Score don

Maryem El Braij, Paolie Amoussou,
Xenia Zaricinii, Benoît Grand

Sommaire

Introduction

Nous menons cette étude dans le but d'aider une grande association à trouver parmi sa liste de donateurs ceux qui ont le plus de chance de faire un don à l'année N compte tenu du profil et des comportements passés de ces derniers. Plus explicitement, notre objectif est de pouvoir prédire à travers une modélisation pour chacun des individus de la base de données fournie si il fera un don ou pas à l'année N en nous basant sur les caractéristiques (informations) de ce dernier. Pour atteindre cet objectif nos travaux sont structurés comme suit:

- exploration de nos données d'études , transformations des variables et construction de quelques agrégats, des indicateurs et de la cible
- sélection des variables éligibles pour le modèle grâce à croisement cible et une analyse de corrélation
- présentation des résultats de la modélisation et du test de performance.

1- Présentation de la base de données

La base de données “donnees_score” mise à disposition contient 3 grandes catégories de variables à savoir:

- des variables relatives aux comportements de don sur chacune des 11 dernières années (nombre de don et montant des dons par exemple) précédent l’année N.

Il faut noter que nous avons également ces variables pour l’année N que nous isolerons pour servir de variable cible

- des variables personnelles propres à chaque donateur (civilité et tranche d’âge par exemple) et à son mode règlement des dons.
- des variables relatives au statut du donateur au sein de l’organisation et aux informations que détient cette dernière concernant le donateur en question (Exemple: top portable renseigné ou top membre).

2- Description du DataFrame

Statistiques descriptives

de Sas

Variable	N	Minimum	Lower Quartile	Mean	Median	Upper Quartile	Maximum	Std Dev	Skewness	Kurtosis
nb_don_annee_N_moins_11	73124	0	0	0	0	0	0	0	.	.
nb_don_annee_N_moins_10	73124	0	0	0.0309748	0	0	3.0000000	0.1804423	6.1350855	41.4144090
nb_don_annee_N_moins_09	73124	0	0	0.0309885	0	0	4.0000000	0.1829613	6.3957796	47.2461435
nb_don_annee_N_moins_08	73124	0	0	0.0931568	0	0	5.0000000	0.3060105	3.4320391	13.8823559
nb_don_annee_N_moins_07	73124	0	0	0.0834746	0	0	4.0000000	0.2848293	3.3751549	11.6500566
nb_don_annee_N_moins_06	73124	0	0	0.1144084	0	0	4.0000000	0.3316487	2.8365375	8.0260165
nb_don_annee_N_moins_05	73124	0	0	0.1710246	0	0	4.0000000	0.3946562	2.1337935	4.0344820
nb_don_annee_N_moins_04	73124	0	0	0.1810760	0	0	3.0000000	0.3961815	1.8860830	2.4085138
nb_don_annee_N_moins_03	73124	0	0	0.2304442	0	0	5.0000000	0.4340099	1.5299736	1.4495235
nb_don_annee_N_moins_02	73124	0	0	0.3802855	0	1.0000000	5.0000000	0.5094866	0.8198150	-0.1938635
nb_don_annee_N_moins_01	73124	0	0	0.3382200	0	1.0000000	19.0000000	0.5106947	1.8206827	25.9307185
nb_don_annee_N	73124	0	0	0.2608993	0	1.0000000	4.0000000	0.4506249	1.2744754	0.2635059
Montant_Don_N_moins_11	73124	0	0	0	0	0	0	0	.	.
Montant_Don_N_moins_10	73124	0	0	1.5879667	0	0	6097.96	25.6189981	186.3562766	43890.63
Montant_Don_N_moins_09	73124	0	0	1.5119317	0	0	1524.49	13.5035138	35.4650805	2802.42
Montant_Don_N_moins_08	73124	0	0	4.7513060	0	0	5703.79	31.0512355	91.2316272	15698.13
Montant_Don_N_moins_07	73124	0	0	4.4125362	0	0	5000.00	35.2222477	79.7832468	9912.19
Montant_Don_N_moins_06	73124	0	0	6.0801607	0	0	10000.00	47.2262305	135.9169192	27644.55
Montant_Don_N_moins_05	73124	0	0	9.4178040	0	0	3050.00	37.0860939	23.8804168	1342.03
Montant_Don_N_moins_04	73124	0	0	10.0119848	0	0	6000.00	46.0224817	57.2349724	6308.37
Montant_Don_N_moins_03	73124	0	0	13.0731851	0	0	3000.00	41.5849771	15.6625665	680.8726733
Montant_Don_N_moins_02	73124	0	0	23.7294221	0	30.0000000	7500.00	68.1317795	38.1137889	3338.97
Montant_Don_N_moins_01	73124	0	0	23.2669311	0	30.0000000	8000.00	65.2237691	33.2914257	3282.24
Montant_Don_N	73124	0	0	19.3238344	0	10.0000000	10000.00	67.8656627	53.2000488	6748.28
TR_age_65_75_ans	73124	0	0	0.1214102	0	0	12.0000000	0.6641079	7.1881605	64.3370172
NB_don_par_Telephone	73124	0	0	0.4297768	0	1.0000000	14.0000000	0.8617683	3.7179645	21.6177499
NB_don_par__INTERNET	73124	0	1.0000000	1.1612193	1.0000000	1.0000000	19.0000000	1.0315270	1.8476712	7.9071914
NB_don_par_Autre_Canal	73124	0	0	0.0630573	0	0	17.0000000	0.3852412	13.1353165	294.7363769
NB_paiement_CB	73124	0	1.0000000	1.1906351	1.0000000	2.0000000	19.0000000	1.0392944	1.8477548	7.8243543
NB_paiement_CHEQUE	73124	0	0	0.4564302	0	1.0000000	19.0000000	0.9389425	4.0602089	27.3406381
NB_paiement_AUTRE	73124	0	0	0.0069881	0	0	6.0000000	0.0933660	18.6449888	573.2614384
top_telephone_renseigne	73124	0	0	0.4009901	0	1.0000000	1.0000000	0.4901024	0.4040487	-1.8367949
top_portable_renseigne	73124	0	0	0.1593594	0	0	1.0000000	0.3660134	1.8614083	1.4648736
top_membre	73124	0	0	0.0176686	0	0	1.0000000	0.1317447	7.3224114	51.6191212
top_adresse_email_FAI	73124	0	0	0.5706881	1.0000000	1.0000000	1.0000000	0.4948813	-0.2856272	-1.9184695
top_email_avec_nom	73124	0	0	0.6688365	1.0000000	1.0000000	1.0000000	0.4706350	-0.7175035	-1.4852294
top_email_avec_prenom	73124	0	0	0.4516985	0	1.0000000	1.0000000	0.4976649	0.1941179	-1.9623719

