Envoi de catalogue pour une campagne marketing (levée de fonds)

Introduction:

Nous avons effectué de l'apprentissage supervisé afin de prédire quels membres d'une fondation devraient recevoir des trousses de remerciement, dans le but de maximiser les dons en 2024. N'ayant pas les renseignements sur les dons pour l'année susmentionnée, nous avons dû sélectionner des prédicteurs et une variable cible à partir des données disponibles, afin de les utiliser et prédire les dons pour l'année 2024.

Initialement, nous avions planifié d'entraîner des modèles prédictifs pour chacune des années où des individus ont fait des dons, afin d'identifier les variables avec le meilleur pouvoir prédictif pour chacune des années. Par la suite, nous aurions comparé l'efficacité de ces modèles dans un processus similaire à une méta-analyse, dans le but d'identifier quelles variables permettent de prédire le mieux les dons de manière stable au fil des années. Cependant, nous avions sous-estimé le temps requis pour entreprendre une telle démarche, ce qui nous a poussé à simplifier notre approche en ne sélectionnant qu'une seule année de référence.

Nous avons donc sélectionné comme variable cible l'année pour laquelle nous avions le plus d'informations: l'année 2023. Notre raisonnement repose sur l'hypothèse qu'il doit exister une structure latente dans les données témoignant de la propension des individus à faire des dons au fil des années, à l'instar d'un trait de personnalité qui influence de manière stable les comportements des individus tout au long de leur vie. Dans cette optique, un modèle capable de découvrir cette structure pour une année donnée devrait pouvoir prédire les dons pour une autre année. Afin de découvrir ce modèle, nous avons essayé plusieurs méthodes en appuyant nos choix méthodologiques sur les notes du cours. Nous avons donc testé 7 méthodes différentes.

1. L'envoi de trousses à tous les membres de la fondation.
2. Les variables de base;
3. Toutes les variables, incluant les termes quadratiques et les interactions d'ordre 2
4. Une sélection séquentielle avec l'AIC
5. Une sélection séquentielle avec le BIC
6. Une recherche exhaustive avec algorithme génétique (sélection selon BIC)
7. La méthode LASSO avec une pénalité optimale déterminée selon le critère de l'aire sous la courbe.

Analyse exploratoire.

Description des bases de données utilisées:

Pour entraîner notre modèle, nous avons réuni toutes les données dont nous disposions, à l'exception de celles contenues dans la base de données "SocialNetworkUsage".

Il a été facile de réunir les observations des bases de données "DonationHistory" et "MembersList", en transformant la première en format "wide" et en les agrégeant par la suite avec l'identifiant unique des membres de la fondation.

La base de données "NewsLetterRead" n'utilisait pas d'identifiant unique auquel il aurait été possible de se fier pour faire correspondre les informations des membres. Toutefois, puisque nous savions que toutes les

observations qu'elle contenait étaient celles de membres de la fondation, nous avons pu agréger les informations des membres en substituant l'identifiant unique de ces derniers pour leur courriel. Ceci dit, nous nous étions assurés au préalable qu'il n'existait pas de doublons dans les courriels.

Nous avons toutefois décidé de ne pas utiliser la quatrième base de données ("SocialNetworkUsage") à cause du manque d'information nous permettant de relier ses données à celles des trois premières. Voici pourquoi:

Comme le démontre le tableau 1, les données ne sont pas indépendantes les unes des autres, puisqu'elles se répètent à toutes les 400k lignes. Ce tableau comporte effectivement trois fois les données du répondant "The MILLICANs". Par ailleurs, il y a des cas où les observations sont presque identiques, mais tout de même différentes. Par exemple, les deux dernières lignes du tableau 1 présentent un profil identique, celui de "H MILLICAN", mais ce dernier est classifié comme un membre dans la dernière ligne alors qu'il ne l'est pas dans l'avant dernière ligne.

