respuestas observadas

 Concordancia de porcentaje          66.5    D de Somers      0.333
 Discordancia de porcentaje          33.2    Gamma            0.334
 Porcentaje ligado                    0.4    Tau-a            0.167
 Pares                          271940981    c                0.667
```

We note that these measures are higher than in the previous case, so we conclude that the fact of introducing two new variables improves the model significantly.

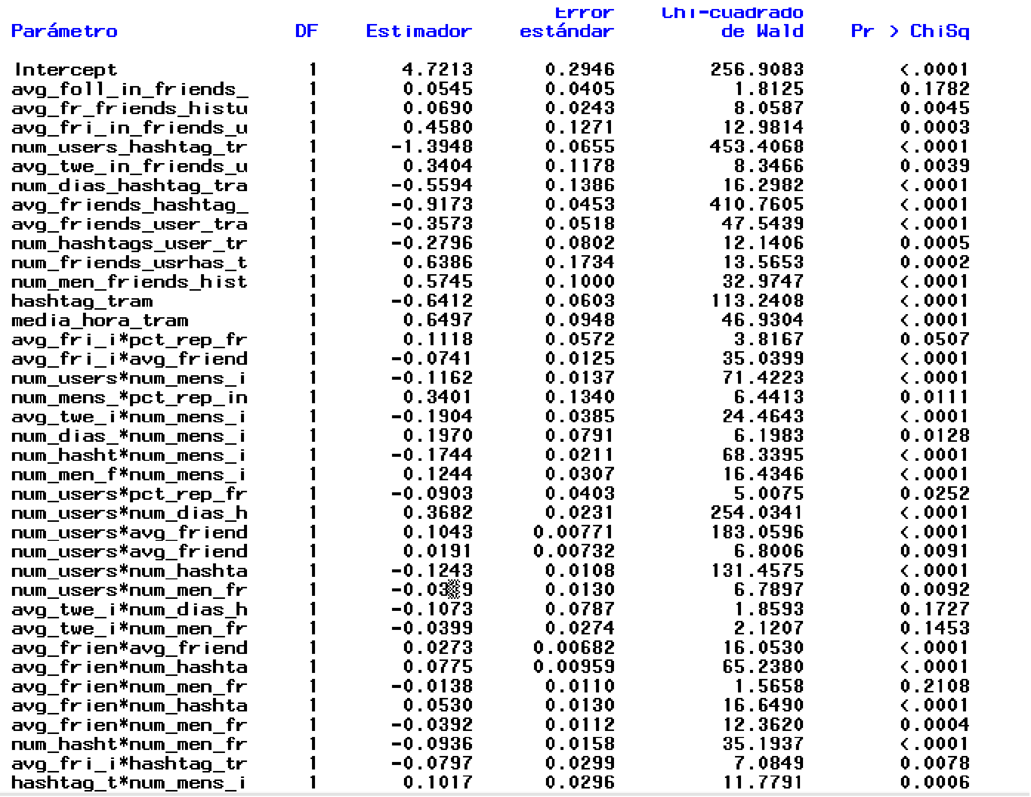
Furthermore enclosed area under the ROC curve for the test table is given by:



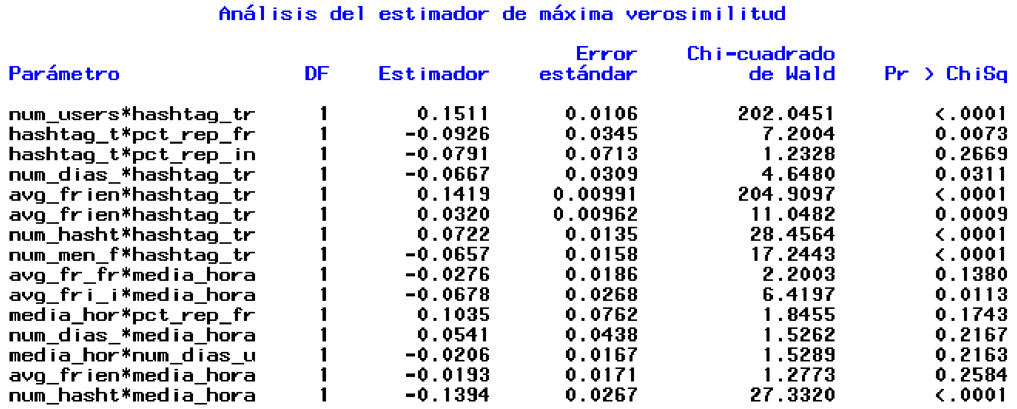
INPUTTING INITIALS VARIABLES, NEW VARIABLES AND ITERATIONS

In this model we have introduced, in addition to all the proposed variables to enter into the model, all possible combinations two by two and a stepwise technique is responsible to discriminate the non-significant variables.

With the same procedure we can observe that the variables that enter into the model are the followings:



Parámetro	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
Intercept	1	4.7213	0.2946	256.9083	<.0001
avg_foll_in_friends_	1	0.0545	0.0405	1.8125	0.1782
avg_fr_friends_histu	1	0.0690	0.0243	8.0587	0.0045
avg_fri_in_friends_u	1	0.4580	0.1271	12.9814	0.0003
num_users_hashtag_tr	1	-1.3948	0.0655	453.4068	<.0001
avg_twe_in_friends_u	1	0.3404	0.1178	8.3466	0.0039
num_dias_hashtag_tra	1	-0.5594	0.1386	16.2982	<.0001
avg_friends_hashtag_	1	-0.9173	0.0453	410.7605	<.0001
avg_friends_user_tra	1	-0.3573	0.0518	47.5439	<.0001
num_hashtags_user_tr	1	-0.2796	0.0802	12.1406	0.0005
num_friends_usrhas_t	1	0.6386	0.1734	13.5653	0.0002
num_men_friends_hist	1	0.5745	0.1000	32.9747	<.0001
hashtag_tram	1	-0.6412	0.0603	113.2408	<.0001
media_hora_tram	1	0.6497	0.0948	46.9304	<.0001
avg_fri_i*pct_rep_fr	1	0.1118	0.0572	3.8167	0.0507
avg_fri_i*avg_friend	1	-0.0741	0.0125	35.0399	<.0001
num_users*num_mens_i	1	-0.1162	0.0137	71.4223	<.0001
num_mens_*pct_rep_in	1	0.3401	0.1340	6.4413	0.0111
avg_twe_i*num_mens_i	1	-0.1904	0.0385	24.4643	<.0001
num_dias_*num_mens_i	1	0.1970	0.0791	6.1983	0.0128
num_hasht*num_mens_i	1	-0.1744	0.0211	68.3395	<.0001
num_men_f*num_mens_i	1	0.1244	0.0307	16.4346	<.0001
num_users*pct_rep_fr	1	-0.0903	0.0403	5.0075	0.0252
num_users*num_dias_h	1	0.3682	0.0231	254.0341	<.0001
num_users*avg_friend	1	0.1043	0.00771	183.0596	<.0001
num_users*avg_friend	1	0.0191	0.00732	6.8006	0.0091
num_users*num_hashta	1	-0.1243	0.0108	131.4575	<.0001
num_users*num_men_fr	1	-0.0339	0.0130	6.7897	0.0092
avg_twe_i*num_dias_h	1	-0.1073	0.0787	1.8593	0.1727
avg_twe_i*num_men_fr	1	-0.0399	0.0274	2.1207	0.1453
avg_frien*avg_friend	1	0.0273	0.00682	16.0530	<.0001
avg_frien*num_hashta	1	0.0775	0.00959	65.2380	<.0001
avg_frien*num_men_fr	1	-0.0138	0.0110	1.5658	0.2108
avg_frien*num_hashta	1	0.0530	0.0130	16.6490	<.0001
avg_frien*num_men_fr	1	-0.0392	0.0112	12.3620	0.0004
num_hasht*num_men_fr	1	-0.0936	0.0158	35.1937	<.0001
avg_fri_i*hashtag_tr	1	-0.0797	0.0299	7.0849	0.0078
hashtag_t*num_mens_i	1	0.1017	0.0296	11.7791	0.0006

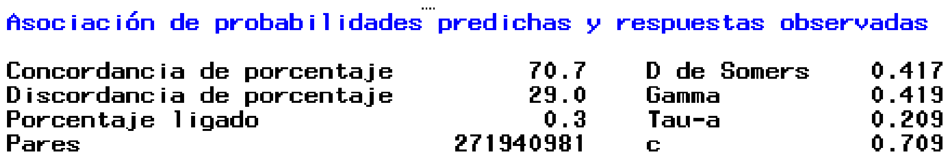


Análisis del estimador de máxima verosimilitud

Parámetro	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
num_users*hashtag_tr	1	0.1511	0.0106	202.0451	<.0001
hashtag_t*pct_rep_fr	1	-0.0926	0.0345	7.2004	0.0073
hashtag_t*pct_rep_in	1	-0.0791	0.0713	1.2328	0.2669
num_dias_*hashtag_tr	1	-0.0667	0.0309	4.6480	0.0311
avg_frien*hashtag_tr	1	0.1419	0.00991	204.9097	<.0001
avg_frien*hashtag_tr	1	0.0320	0.00962	11.0482	0.0009
num_hasht*hashtag_tr	1	0.0722	0.0135	28.4564	<.0001
num_men_f*hashtag_tr	1	-0.0657	0.0158	17.2443	<.0001
avg_fr_fr*media_hora	1	-0.0276	0.0186	2.2003	0.1380
avg_fri_i*media_hora	1	-0.0678	0.0268	6.4197	0.0113
media_hor*pct_rep_fr	1	0.1035	0.0762	1.8455	0.1743
num_dias_*media_hora	1	0.0541	0.0438	1.5262	0.2167
media_hor*num_dias_u	1	-0.0206	0.0167	1.5289	0.2163
avg_frien*media_hora	1	-0.0193	0.0171	1.2773	0.2584
num_hasht*media_hora	1	-0.1394	0.0267	27.3320	<.0001

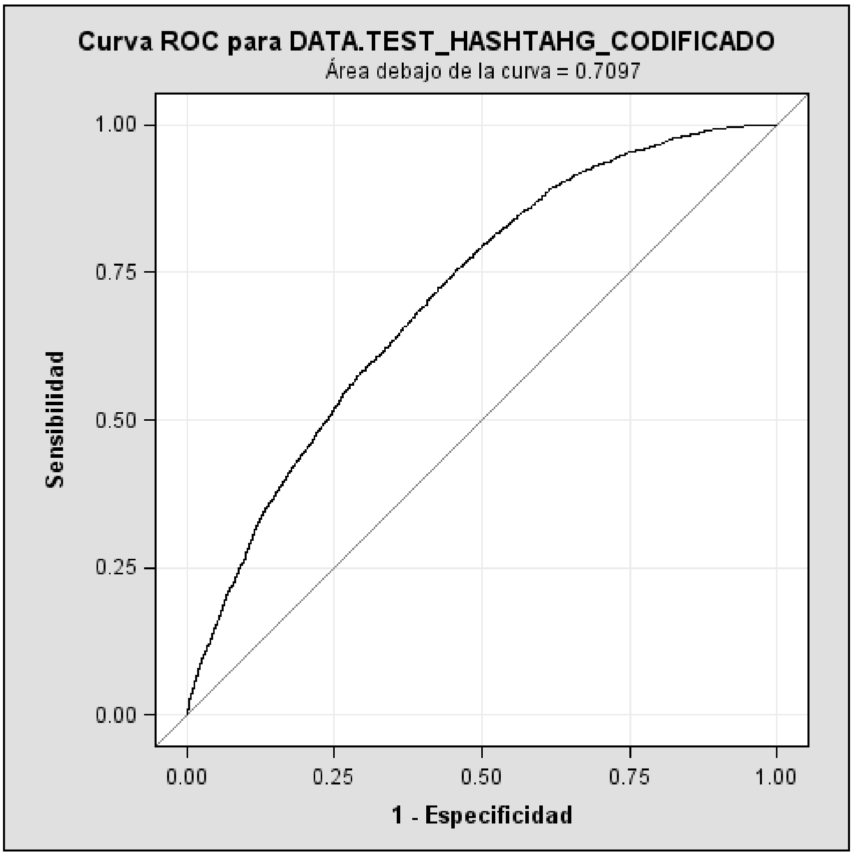
It’s important to say that there are a large number of iterations, so the combinations are significant to our model.