Les variables représentatives: Montant_don_N_moins_1 , Montant_don_N_moins_2, Nb_don_N_moins_1 et Nb_don_N_moins_2- ces variables ont plus de valeurs différentes de 0 que les autres variables qui représentent l'historique des dons. Plus que 25% des observations différents de 0. Le reste de ce type des variables contiennent plus que 75% des valeurs nulles.

Montant_don_N_moins_11 et Nb_don_N_moins_11- ne contiennent que des 0 et n'ont pas d'importance pour notre analyse.

3- Fréquence des variables catégorielles. Civilité et l'âge.

Civilite	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Madame	31123	42.56	31123	42.56
Mademoiselle	1476	2.02	32599	44.58
Monsieur	36589	50.04	69188	94.62
Non renseigné	3726	5.10	72914	99.71
Sociéte	210	0.29	73124	100.00

age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
18_25_ans	915	1.25	915	1.25
25_35_ans	3677	5.03	4592	6.28
35_45_ans	4859	6.64	9451	12.92
45_55_ans	6092	8.33	15543	21.26
55_65_ans	6866	9.39	22409	30.65
65_75_ans	3365	4.60	25774	35.25
Moins de 18 ans	115	0.16	25889	35.40
Non renseigné	46263	63.27	72152	98.67
Plus_75_ans	972	1.33	73124	100.00

- Les donneurs sont majoritairement des hommes et des femmes avec une répartition plutôt égalitaire. Très peu de sociétés (0.29%) et 5.10% des personnes inconnus.
- Majoritairement la tranche d'âge est non renseignée (63.27%) . La catégorie d'âge la plus représentée est la catégories des 35-65 ans. Il est difficile de déterminer l'âge important des donneurs.

3-1 Fréquence des variables catégorielles. Origine et abonnement.