Tableau 1

Name	email	Shares	Likes	Supporter	Member
The MILLICANs		0	1	FALSE	NA
JOHN MILLICAN		0	1	FALSE	NA
The MILLICANs		0	1	FALSE	NA
JOHN MILLICAN		0	1	FALSE	NA
The MILLICANs		0	1	FALSE	NA
JOSEPH MILLICAN	cecelia.beer@benefits.org	1	3	FALSE	NA
JOSEPH MILLICAN	cecelia.beer@benefits.org	1	3	FALSE	NA
TOMMY MILLICAN	dana.496@youppi.fr	7	120	TRUE	TRUE
TOMMY MILLICAN	dana.496@youppi.fr	7	120	TRUE	FALSE
H MILLICAN	muriel.157@myschool.edu	0	6	FALSE	FALSE
H MILLICAN	muriel.157@myschool.edu	0	6	FALSE	TRUE

Nous avons tenté d'identifier ce "H MILLICAN" dans l'optique de trouver une logique nous permettant d'agréger les données de cette base de données aux trois autres, mais sans succès. Nous avons répertorié tous les individus dont le nom contenait "MILLICAN" dans le tableau 2, mais personne n'avait le même courriel que "H MILLICAN". Nous avons donc cherché le courriel de "H MILLICAN" dans la base de données "SocialNetworkUsage" et "DonationHistory" pour découvrir que les courriels et les noms ne correspondaient pas, comme en témoigne le tableau 3. Ainsi, nous en avons conclu qu'il n'y avait pas d'information fiable sur laquelle nous fier pour relier les individus des trois premières bases de données. Compte tenu de toutes ces raisons, nous avons pris la décision d'exclure ces données par crainte qu'elles puissent biaiser nos résultats.

Tableau 2

ID	LastName	FirstName	email	Woman	Age	Salary	Education	City	Joined
7592734	MILLICAN	AMANDA	amanda.1422@myschool.edu	1	48	115500	High School	Suburban	2014
7630746	MILLICAN	AMANDA	amanda.1515@email.com	1	50	71200	University / College	City	2014
7933608	MILLICAN	CARMEN	carmen.1062@mymail.ca	1	81	68600	High School	City	2016
7889600	MILLICAN	CHRIS	chris.1076@proudof.me	0	32	128500	University / College	Suburban	2017
7142699	MILLICAN	DAVID	david.1776@mail.me	0	46	67600	University / College	Suburban	2020
7251560	MILLICAN	DEBRA	debra.558@mymail.ca	1	42	154400	University / College	Rural	2017
7438417	MILLICAN	GREGORY	gregory.1040@fakemail.com	0	64	43200	University / College	City	2021
7697814	MILLICAN	INGRID	ingrid.86@target.me	1	19	66900	High School	Downtown	2019
7879304	MILLICAN	JAMES	james.15600@myschool.edu	0	16	7700	University / College	Downtown	2022
7287768	MILLICAN	JESUS	jesus.255@mail.me	0	90	10800	High School	Suburban	2013
7003005	MILLICAN	JOSEPHINE	josephine.4@overview.com	1	22	34200	High School	City	2016
7485080	MILLICAN	MARK	mark.2441@paradise.com	0	20	10900	University / College	Suburban	2022
7344146	MILLICAN	MATTHEW	matthew.1214@youppi.fr	0	47	15100	University / College	City	2020
7248148	MILLICAN	MAUREEN	maureen.140@youppi.fr	1	58	32200	High School	City	2022
7453448	MILLICAN	PIERRE	pierre.53@target.me	0	37	209800	University / College	Downtown	2018
7256587	MILLICAN	STEPHANIE	stephanie.559@target.me	1	42	84200	University / College	City	2017
7346226	MILLICAN	STEVEN	steven.1524@target.me	0	44	33700	High School	Downtown	2022
7095757	MILLICAN	TARA	tara.57@target.me	1	45	16800	High School	Suburban	2014

Tableau 3

Name	email
H MILLICAN	muriel.157@myschool.edu
MURIEL GARCIA	muriel.157@myschool.edu

Création de variables

À cette étape, la base de données contenait environ 35 colonnes, représentant chacune une variable. Parmi celles-ci, 11 colonnes représentaient les dons à travers les années et 12 autres l'ouverture de l'infolettre pour chacun des mois de l'année. Pour réduire le nombre de variables, nous avons décidé de faire la somme de ces colonnes pour créer deux nouvelles variables, nous permettant de minimiser le temps requis pour les analyses tout en conservant le plus d'information possible. Deux nouvelles variables ont donc été créées, nommées frq_ouverture (nombre total d'ouverture de l'infolettre) et don_no2023 (total des dons à travers les années sauf 2023). Nous avons également créé deux autres nouvelles variables. La première, appelée ancientete, est la même chose que la variable Joined, mais plus facilement interprétable. Nous avons également créé la variable don_2023, une variable binaire représentant le fait d'avoir donné à la fondation en 2023 ou non. Cette dernière nous sera utile pour connaître l'espérance du montant des dons des individus qui ont effectivement fait un don, afin de connaître le nombre de trousses à envoyer pour maximiser nos profits nets.