Seeing the association of predicted probabilities and observed responses, this measures have grown considerably by introducing combinations, so we conclude that this model is the best of the three studied.



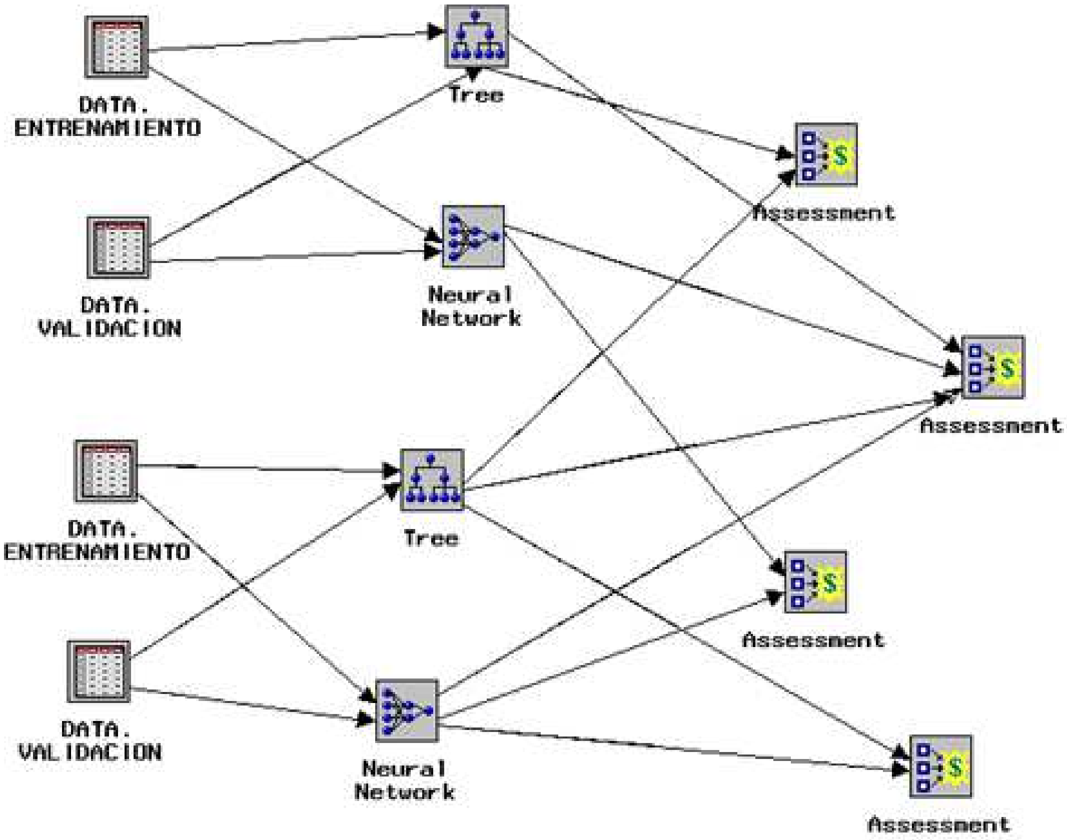
```
Asociación de probabilidades predichas y respuestas observadas

Concordancia de porcentaje          70.7    D de Somers    0.417
Discordancia de porcentaje          29.0    Gamma          0.419
Porcentaje ligado                    0.3    Tau-a          0.209
Pares                          271940981    c              0.709
```

In this case, this is the Roc Curve that we have obtained:



We have used **ENTERPRISE MINER** to create the decision tree models and the neuronal network models.

This is the structure of the diagram made in **ENTERPISE MINER**:



The nodes: **DATA.ENTRENAMIENTO / DATA.VALIDACION** refer to training and validation tables for the models with the original variables and for the models with the original variables and the new two variables (*hashtag content* and *user's activity time*), respectively.

The nodes: **TREE / NEURONAL NETWORK** refer to the decision tree models and the neuronal network models, respectively.

The nodes of **ASSESSMENT** have been used to compare more easily the different models we have made.

DECISION TREE MODEL

For the second model, we have predicted the binary target via a **DECISION TREE**.

Decision tree models are perhaps the models easiest to use and understand. They have the advantage that you can see all the possible consequences of a decision.

In our particular case, it may be very appropriate to use a binary decision tree model because is very easy to apply it to a new hashtag and a new user to know if it will be retweeted. Just follow the branches of the tree to reach the appropriate sheet for the answer YER or NO.

- **TRAINING TABLE**: used to fit the initial tree

- **VALIDATION TABLE**: used by default for assessment of the tree.

- **TEST TABLE**: additional "hold out" data set than we can use for model assessment.

DECISION TREE FOR THE MODEL WITH THE ORIGINAL VARIABLES

For this model we have considerated the following variables:

- **MODEL ROLE TARGET**: *Target*
- **MODEL ROLE ID**: *User_ID* and *Hashtag*
- **MODEL ROLE INPUT**: The **original variables** but **categorized**

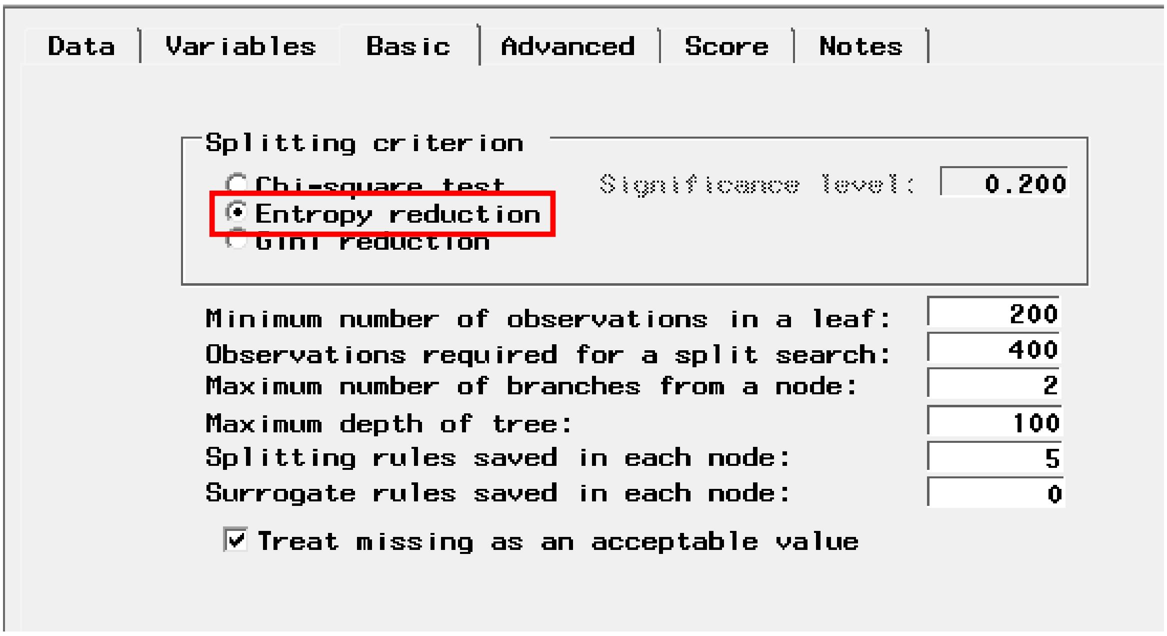
The measurement of the target variable is BINARY, because it can take only two different values (1 or 0).

The measurement of the categorized variables *antiquity_months, avg_fol_friends_histusrhas, avg_foll_in_friends_usrhas, avg_tw_friends_histusrhas, avg_fr_friends_histusrhas, avg_fri_in_friends_usrhas, num_friends_histusrhas, num_mens_in_friends_usrhas, num_users_hashtag, pct_replies_hashtag, avg_twe_in_friends_usrhas, num_dias_user, avg_friends_hashtag, avg_friends_user, num_hashtags_user, num_men_friends_histusrhas, num_friends_usrhas* is NOMINAL, because they can take between 3 and 10 different values.

The measurement of the categorized variables *pct_rep_friends_histusrhas, pct_rep_in_friends_usrhas, pct_replies_user, ind_description, ind_geo_enabled, ind_location, ind_url_in_description, num_dias_hashtag* is BINARY, because they can take only two different values.

The Splitting Criterion we get the best results is **ENTROPY REDUCTION** (the reduction in the entropy measure of node impurity).

The rest of the options in the decision tree model (maximum depth of tree, ...) are shown in the following image:



Splitting criterion

○ Chi-square test

● Entropy reduction

○ Gini reduction

Significance level: 0.200

Minimum number of observations in a leaf: 200

Observations required for a split search: 400

Maximum number of branches from a node: 2

Maximum depth of tree: 100

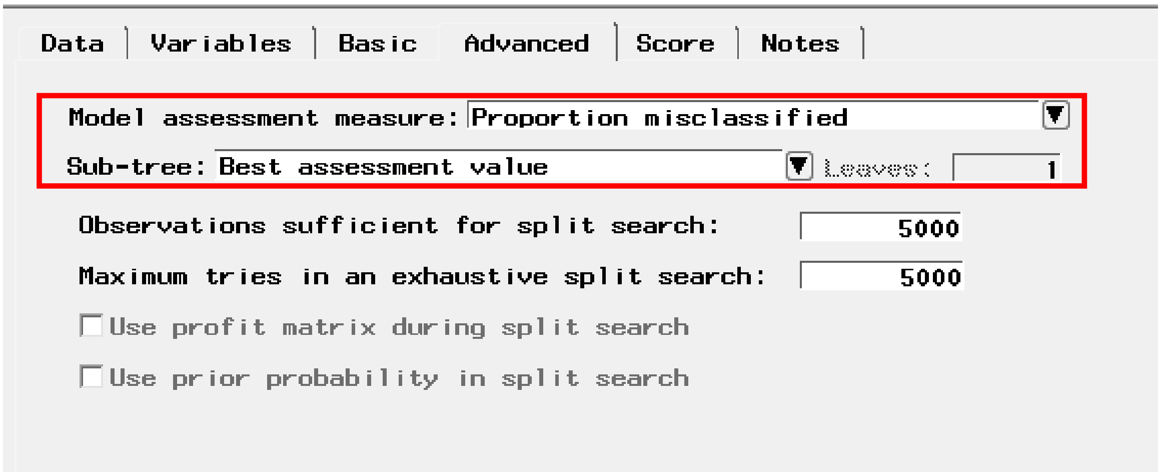
Splitting rules saved in each node: 5

Surrogate rules saved in each node: 0

☑ Treat missing as an acceptable value

The model assessment measure we have chosen is **PROPORTION MISSCLASIFIED** (it's the default option; chooses the tree with the smallest misclassification rate).