- Selon l'origine des donneurs, nous avons soit des donneurs BDD soit des Web (en total 92.7%)
- Selon l'abonnement autour de 42% ont un abonnement newsletter générale et près de 73% des donneurs ont l'abonnement quelque soit.

origine	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Autre	106	0.14	106	0.14
BDD	33324	45.57	33430	45.72
Bénévole	1262	1.73	34692	47.44
Web	34464	47.13	69156	94.57
Web actif	3968	5.43	73124	100.00

abonnement	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Newsletter générale	30695	41.98	30695	41.98
Newsletter spécifique	22800	31.18	53495	73.16
Pas d'abonnement	19629	26.84	73124	100.00

4- Calcul de nouveaux indicateurs et agrégats

Afin d'améliorer la qualité de notre analyse, nous avons créé de nouvelles variables indicateurs, tout en utilisant les données à disposition:

- montant_don_total: correspond au montant total des dons sur la période étudiée.
- nb_don_total:correspond au nombre total des dons sur la période étudiée.
- recence: indique le nombre d'année écoulées entre le dernier don et l'année N.
- max_mnt : est montant de don maximal sur les 11 années dernière années.
- min_mnt: est au montant de don minimal sur les 11 années dernière années.
- mean_mnt: est la moyenne des montants de don sur la période étudiée.
- pct_annee_avec_don: fait référence au pourcentage des années avec au moins un dons (exemple :si l'individu fait des dons sur 5 années alors pct_annee_avec_don=(5/11)*100=45,45)

5- Nettoyage et transformation des données

Créations de variables binaires selon les valeurs des colonnes :

- age
- Civilite
- origine
- abonnement

On supprime ensuite les colonnes précédentes et les colonnes non nécessaires :

- age_Non_renseign_
- Civilite_Non_renseign_
- identifiant_personne

5-1 Les variables exclus de l'analyse.

- Montant_don_N_moins_11: cette variable ne contienne que des 0 et n'apporte pas d'information dans notre base.
- Nb_don_N_moins_11: cette variable ne contienne que des 0 et n'apporte pas d'information dans notre base.
- Montant_don_N : représente la même information que notre variable cible mais dans une autre unité présentée.
- identifiant_personne: identifiant unique qui n'est pas utile à la prédiction
- nb_don_annee_N qui a été remplacé par la nouvelle variable cible.

5-2 Les variables exclus de l'analyse.

- TR_age_65_75_ans : l'information de cette variable se trouve dans la variable âge. De plus, cette variable contient des modalités 2,3 et 4 qui sont interprétables, car ces modalités sont présents pour les personnes particulières et pas pour les sociétés.
- min_mnt: cette variable qui représente le montant minimum des dons pour chaque individu contient uniquement des 0 pour toutes les observations; elle n'apporte donc aucune information. Cela s'explique par le fait que sur les 11 années observées il y a au moins une année où chaque donateur n'a pas fait un don.

5-3 Calcul de la variable cible.

Nous calculons la variable target à partir de la variable nb_don_annee_N qui représente le nombre de dons l'année N.

- Si nb_don_annee_N est égal à 0, alors la cible est aussi 0.
- Si nb_don_annee_N = ou > à 1, alors notre cible vaut 1.

Nous obtenons alors une cible binaire don_annee_N.

don_annee_N	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	54402	74.40	54402	74.40
1	18722	25.60	73124	100.00

Selon la répartition des modalités de la cible, nous constatons que dans 74.40% des cas il n'y avait pas de don l'année N et dans 25.60% des cas il y avait au moins 1 don. La classe 0 (pas de don l'année N) est majoritaire.