1. ancientete: aussi définie comme "Joined" dans le code, cette variable mal orthographiée représente le nombre d'années qui se sont écoulées entre l'adhésion des membres de la base de données à la fondation et aujourd'hui.
2. don_no2023: Le montant total des donations des membres de la base de données, de leur adhésion à la fondation jusqu'à l'année 2022.
3. frq_ouverture: Le nombre total de fois que les membres de la base de données ont ouvert l'infolettre pendant l'année 2023.
4. don_2023: (Variable dépendante) Le fait d'avoir fait ou non un don en 2023.

Note: Nous sommes conscients que la sommation des données des ouvertures de l'infolettre et des dons par année a réduit la granularité de l'information dont nous disposions et probablement l'exactitude des prévisions. Cependant, nous croyons que cela sera compensé par une meilleure interprétabilité des résultats.

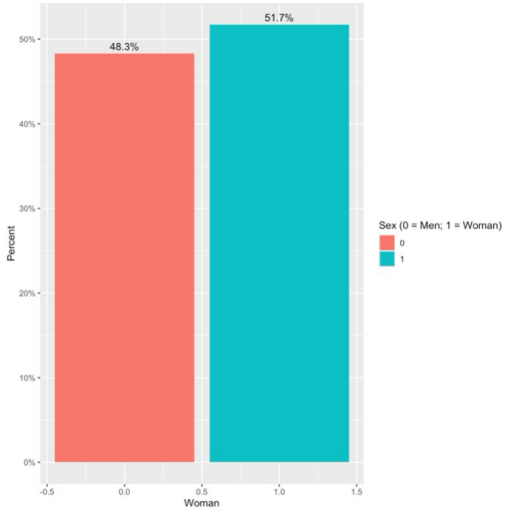
Exploration des données:

Afin de mieux connaitre les données avec lesquelles nous allions travailler, nous avons fait une exploration visuelle de celles-ci. Nous avons commencé par une exploration univariée des données, dont nous présentons ici les faits saillants:

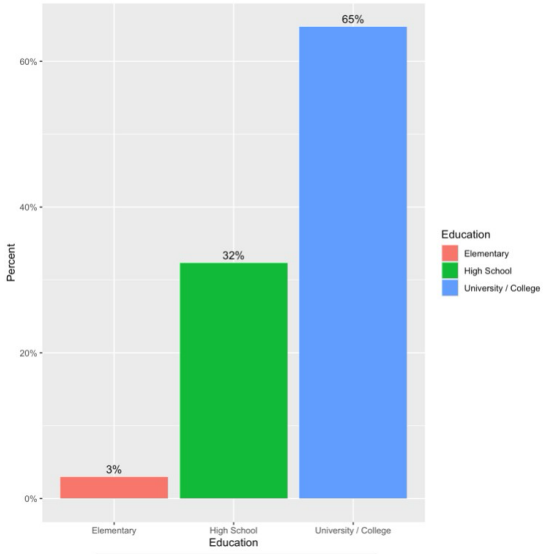
- L'échantillon contient les données de personnes mineures. En temps normal, nous aurions dû les éliminer pour nous conformer aux lois en vigueur.
- Presqu'un membre sur cinq (~18%) a rejoint la fondation l'année dernière. Cette proportion représente le double de celle de l'année précédente (~9%).
- Il est extrêmement difficile de représenter graphiquement la distribution des dons compte tenu de l'énorme quantité d'observations et des outliers. Il est intéressant de noter que les dons présentent 48 montants différents.

Bien que la composition de certaines catégories sociodémographiques soit disproportionnée (e.g. 3% de personnes dont le plus haut niveau de scolarité atteint soit l'école primaire par rapport à 65% qui ont atteint l'université) elle ne devrait pas poser de problème lors de la création d'échantillons de test et de validation, compte tenu de la grande taille de l'échantillon.