The Tree node of Enterprise Miner evaluates all possible sub-trees of the constructed tree, and reports the results for the best sub-tree for each possible number of leaves. The node supports different sub-tree methods, and we have chosen **BEST ASSESSMENT VALUE** (the sub-tree which produces the best results according to the selected Model assessment measure chosen. Validation data is used if it is available)**,** how you can see in the image below:



Model assessment measure: Proportion misclassified

Sub-tree: Best assessment value

Leaves: 1

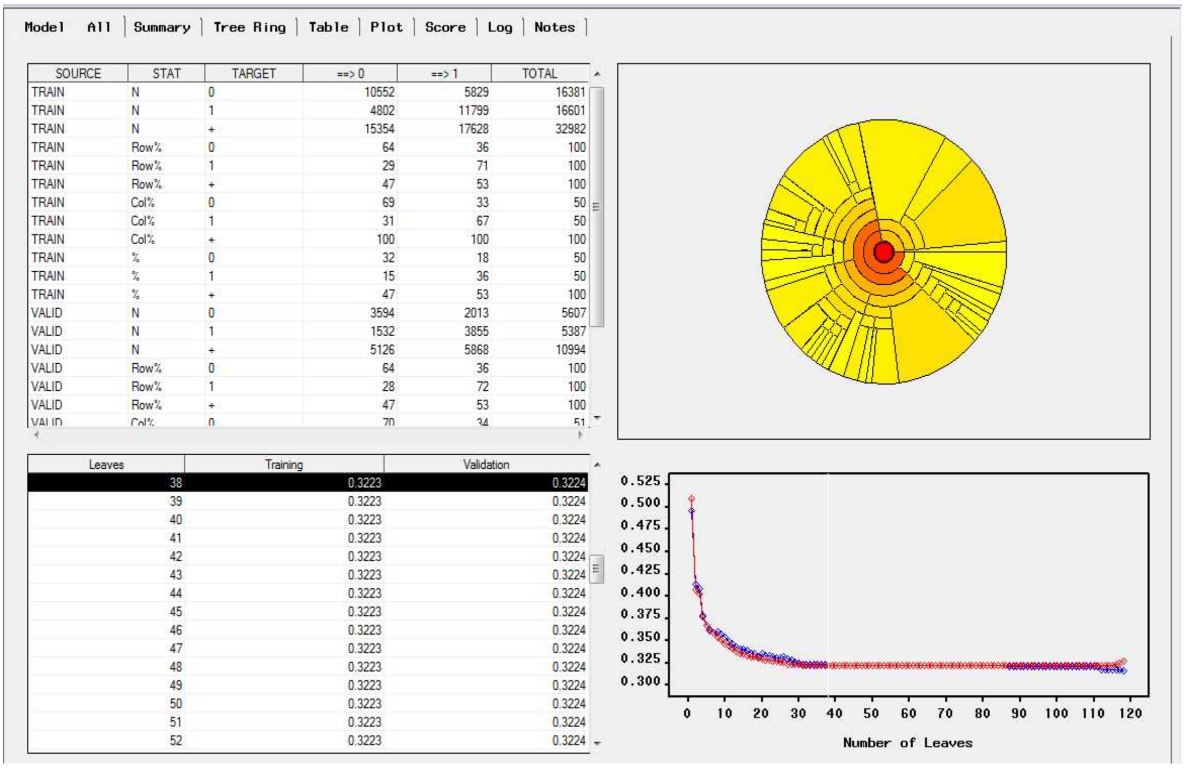
Observations sufficient for split search: 5000

Maximum tries in an exhaustive split search: 5000

□ Use profit matrix during split search

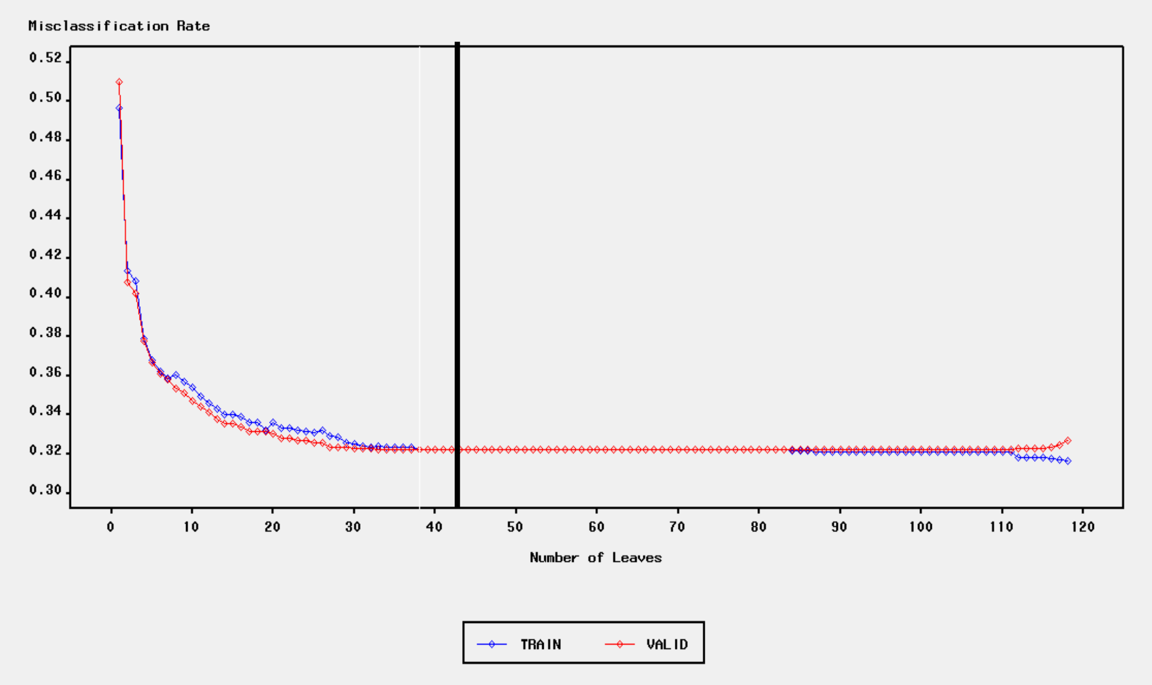
□ Use prior probability in split search

The results obtained can be summarized in the following image:



SOURCE	STAT	TARGET	==> 0	==> 1	TOTAL
TRAIN	N	0	10552	5829	16381
TRAIN	N	1	4802	11799	16601
TRAIN	N	+	15354	17628	32982
TRAIN	Row%	0	64	36	100
TRAIN	Row%	1	29	71	100
TRAIN	Row%	+	47	53	100
TRAIN	Col%	0	69	33	50
TRAIN	Col%	1	31	67	50
TRAIN	Col%	+	100	100	100
TRAIN	%	0	32	18	50
TRAIN	%	1	15	36	50
TRAIN	%	+	47	53	100
VALID	N	0	3594	2013	5607
VALID	N	1	1532	3855	5387
VALID	N	+	5126	5868	10994
VALID	Row%	0	64	36	100
VALID	Row%	1	28	72	100
VALID	Row%	+	47	53	100
VALID	Col%	0	70	34	51

Leaves	Training	Validation
38	0.3223	0.3224
39	0.3223	0.3224
40	0.3223	0.3224
41	0.3223	0.3224
42	0.3223	0.3224
43	0.3223	0.3224
44	0.3223	0.3224
45	0.3223	0.3224
46	0.3223	0.3224
47	0.3223	0.3224
48	0.3223	0.3224
49	0.3223	0.3224
50	0.3223	0.3224
51	0.3223	0.3224
52	0.3223	0.3224



You can see that the number of sheets that have a higher number of cases correctly classified in the validation sample is 38. The lower the **MISCLASSIFICATION RATE**, the higher the number of cases correctly classified.

As the decision tree for the model with the original variables and the decision tree for the model with the original variables and the two new variables are similar, we will present the tree for the model with the original variables and the two new variables because it will have better results.

In order to compare the two decision tree models we will apply and the different statistical models we will do, we obtain the value of the area under the ROC curve using the following SAS macro:

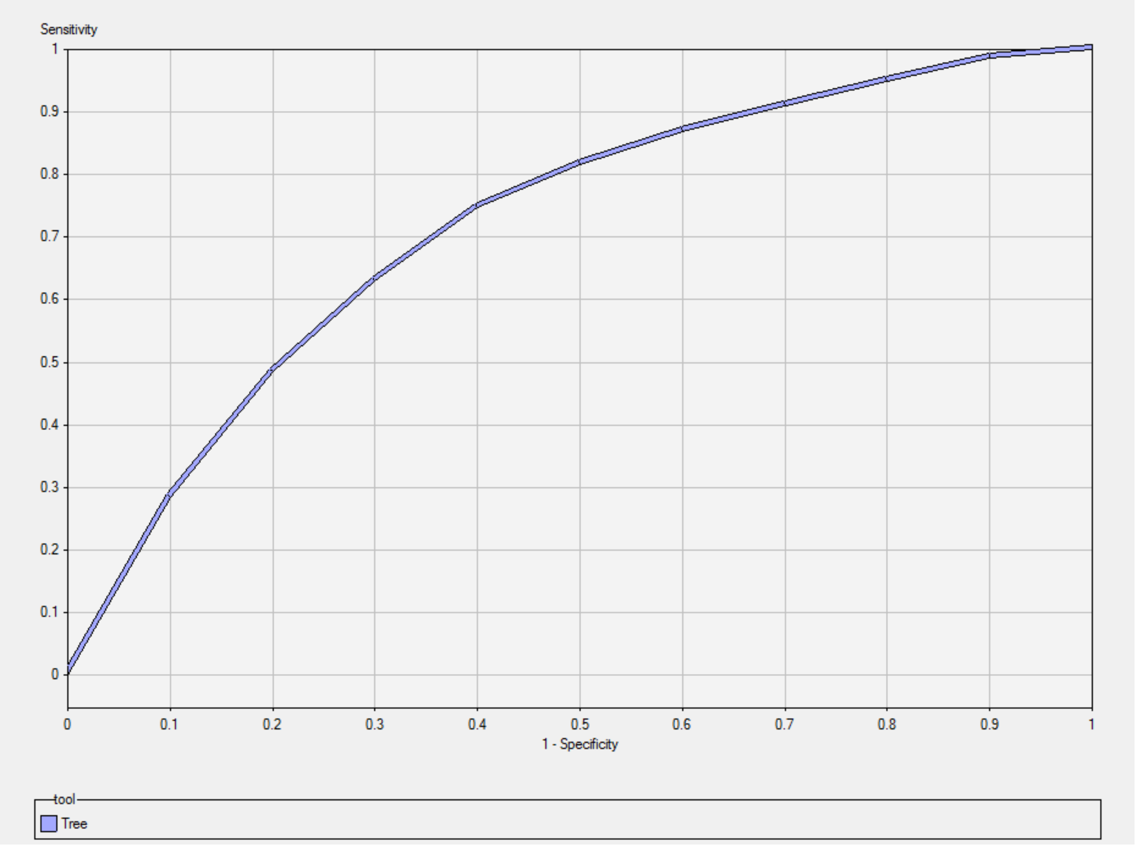
```
%macro obtenerRoc(tabla=, target=, predict=);
 %global area_roc;
 proc sort data=&tabla(keep=&target &predict) out=roc_temp_01;
  by descending &predict;
 run;
 proc sql noprint;
  select count(*) into :n
  from roc_temp_01;
  select count(*) into :n_positivos
  from roc_temp_01
  where &target = 1;
 quit;
```

```
  %let n_negativos = %eval(&n - &n_positivos);
  data roc_temp_02;
    set roc_temp_01 end=fin;
    retain sen_roc fp_roc 0;
    retain fp_roc_ant 0;
    if &target=1 then sen_roc=sen_roc+(1/&n_positivos);
    if &target ne 1 then do;
      fp_roc=fp_roc+(1/&n_negativos);
    end;
    output;
    if _N_>1 then do;
      fp_roc_ant=fp_roc;
    end;
  run;
  data roc_temp_03(drop=area_roc);
    set roc_temp_02 end=fin;
    posicion=_N_;
    retain area_roc 0;
    area_roc=area_roc+((fp_roc-fp_roc_ant)*sen_roc);
    if fin then do;
      call symput('area_roc',put(area_roc,8.5));
    end;
  run;
  %put area_roc= &area_roc;
%mend;
```

And this is the value we have obtained for the test table:

AREA UNDER THE ROC CURVE: 0.7

This is the ROC CURVE (we are modeling the probability of being 1):



DECISION TREE FOR THE MODEL WITH THE ORIGINAL VARIABLES AND THE TWO NEW VARIABLES

For this model we have considerated the following variables:

- **MODEL ROLE TARGET**: *Target*
- **MODEL ROLE ID**: *User_ID*
- **MODEL ROLE INPUT**: The **original variables** and the two new variables (*hashtag content* and *user's activity time*) but **categorized**

The measurement of the target variable is BINARY, because it can take only two different values (1 or 0).