6- Corrélation avec notre cible

NAME	don_annee_N
don_annee_N	1
age_18_25_ans	0,065039557
age_25_35_ans	0,139165547
age_35_45_ans	0,148822585
age_45_55_ans	0,185073698
age_55_65_ans	0,235807757
age_65_75_ans	0,162328967
age_Moins_de_18_ans	0,017836517
age_Plus_75_ans	0,077740692
nb_don_annee_N_moins_10	0,029710233
nb_don_annee_N_moins_09	0,013501259
nb_don_annee_N_moins_08	0,051394388
nb_don_annee_N_moins_07	-0,019664354
nb_don_annee_N_moins_06	-0,026828064
nb_don_annee_N_moins_05	-0,042550228
nb_don_annee_N_moins_04	-0,044931757
nb_don_annee_N_moins_03	-0,005730709
nb_don_annee_N_moins_02	0,105062875
nb_don_annee_N_moins_01	0,208233241
nb_don_total	0,098487621
NB_don_par_Telephone	0,015370219
NB_don_par__INTERNET	0,110924687
NB_don_par_Autre_Canal	0,010853714
NB_paiement_CB	0,104646979
NB_paiement_CHEQUE	0,02338414
NB_paiement_AUTRE	0,012138298
origine_Autre	-0,022351499
origine_BDD	-0,476204168
origine_B_n_vole	0,064935524
origine_Web	0,447570499
origine_Web_actif	0,02711976
abonnement_Nl_g	0,530538022
abonnement_Nl_s	-0,283186689
pas_dabonnement	-0,294827628
top_telephone_renseigne	0,120302472
top_portable_renseigne	0,235723627
top_membre	0,067596259
top_adresse_email_FAI	0,068404726
top_email_avec_nom	0,076901254
top_email_avec_prenom	0,0493816
Civilite_Madame	0,025004022
Civilite_Mademoiselle	0,005369335
Civilite_Monsieur	0,030587498
Civilite_Soci_te	-0,00513318
pct_annee_avec_don	0,097044342
recence	-0,278133025
Montant_Don_N_moins_10	0,013825634
Montant_Don_N_moins_09	0,021061528
Montant_Don_N_moins_08	0,03533748
Montant_Don_N_moins_07	-0,002729361
Montant_Don_N_moins_06	0,000740671
Montant_Don_N_moins_05	-0,011444965
Montant_Don_N_moins_04	-0,002902035
Montant_Don_N_moins_03	0,020080217
Montant_Don_N_moins_02	0,088951445
Montant_Don_N_moins_01	0,152659911
max_mnt	0,001467824
mean_mnt	0,080479174
montant_don_total	0,080479174

6-1 Analyse matrice de corrélations

On retient quelques variables directement corrélées aux dons à l'année N :

- originie_Web
- abonnement_Nl_g
- origine_BDD

Au contraire, voici quelques variables très peu corrélées à notre cible:

- max_mnt
- NB_don_par_Telephone
- Civilite_soci_te

7- Analyse du croisement cible

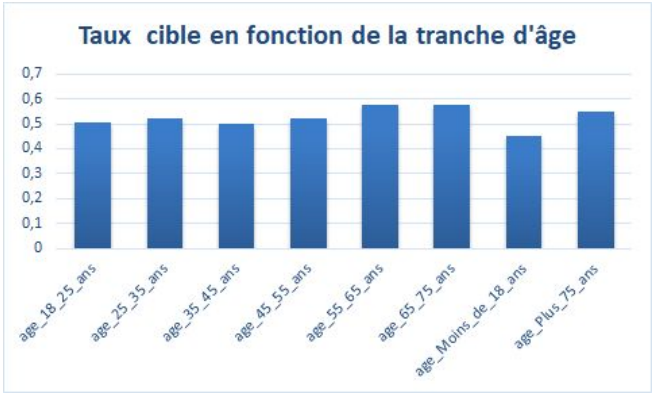
variable	modalite	COUNT	count_tot	tx_cible	impact
age_18_25_ans	0	18257	72209	0,253	-
age_18_25_ans	1	465	915	0,508	+
age_25_35_ans	0	16810	69447	0,242	-
age_25_35_ans	1	1912	3677	0,520	+
age_35_45_ans	0	16295	68265	0,239	-
age_35_45_ans	1	2427	4859	0,499	+
age_45_55_ans	0	15530	67032	0,232	-
age_45_55_ans	1	3192	6092	0,524	+
age_55_65_ans	0	14769	66258	0,223	-
age_55_65_ans	1	3953	6866	0,576	+
age_65_75_ans	0	16775	69759	0,240	-
age_65_75_ans	1	1947	3365	0,579	+
age_Moins_de_18_ans	0	18670	73009	0,256	-
age_Moins_de_18_ans	1	52	115	0,452	+
age_Plus_75_ans	0	18189	72152	0,252	-
age_Plus_75_ans	1	533	972	0,548	+
origine_Autre	0	18722	73018	0,256	+
origine_Autre	1	0	106	0,000	-
origine_BDD	0	17759	39800	0,446	+
origine_BDD	1	963	33324	0,029	-
origine_B_n_vole	0	18129	71862	0,252	-
origine_B_n_vole	1	593	1262	0,470	+