Tableau 4 & 5.

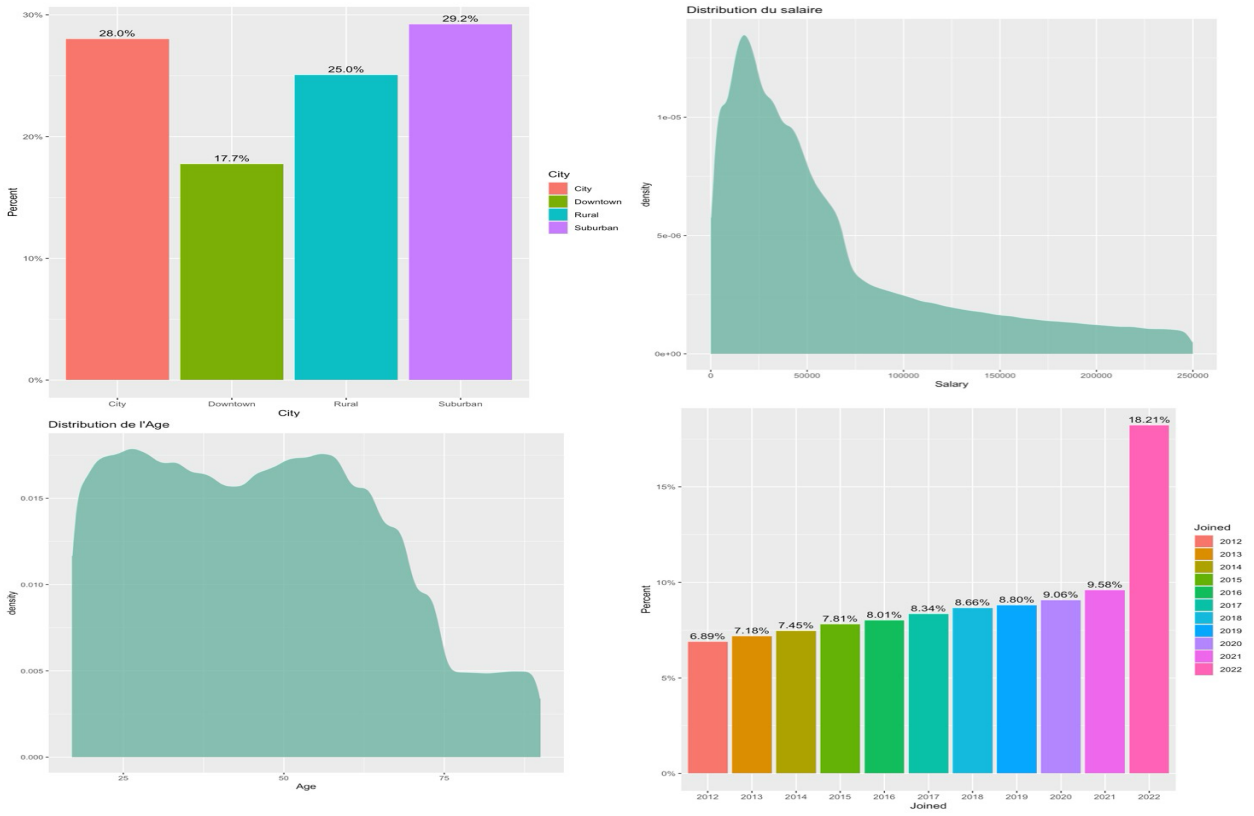


Woman	N	Frequency
0	483044	0.48
1	516956	0.52



Education	N	Frequency
Elementary	29770	0.03
High School	323025	0.32
University / College	647205	0.65

~Tableau 4 (à gauche) distribution de la variable Sexe;
Tableau 5 (à gauche) distribution du milieu de vie.~



~Tableau 6 (en haut à gauche) représentant la distribution de la variable “city”;
Tableau 7 (en haut à droite) distribution du salaire;
Tableau 8 (en bas à gauche) distribution de l’âge ;
Tableau 9 (en bas à droite) distribution de l’année d’adhésion des membres.~