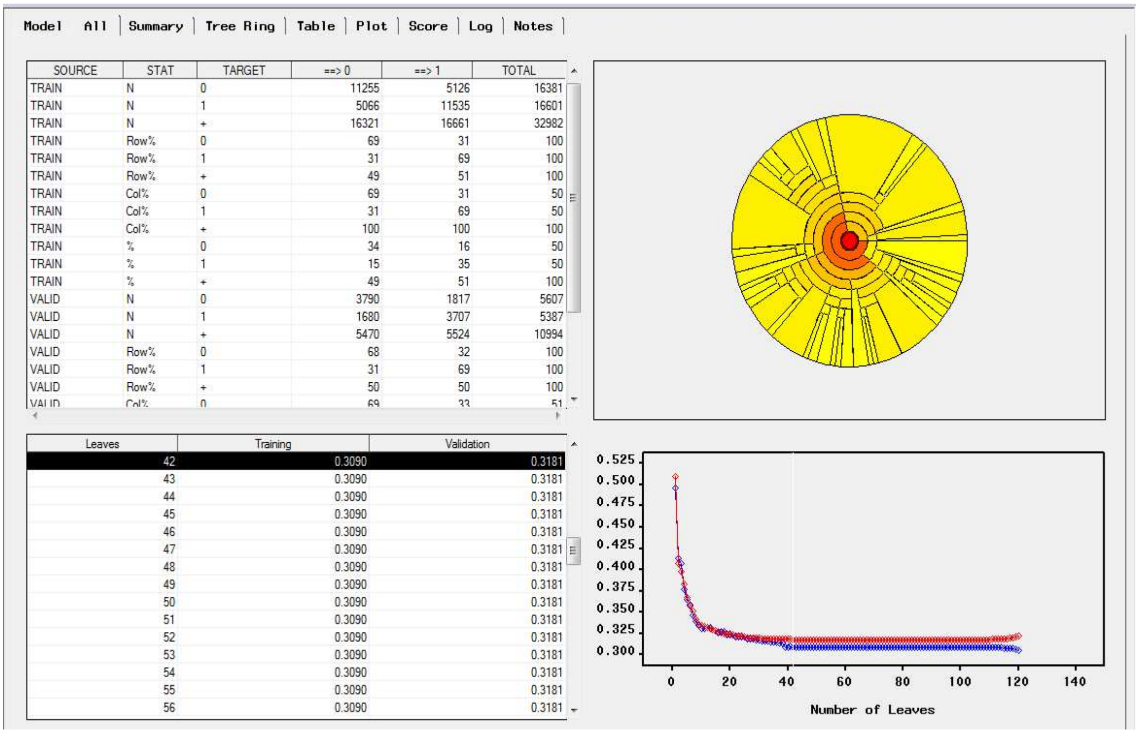
The measurement of the categorized variables *antiquity_months, avg_fol_friends_histusrhas, avg_foll_in_friends_usrhas, avg_tw_friends_histusrhas, avg_fr_friends_histusrhas, avg_fri_in_friends_usrhas, num_friends_histusrhas, num_mens_in_friends_usrhas, num_users_hashtag, pct_replies_hashtag, avg_twe_in_friends_usrhas, num_dias_user, avg_friends_hashtag, avg_friends_user,* *num_hashtags_user, num_men_friends_histusrhas, num_friends_usrhas, hashtag, media_hora* is NOMINAL, because they can take between 3 and 10 different values.

The measurement of the categorized variables *pct_rep_friends_histusrhas, pct_rep_in_friends_usrhas, pct_replies_user, ind_description, ind_geo_enabled, ind_location, ind_url_in_description, num_dias_hashtag* is BINARY, because they can take only two different values.

In order to compare the two decision trees models we have applied, we have considered in this case the same options for the decision tree with the original variables, that is to say:

- **SPLITTING CRITERION**: Entropy reduction
- **MODEL ASSESSMENT MEASURE**: Proportion misclassified
- **SUB-TREE**: Best assessment value

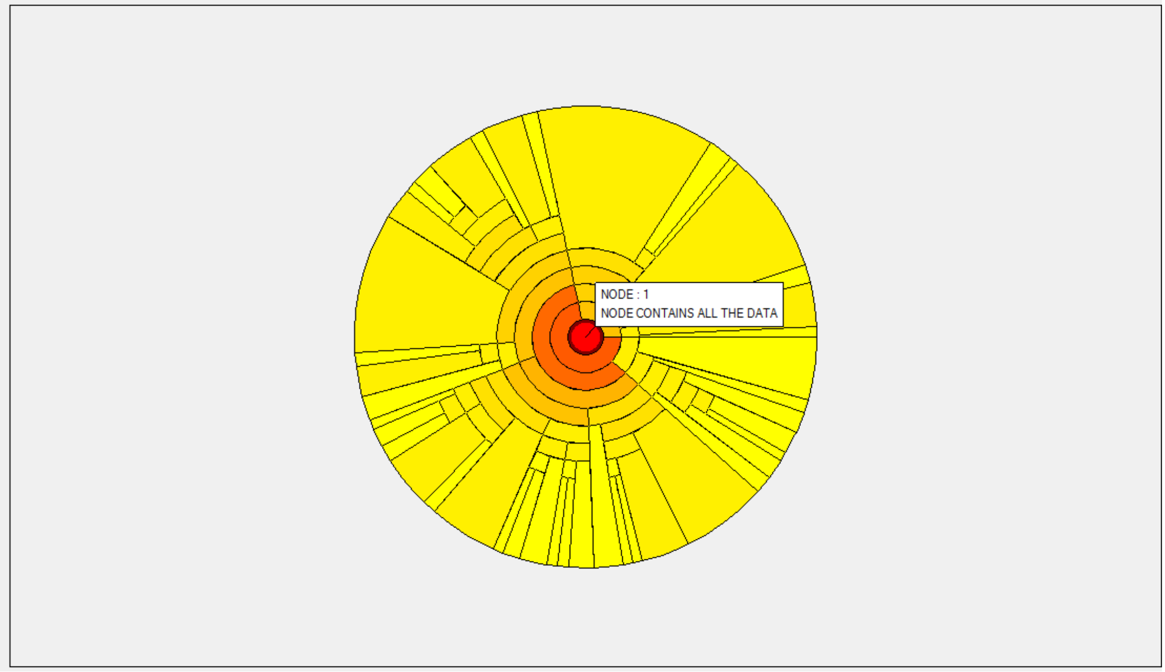
The results obtained can be summarized in the following image:



SOURCE	STAT	TARGET	==> 0	==> 1	TOTAL
TRAIN	N	0	11255	5126	16381
TRAIN	N	1	5066	11535	16601
TRAIN	N	+	16321	16661	32982
TRAIN	Row%	0	69	31	100
TRAIN	Row%	1	31	69	100
TRAIN	Row%	+	49	51	100
TRAIN	Col%	0	69	31	50
TRAIN	Col%	1	31	69	50
TRAIN	Col%	+	100	100	100
TRAIN	%	0	34	16	50
TRAIN	%	1	15	35	50
TRAIN	%	+	49	51	100
VALID	N	0	3790	1817	5607
VALID	N	1	1680	3707	5387
VALID	N	+	5470	5524	10994
VALID	Row%	0	68	32	100
VALID	Row%	1	31	69	100
VALID	Row%	+	50	50	100
VALID	Col%	0	69	33	51

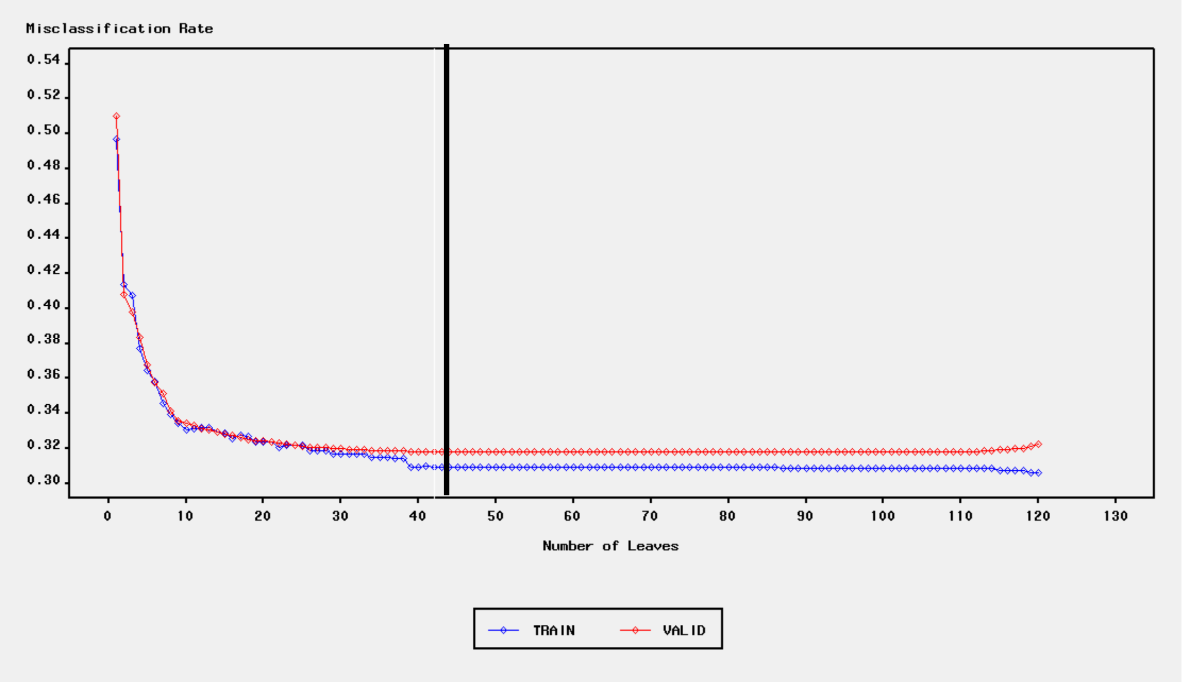
Leaves	Training	Validation
42	0.3090	0.3181
43	0.3090	0.3181
44	0.3090	0.3181
45	0.3090	0.3181
46	0.3090	0.3181
47	0.3090	0.3181
48	0.3090	0.3181
49	0.3090	0.3181
50	0.3090	0.3181
51	0.3090	0.3181
52	0.3090	0.3181
53	0.3090	0.3181
54	0.3090	0.3181
55	0.3090	0.3181
56	0.3090	0.3181

The **TREE RING** Navigator is a graphical display of tree complexity, split balance, and discriminatory power. With the Tree Ring tab we can see an enlarged version of the Tree Ring. The center region represents the entire data set (the root node of the tree).