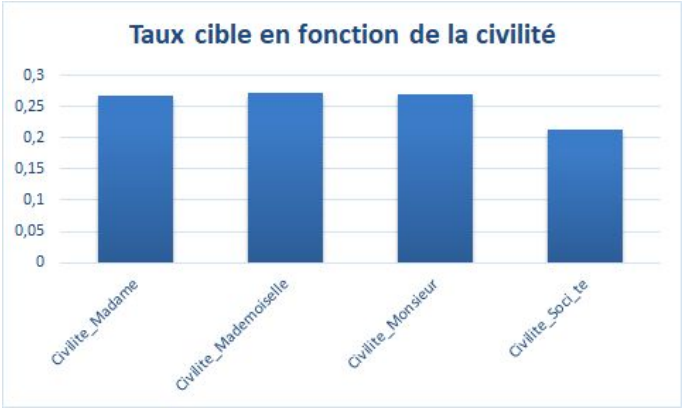
Ce tableau issus du croisement entre la variable cible et chaque variable catégorielle a pour but de mesurer l'impact sur la cible à travers le calcul d'un taux de cible. La dernière colonne "impact" nous permet de voir dans quel sens (positif ou négatif) va l'impact d'une variable donnée sur la cible en fonction de la modalité qu'elle prend. Les cellules en violet matérialisent les taux de cible les 10% plus élevés parmis l'ensemble des variables considérées pour cette analyse et les cellules surlignées en orange matérialisent les 10% les moins élevés. On voit que par exemple les personnes d'un âge compris entre 18 et 25 ans font partie de ceux qui ont plus de chance de faire un don à l'année N car ils ont un taux de cible largement dessus au taux de cible moyen.

A contrario, le fait d'avoir une origine BDD pour un donateur à pour effet d'influencer négativement la probabilité de ce dernier à faire un don à l'année N.

7-1 Quelques statistiques descriptives issues du tableau croisé

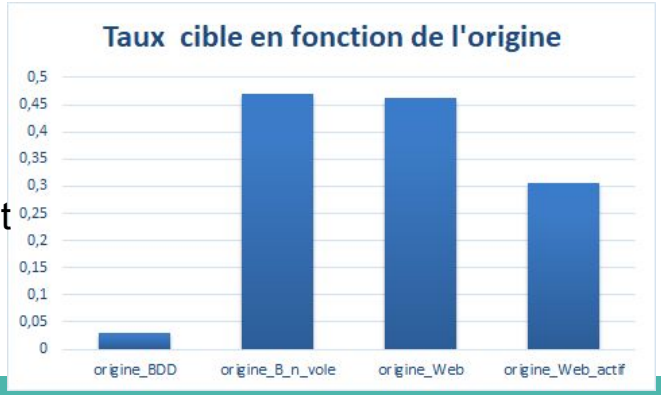
Les personnes âgées (+ de 55 ans) donnent plus que toutes les autres catégories





La civilité importe peu

origine_BDD est loin derrière origine_B_n_voie, Web et Web_actif, Ceci est assez logique



8- Sélection de variables pour la modélisation

- La matrice de corrélation nous a permis en plus d'identifier les variables qui sont fortement corrélées à la variable cible de voir les variables explicatives qui ont de fortes corrélations entre elles. Afin d'éviter d'avoir un modèle dans lequel certaines informations sont surpondérées du fait des corrélations élevées, nous avons décidé d'en supprimer à chaque fois que le taux de corrélation est supérieur en valeur absolue à 0,70. Ainsi les variables qui étaient fortement corrélées à d'autres et qui ont été exclues pour la suite de la modélisation sont :

montant_don_total, civilite_madame, origine_B_n_vole, origine_Web, NB_don_par_Telephone, NB_paiement_CB nb_don_total.

8-1 Sélection de variables pour la modélisation

- Aussi, pour éviter le problème de corrélation avec la constante pendant la modélisation nous avons éliminé les variables: moins_de_18_ans et pas_dabonnement.

Notre choix de suppression de ces variables s'est principalement porté sur les variables explicatives qui les moins discriminantes.

- Nous avons également décidé de ne pas garder la variable "récence" car celle-ci a généré des valeurs manquantes pour les donateurs qui n'ont fait aucun don pendant les 11 dernières années mais qui en ont quand même fait à l'année N. Ces observations étaient d'un nombre important et nous avons tenu à garder ce cas de figure pour la modélisation.