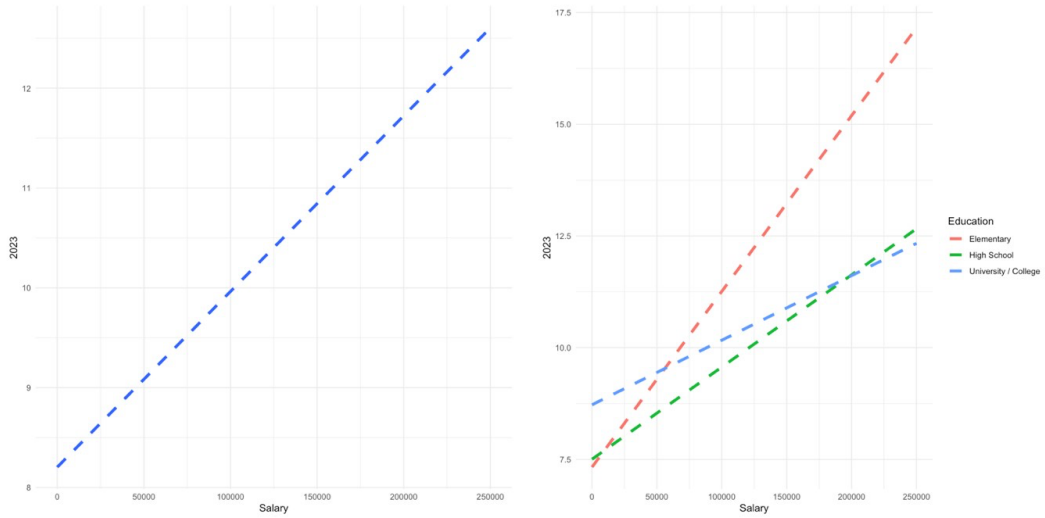
Nous avons ensuite mené une analyse exploratoire bivariée des données. Cette démarche avait notamment pour objectif de détecter la présence de caractéristiques qui auraient été susceptibles de biaiser les données et de nous guider pour la création de variables supplémentaires. Concernant les caractéristiques susceptibles de biaiser nos données, outre les constats que nous avons émis par rapport à la base de données “SocialNetworkUsage”, rien d’étrange n’a été remarqué. Concernant la création de variables, les analyses bivariées n’ont pas été d’une grande aide.

Cette démarche n’a toutefois pas été complètement inutile puisqu’elle nous a permis de justifier le choix de certaines de nos méthodes, notamment la méthode exhaustive avec les effets d’interaction. Effectivement, le tableau 10 semble démontrer la présence d’une relation linéaire entre le salaire et le montant des dons pour l’année 2023. Cependant, les tableaux suivants suggèrent la présence d’effets d’interaction entre le salaire et:

1. Le plus haut niveau d’éducation atteint (Tableau 11);
2. Le milieu de vie (rural, au centre-ville, en campagne) (Tableau 12)
3. La nombre d’ouverture de l’infolettre (Tableau 13)

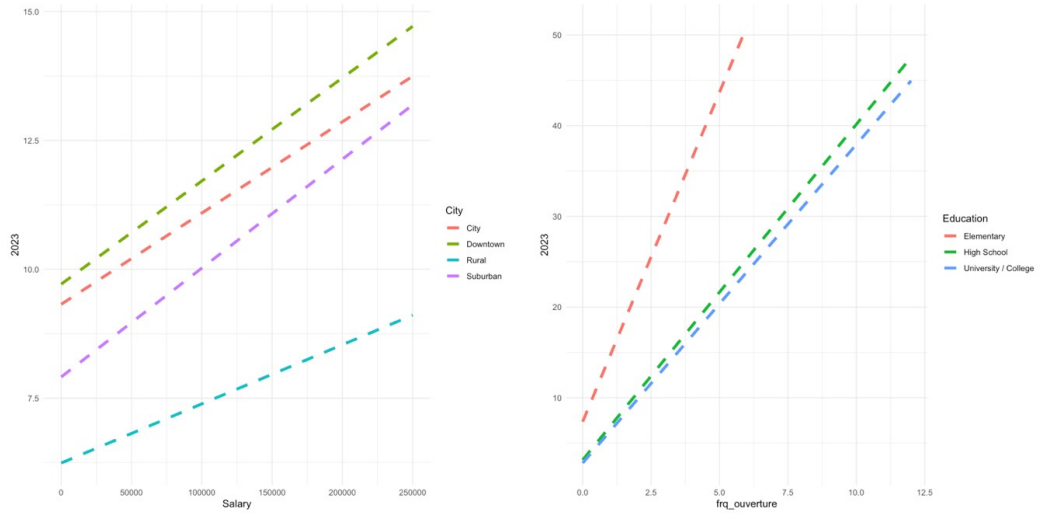
Ainsi que la présence d’un effet fixe relié au fait d’être une femme (Tableau 14).