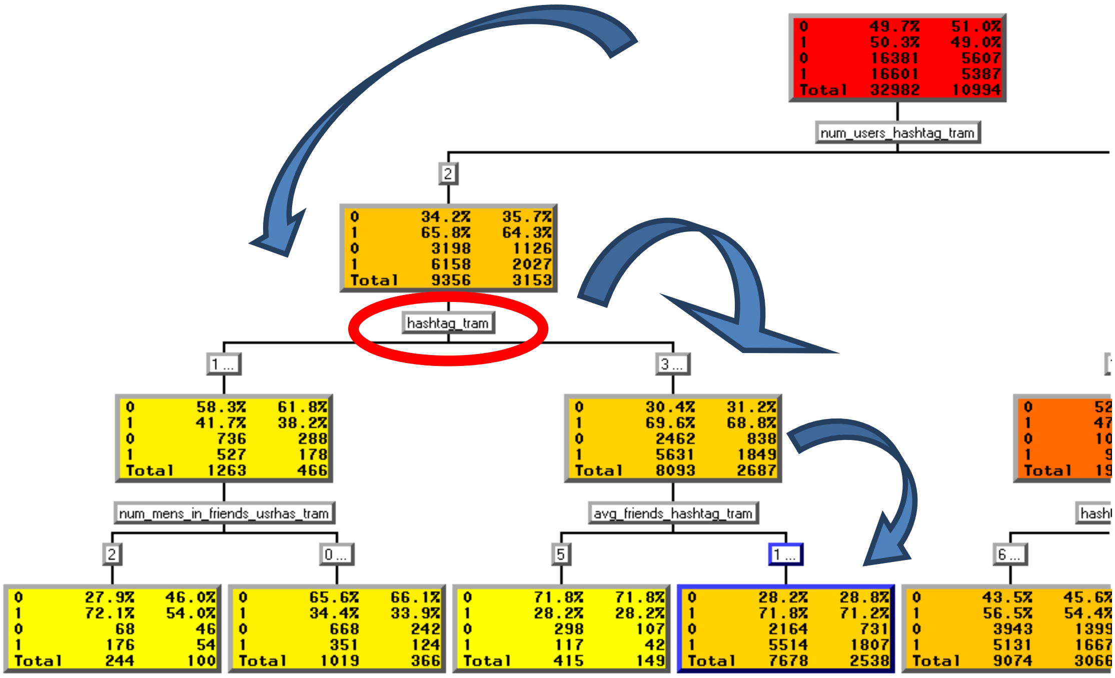


The ring surrounding the center represents the initial split. Successive rings represent successive stages of splits. The sizes of displayed segments in one ring are proportional to the number of training observations in the segments. For nominal targets, the color hues in the Tree Ring segments correspond to the assessment values. The default Tree Ring is segmented by different shades of orange/yellow. The legend window displays the assessment values for each hue. Nodes that have similar assessment values have a similar color.

We can see in the chart below that the number of sheets that have a higher number of cases correctly classified in the validation sample is 41:



This is a fragment of the decision tree made:



We observe that, originally, the first column of the first sheet represents the training table and it has approximately 50.3% of ones and 48.7% of zeros, while the second column of the first sheet, which represents the validation table, has 49% of ones and 51% of zeros.

The variable that best describes our data is *num_user_hashtag_tram*, which is the variable that has a higher number of ones in the target variable.

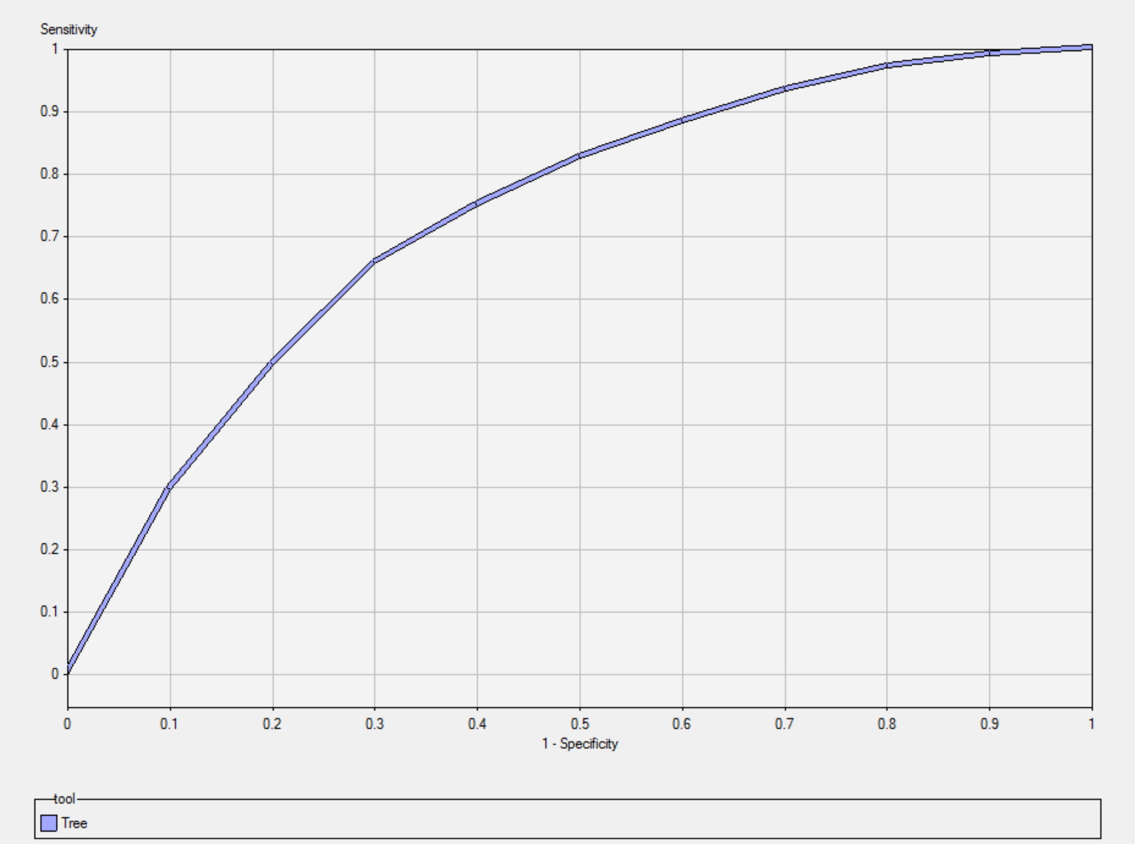
From there, one of the two variables in which the tree branches is just one of the new variables included in the model, *hashtag_tram*, picking in the corresponding sheet approximately 30% of the original data. This could be an indication that the meaning of the hashtag can be an important issue to explore in predicting the binary target.

It is important to remember that the variable *hashtag_tram* has been coded depending on the topic of the hashtag in sis categories, where the category of politics represents near 50% of all the hashtags.

Finally, following the structure of the tree, we can reach a node that contains approximately 24% of the data and having an approximately 71.8% and 28.2% of zeros, representing a high percentage of ones.

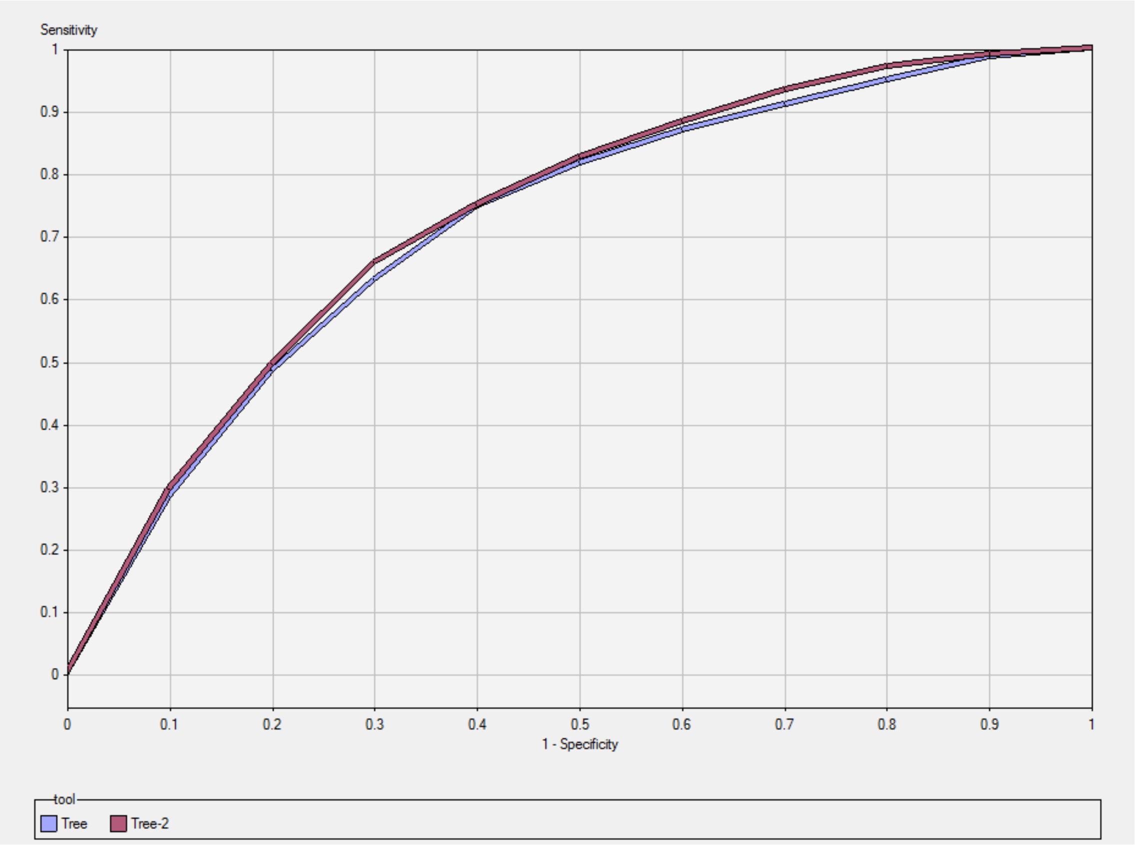
Using the SAS macro we have described above for the test table to calculate the area under the ROC curve:

AREA UNDER THE ROC CURVE: 0.72



We can observe that we significantly improve because the area under the ROC curve increases in 0.02 with the inclusion of the two new variables in the model (0.7 to 0.72).