- Enfin nous avons supprimé les variables top_email_avec_nom et top_email_avec_prenom car nous les jugeons non pertinentes pour expliquer la variable cible.

9- Résultats de la modélisation

Significativité globale du modèle :

Les p-values associées aux tests de ratio de vraisemblance, de Score et de Wald indiquent que l'hypothèse nulle d'absence de significativité globale des paramètres est rejetée.Autrement dit le modèle est globalement significatif au seuil de 1%, ce qui signifie qu'il existe au moins un coefficient significativement différent de 0 dans la régression .

Significativité des coefficients :

En examinant les p-value associés à chacun des paramètres étudiés, on constate que les variables précédemment sélectionnées sont statistiquement significatives au seuil de 1% (p-value < 0.001) , à l'exception des variables civilite_monsieur,civilite_madame , montant de don pour les années supérieures à N-2 ,nombre de paiement "autre",le nombre de don pour l'année N-2 ainsi que le nombre de don pour les années supérieures à N-3 et pourcentage année avec don et enfin la variable top téléphone renseigné.

9-1 Résultats de la modélisation

Analyse des coefficients :

Les résultats de la modélisation est en total adéquation avec l'analyse comparative effectué auparavant sauf pour la variable civilite_mademoiselle où le signe du coefficient estimé est négatif , ce qui est incohérent avec l'impact attendu de cette variable.

9-2 Evaluation de la qualité du modèle

rang	eff	eff_test	eff_target	eff_target_test	seuil_min	seuil_max	proba_min	proba_max	taux_target	taux_target_test
0	5135	2172	38	31	-81.5453733	-5.019614379	3.84853E-36	0.006563707	0.740019474	1.427255985
1	5136	2172	47	28	-5.019459778	-4.548064794	0.006564715	0.010476749	0.915109034	1.289134438
2	5110	2163	281	115	-4.547566657	-3.844493722	0.010481915	0.020948982	5.499021526	5.316689783
3	5180	2191	77	40	-3.844280068	-3.471278022	0.020953364	0.0301406	1.486486486	1.825650388
4	5139	2161	511	216	-3.471233662	-2.040617604	0.030141896	0.115003859	9.943568788	9.995372513
5	5114	2173	361	168	-2.039148723	-1.440294651	0.115153443	0.191499724	7.059053578	7.731247124
6	5169	2172	833	347	-1.440259872	-0.787946725	0.191505109	0.312609717	16.11530277	15.97605893
7	5140	2172	2643	1090	-0.787574587	0.536037567	0.31268969	0.630890174	51.42023346	50.18416206
8	5141	2173	3837	1633	0.536073059	1.39925214	0.630898439	0.802065188	74.63528496	75.14956282
9	5140	2171	4524	1902	1.399297154	21.75641716	0.802072334	1	88.0155642	87.60939659

Afin d'évaluer la qualité de prévision, nous avons séparé le jeu de données en deux échantillons, nous avons utilisé l'échantillon d'apprentissage pour le calcul des coefficients , ces derniers sont appliqués ensuite sur l'échantillon test.
Le tableau ci dessus indique que les performances du modèle sont plutôt bonnes puisque les taux de cible calculés à partir des deux échantillons sont de manière générale assez proches,de plus le taux de cible augmente d'un décile à l'autre ce qui signifie que la probabilité de faire un don l'année N croît aussi avec le décile sauf pour le décile 4 où les taux baissent .

Conclusion

Cette étude portant sur la modélisation de la probabilité de faire un don à l'année N mène à conclure que le fait de faire un don l'année N-1 et N-2 a un impact positive sur l'année N, autrement dit si un individu a fait un don pendant les deux dernières années alors il a plus de chance d'être parmi les donateurs de l'année N, ce qui confirme les constats tirés des analyses exploratoires .

L'âge de l'individu est aussi un facteur majeur , en effet il ressort des résultats de modélisation et du croisement cible que les tranches 18-25 ans et 25-35 ans sont celles qui ont le plus d'effet sur la variable cible. Cela peut être expliqué en partie par le fait que les jeunes individus appartenant à ces tranches d'âge commencent à trouver un situation économique plus aisée et sont plus engagés dans la vie associative.