Tableaux 10 & 11



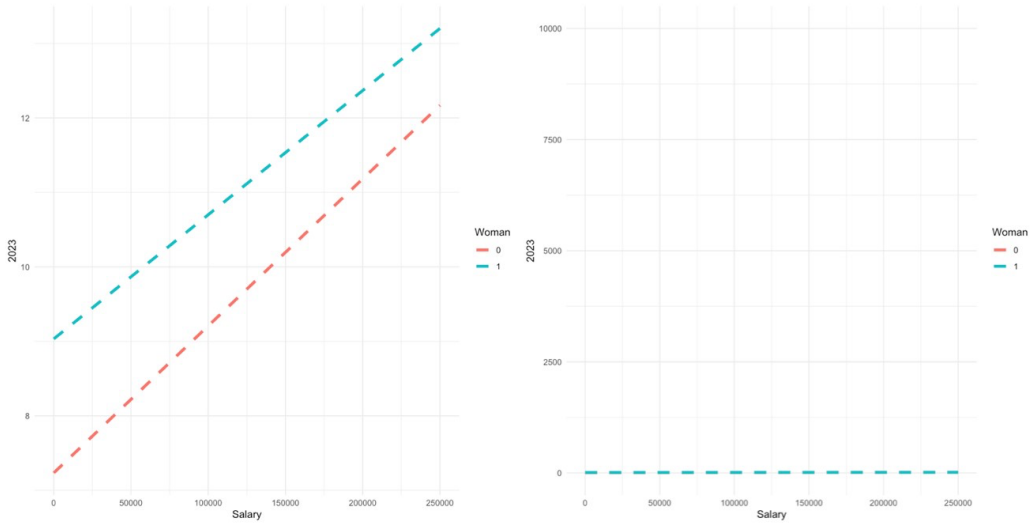
^Tableau 10 (à gauche) relation entre le salaire et le montant des dons en 2023;
Tableau 11 (à droite) effet d'interaction de la variable éducation sur la relation entre le salaire et le montant des dons en 2023.^

Tableaux 12 & 13



~Tableau 12 (à gauche) démontrant l'effet d'interaction de la variable City sur la relation entre le montant des dons en 2023 et le salaire;
Tableau 13 (à droite) démontrant l'effet d'interaction de la variable éducation sur la relation entre la fréquence d'ouverture de l'infolettre et le montant des dons e

Tableaux 14 & 15



~Tableau 14 (à gauche) démontrant l'effet fixe du sexe dans la relation entre le montant des dons en 2023 et le salaire; Tableau 15 (à droite) même chose que le tableau 14 mais montrant l'intervalle complet des dons.~

Nous n'avons cependant pas poussé notre investigation plus loin. Effectivement, l'étendue des dons est si grande (010,000$) qu'elle ne nous permettait pas de bien visualiser les effets des variables étant donné que les tableaux ne montraient pas de valeurs pour les dons plus élevés que 50$. En spécifiant cette étendue à la main, les effets devenaient alors imperceptibles, comme le montre le Tableau 15. En outre, l'exploration visuelle des données demeure assez limitée par rapport aux autres indices que pourront nous procurer la création de nos modèles.

Évaluation de la performance des modèles:

Tout d'abord, nous avons séparé notre échantillon en deux afin d'entraîner et de valider notre modèle. Nous avons ensuite défini un seed pour assurer la reproductibilité et utilisé un ratio de 0,99 pour diviser la base de données principale, soit 1% de l'échantillon dédié à l'entrainement des modèles (n = 10,000) et 99% pour leur validation (n = 990,000). Ce ratio a été sélectionné de manière relativement arbitraire, mais a surtout été influencé par la capacité limité de nos ordinateurs. Il n'est d'ailleurs pas disproportionné par rapport au ratio de l'exemple du Pr. Belzile qui mobilisait 1,000 observations pour l'entrainement et 100,000 observations pour la validation.