If we plot the ROC curve simultaneously for both decision trees with the assistance of the **ASSESSMENT NODE** of Enterprise Miner, we can observe more easily that we improve with the inclusion of the two new variables (*hashtag content* and *user's activity time*) in the model:



- Tree refers to the decision tree for the model with the original variables
- Tree-2 refers to the decision tree for the model with the original variables and the two new variables

NEURAL NETWORKS

WHAT IS A NEURAL NETWORK?

The objective that gives rise to neural networks is to build a model that is able to reproduce the learning method of the human brain

Basically a neural network is defined by a graph consisting of:

- Nodes called **NEURONS** which may be 3 types:

  - Input neurons: Through them to enter the values of the explaining variables

  - Output neurons: Through them to enter the target variable values

  - Intermediate or hidden neurons: It is optional and propagates the information from input neurons to the output neurons

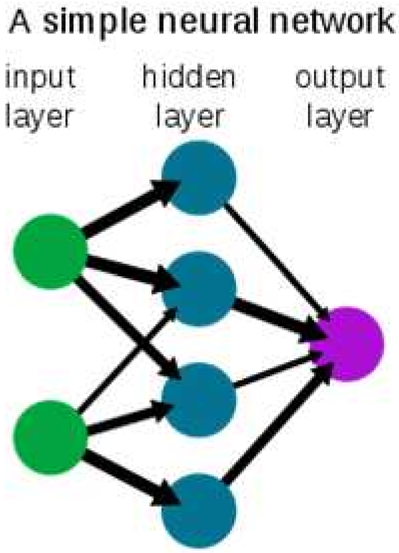
- Arcs called **CONNECTIONS** that are associated with an actual value, the weight which determines the influence of a neuron in another neuron.

Some networks are structured in layers whose definition is referenced to the type composed of neurons. A layered structured network must contain at least:

- An input layer: Associated to input neurons

- An output layer: Associated to output neurons

Optionally, a neural network can be associated with one or more intermediate or hidden layers as we can see in the picture



In the same destination neuron may be several connections in which case weights associated with them are combined and processed using the function entry or spread. The value provided by the above function is handled by a function called the activation function that generates the output ending. A popular function spread is the perceptron that combine weights with the explaining variables

$$f(x_1, x_2, ..., x_n) = \sum_{i=1}^{n} w_i x_i$$

The objective is to search the weight vector $\overline{w}$ yielding the best linear hyperplane that separates areas.

One of the typical problems in neural network models is determine the network architecture, in particular as regards the number of hidden layers. A network without hidden layers acts only as linear classifier. A network with one hidden layer discriminate classes / regions related ("no holes"). In theory, a network with two hidden layers can identify regions of any type.

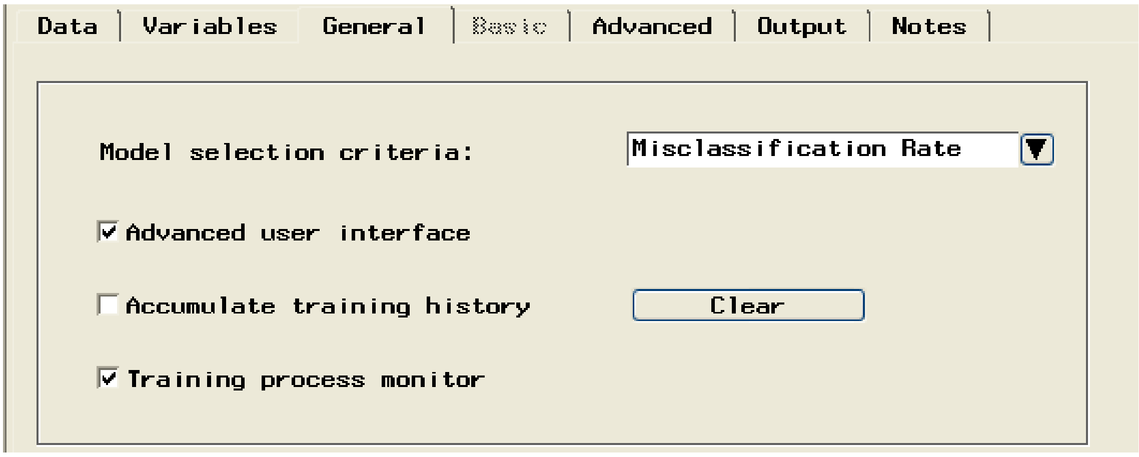
Another typical problem in the process of configuring the architecture of the network is to determine the number of neurons per layer. This number usually responds to different tests performed by trial and error that contribute to the term "black box" that is given to this type of models.

In any case, it is difficult to establish when an input or hidden neuron must be eliminated from the model because the weights are difficult to interpret. Theoretically, if all the weights of a neuron are "close" to zero, the neuron does not influence the model, but there is no measurement (like p-valor) that contrasts this significance.

RESULTS

As it was the case with the decision tree we will have a **TRAINING TABLE, a VALIDATION TABLE** and a **TEST TABLE**

The options used in our neural network model are as follows.



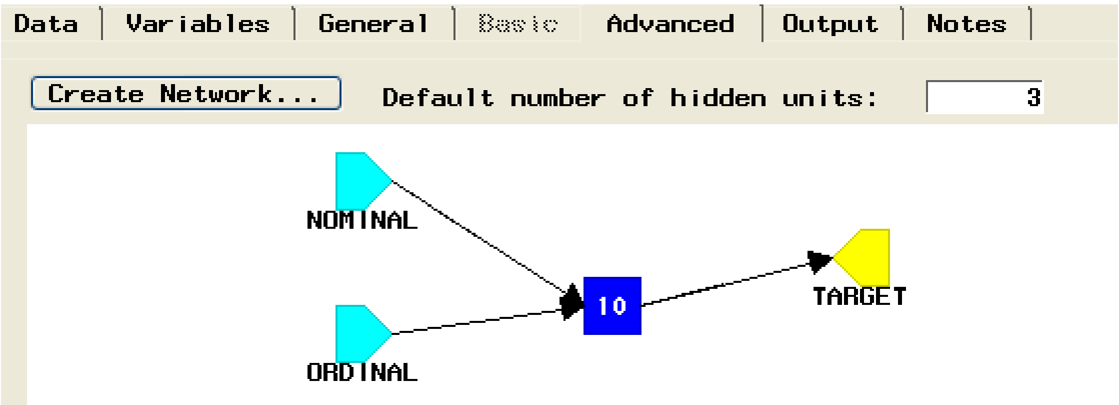
Model selection criteria: Misclassification Rate

☑ Advanced user interface

□ Accumulate training history

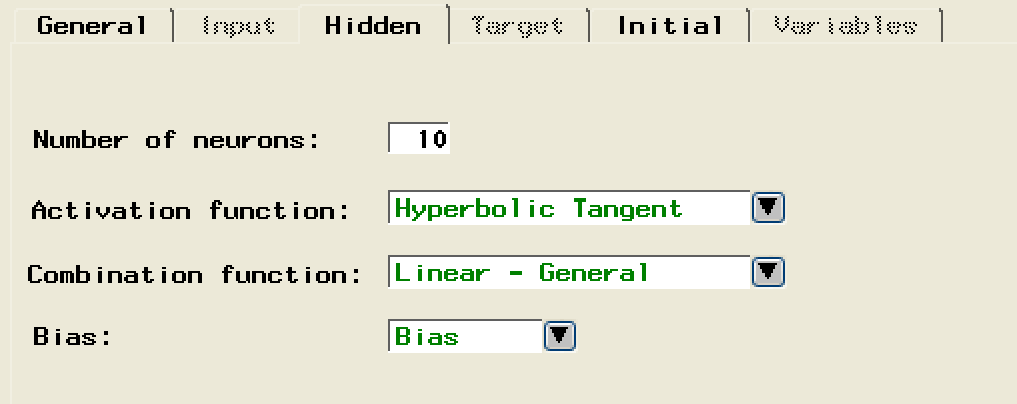
☑ Training process monitor

“Advanced user interface” allows us to customize the architecture of build network and “Training process monitor” allows monitor the process and stop if deemed necessary. In the other hand we have use the model selection criteria “Misclassification Rate”



Default number of hidden units: 3

Our network consists of a hidden layer that has 10 neurons.



Number of neurons: 10

Activation function: Hyperbolic Tangent

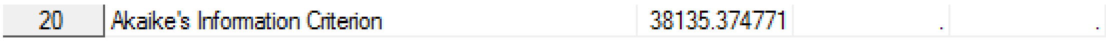
Combination function: Linear - General

Bias: Bias

Activation function used is hyperbolic tangent because it offers better results after testing by trial and error. We have considered existence of bias.

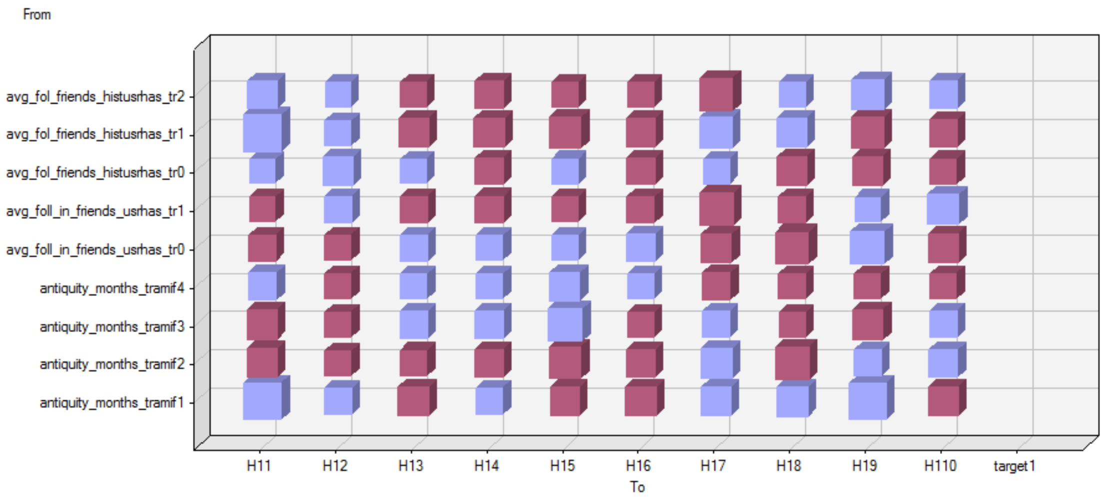
NEURAL NETWORK FOR THE MODEL WITH THE ORIGINAL VARIABLES

With respect to the variables, we will make the same assumptions as we did with the model with the original variables in the decision tree.

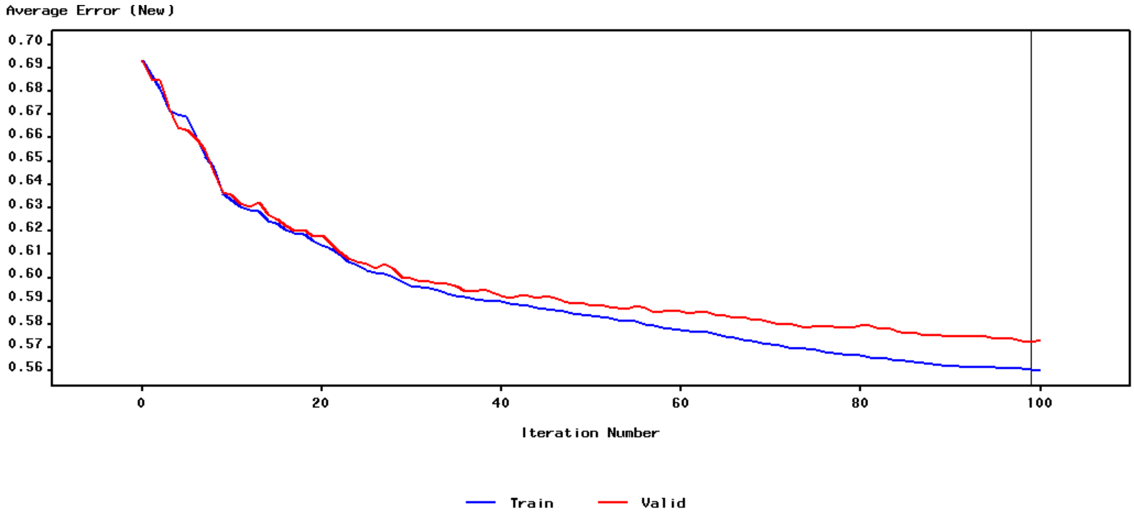


20	Akaike's Information Criterion	38135.374771	.	.

With the neural network, we get a series of statistics, where can observe the value of Akakike's Information Criterion (AIC), the lower value will be the most parsimonious model.



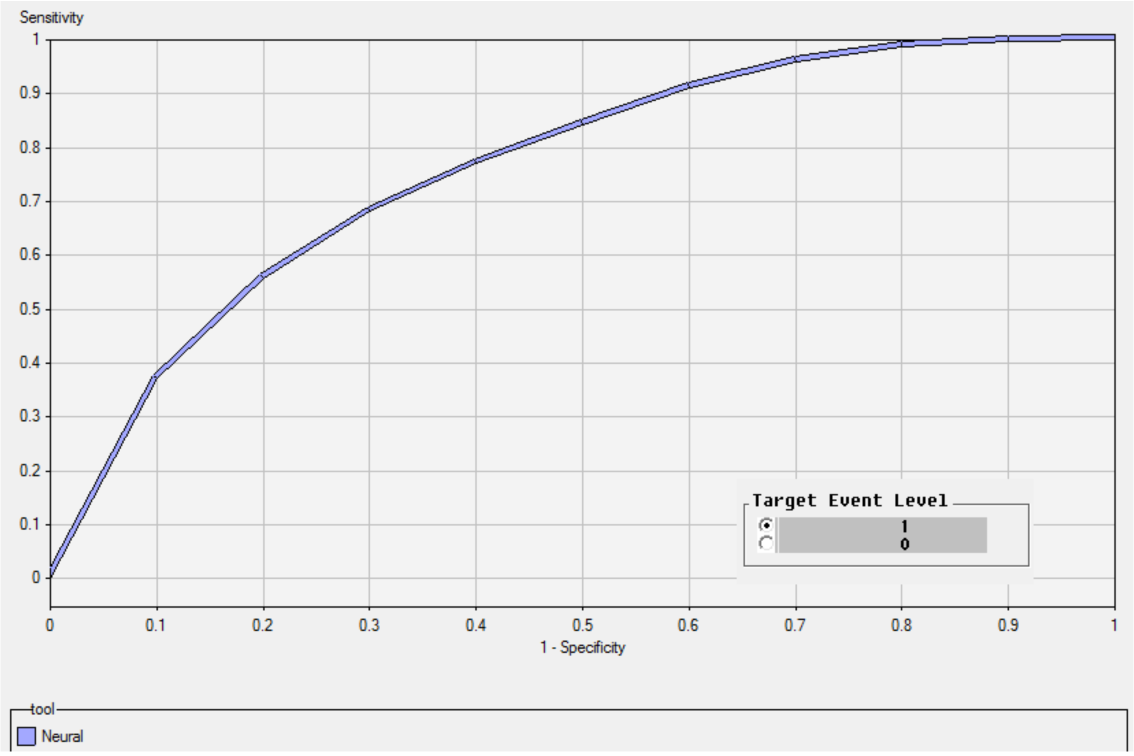
As we mentioned previously, it is difficult to interpret the weights. The blue boxes indicate the weights are positive and the purple boxes are negative weights. The size of the box indicates the influence of weight.



The algorithm must stop at the time when a cycled, the time it takes for the network to go through all the training patterns, stop predicting decrease average error in coordination in the training and validation tables. The optimal model will be around 100 cycles.

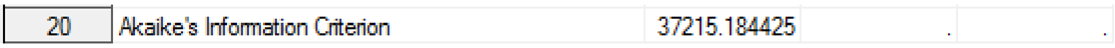
Using the SAS macro we have described above for the test table to calculate the area under the ROC curve:

AREA UNDER THE ROC CURVE: 0.76



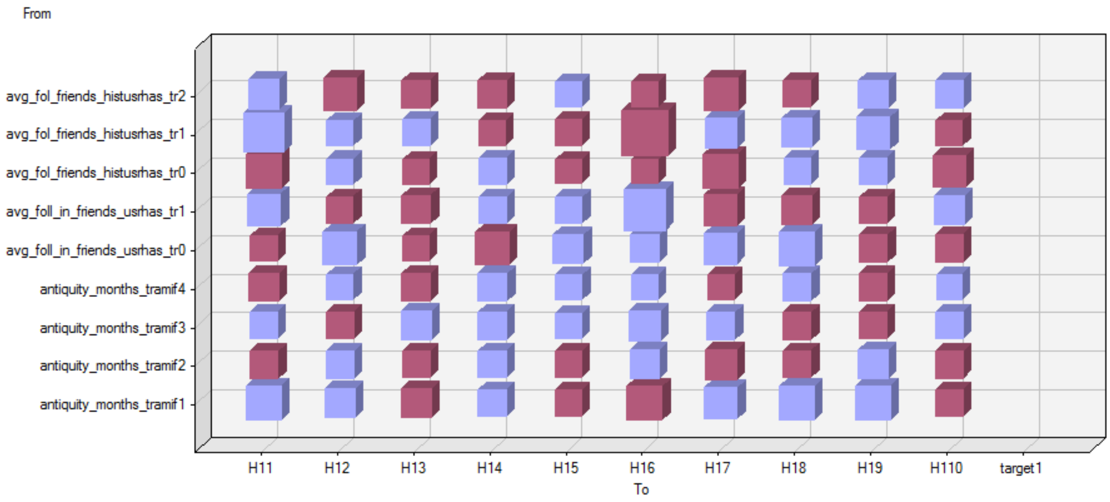
NEURAL NETWORK FOR THE MODEL WITH THE ORIGINAL VARIABLES AND THE TWO NEW VARIABLES

With respect to the variables, we will make the same assumptions as we did with the model with the original variables and the two new variables in the decision tree

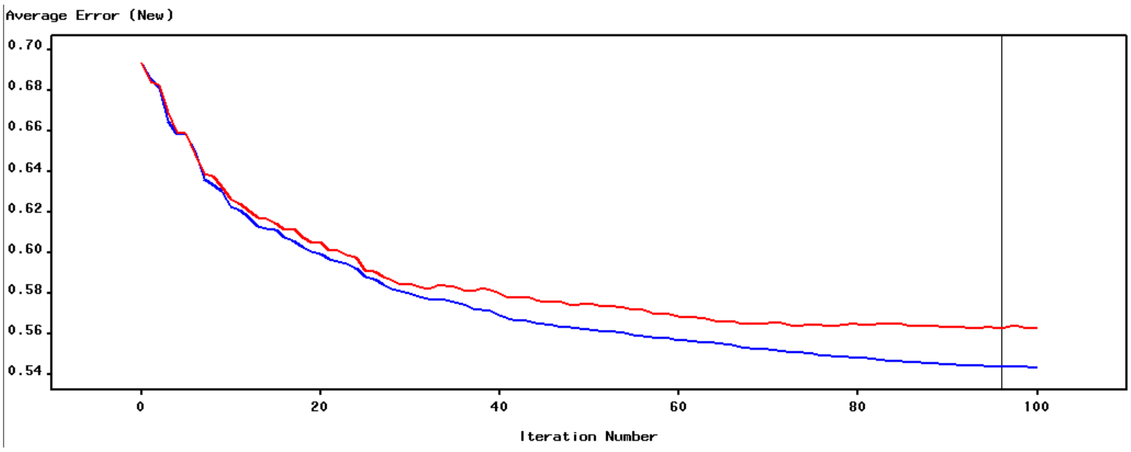


20	Akaike's Information Criterion	37215.184425	.	.

With the neural network, we get again a series of statistics, where can observe that the value of Akakike's Information Criterion (AIC) is lower in this model than the previous, so it will be more parsimonious.



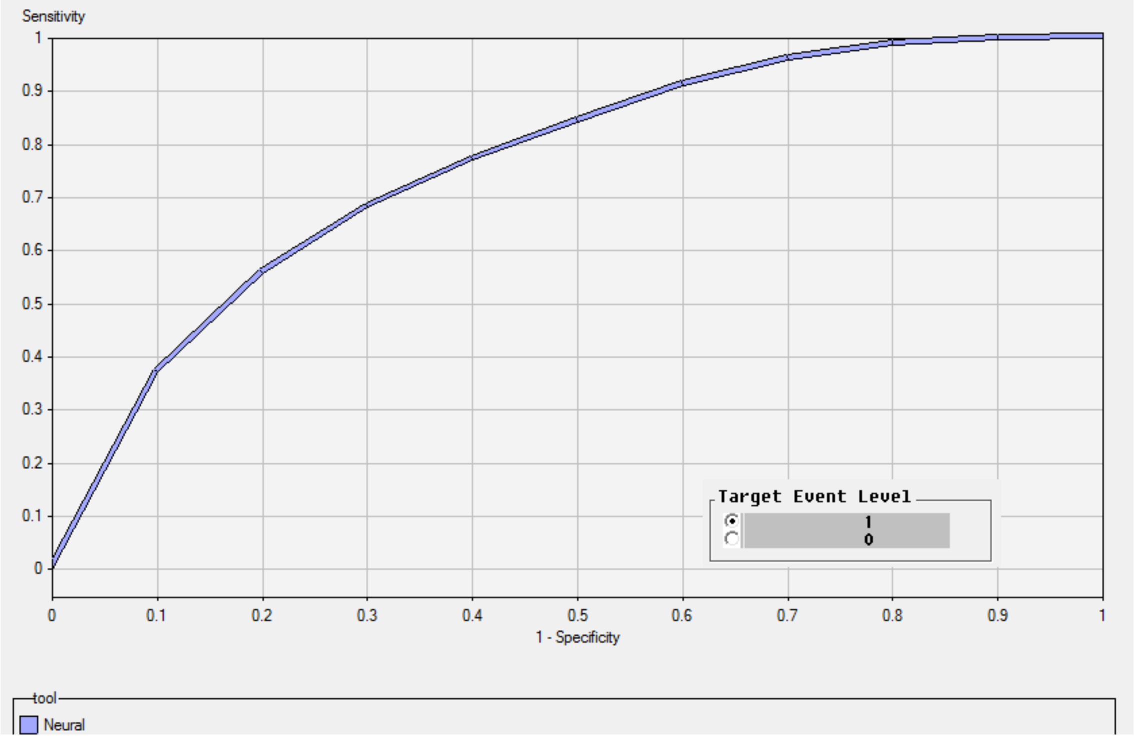
At first glance, it seems that we cannot remove any neuron because it has not low weights for all.



The optimal model will be around 100 cycles again, but it seems to be one or two cycles earlier than in the previous model

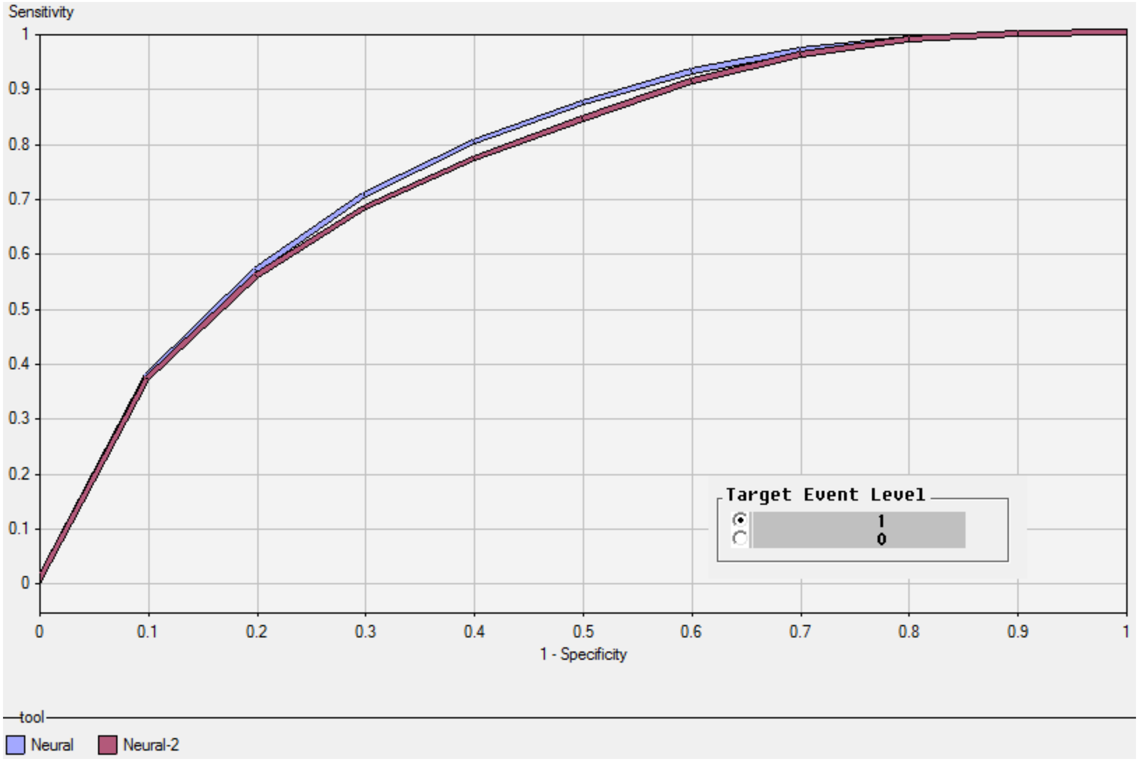
Using the SAS macro we have described above for the test table to calculate the area under the ROC curve:

AREA UNDER THE ROC CURVE: 0.77



We can observe that we significantly improve because the area under the ROC curve increases in 0.01 with the inclusion of the two new variables in the model (0.76 to 0.77).

If we plot the ROC curve simultaneously for both neural networks with the assistance of the **ASSESSMENT NODE** of Enterprise Miner, we can observe more easily that we improve with the inclusion of the two new variables (*hashtag content* and *user's activity time*) in the model:



- Neural-2 refers to the decision tree for the model with the original variables
- Neural refers to the decision tree for the model with the original variables and the two new variables

STRATEGIES

ADVANTAGES AND DISADVANTAGES OF EACH MODEL

- An advantage of logistic regression is that it's easy to interpret. A disadvantage is that the meaning of the variables should be clearly stated to introduce them properly at the model.

- Decision trees have an advantage that it's provide a clear and easy way to interpret and quantify the probability of an outcome. A disadvantage may be that the model is very sensitive with changes on the training data.

- The Neural Networks have advantages like they are capable of adaptive learning. Therefore, there aren't need to specify probability distribution functions. And a disadvantage is that we can't see what variables are part of the model or their relationships selected by the algorithm.

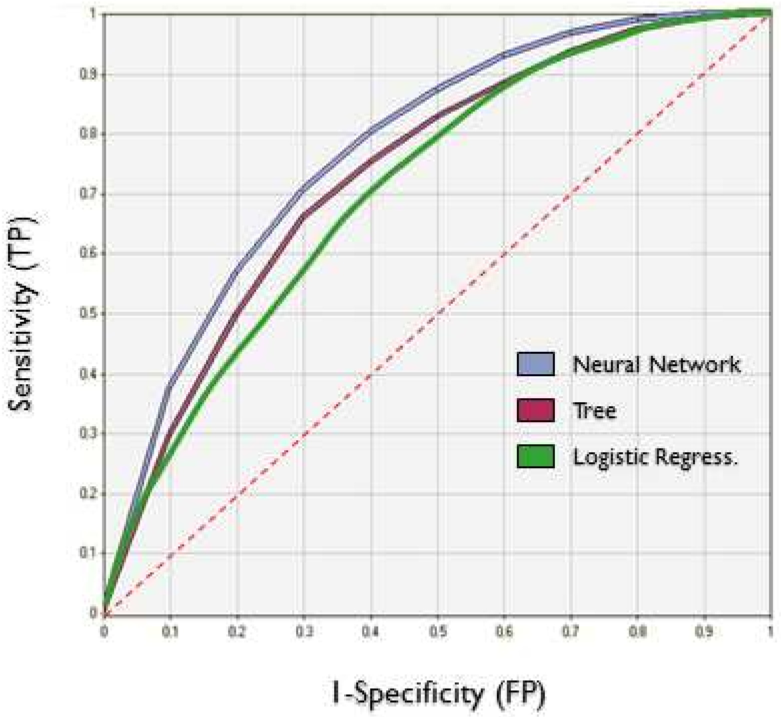
RESULTS

As the model is binary, the results are labeled as positive and negative. We want to predict when a hashtag will be used. If we succeed, it will be a true positive, and if we fail, it will be a false positive.

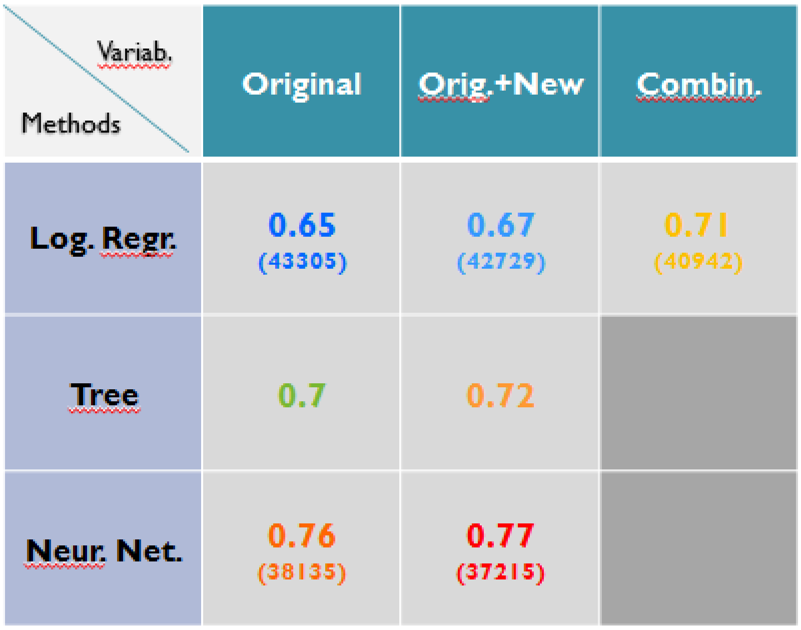
To compare the results in three models, we will use the value of the area under the ROC curve. In this curve, the x-axis represents the ratio of false positives and the y-axis represents the ratio of true positives.

This values moves between 0.5 and 1. When classification is random, we get a point on the diagonal line.

As it can see in the next image, the biggest and best value on validation table is obtained by applying a neural network model.



The following table shows the numerical values obtained with each model on the original variables (first column), the original variables and the two new variables (second column), and combined variables (third column).



Variab. Methods	Original	Orig.+New	Combin.
Log. Regr.	0.65 (43305)	0.67 (42729)	0.71 (40942)
Tree	0.7	0.72	
Neur. Net.	0.76 (38135)	0.77 (37215)	

Using only the original variables, the neural network model gives a better result, when we introduce two new variables, hashtag content and user's activity time, all models improve, and neural network continue be the best (0.77)

It can be seen that the more red are the colors the better the result.

The boxes in the third columns of decision trees and neural network are empty because these models perform their own implicit combination of variables. In the logistic regression model, we interacts some variables and the model improves. We obtain at ROC curve, 0.71

AIC (Akakike's Information Criterion) is a measure of the complexity of the model. The lower the AIC, the better the model is.

In this case, it 38135.374771 with the neural network with original variables, and is better with two new variables, 37215.184425.

CONCLUSIONS

- We wanted to predict a binary variable as to whether a tweet will be retweeted.

- We have considered three models, logistic regressions, decision trees and neural networks, as strategies to predict the binary target variable.

- In addition to the initially proposed variables, we have included two new variables, hashtag content and user's activity time, with which we have obtained best predictions in all models used.