Les variables principales que nous avons choisi d'inclure dans notre modèle étaient les suivantes:

1. Woman: Le sexe biologique des membres de la base de données.
2. Age: L'âge des membres de la base de données.
3. Salary: Le salaire des membres de la base de données.
4. Education: Le plus haut niveau d'éducation obtenu par les membres de la base de données.
5. City: Le type de milieu urbain dans lequel habitent les membres de la base de données.
6. ancientete: aussi définie comme "Joined" dans le code, cette variable mal orthographiée représente le nombre d'années qui se sont écoulées entre l'adhésion des membres de la base de données à la fondation et aujourd'hui.
7. don_no2023: Le montant total des donations des membres de la base de données, de leur adhésion à la fondation jusqu'à l'année 2022.
8. frq_ouverture: Le nombre total de fois que les membres de la base de données ont ouvert l'infolettre pendant l'année 2023.
9. don_2023: (Variable dépendante) Le fait d'avoir fait ou non un don en 2023.
10. 2023: (Variable dépendante) Le montant donné en 2023.

Nous avons ensuite testé plusieurs méthodes déjà mentionnées dans l'introduction de ce travail:

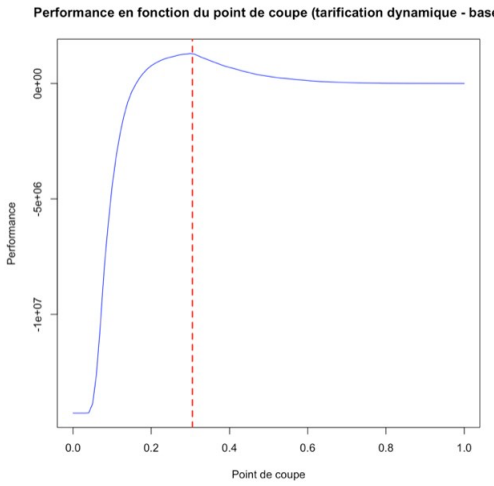
1. L'envoi de trousses à tous les membres de la fondation.
2. Les 8 variables de base;
3. Toutes les variables, incluant les termes quadratiques et les interactions d'ordre 2
4. Une sélection séquentielle avec l'AIC
5. Une sélection séquentielle avec le BIC
6. Une recherche exhaustive avec algorithme génétique (sélection selon BIC)
7. La méthode LASSO avec une pénalité optimale déterminée selon le critère de l'aire sous la courbe.

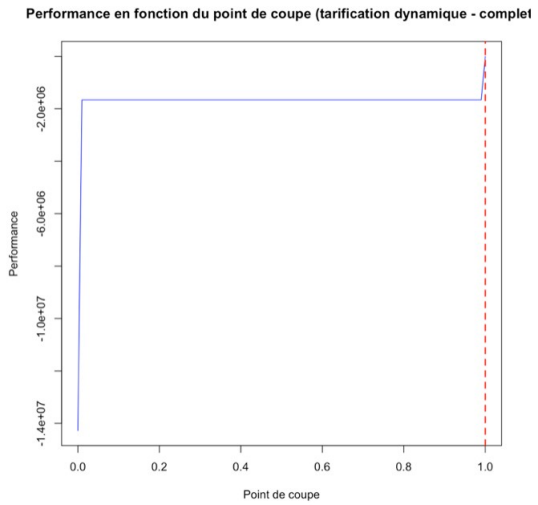
Pour évaluer notre modèle, il a fallu se baser sur une métrique pertinente au problème qu'il nous était demandé de résoudre. Dans notre cas, la fondation nous a mandaté pour maximiser ses profits et c'est donc sur la base des profits que nous baserons la sélection de notre modèle.

Compte tenu du fait que les trousses avaient une structure tarifaire dynamique, nous avons défini une fonction nous permettant de faire des prédictions binaires s'adaptant à cette structure. Cette fonction nous a permis de calculer la performance de notre modèle en prédisant si les individus allaient faire un don ou non. Nous avons ensuite calculé le nombre d'envois optimal, en prenant en compte le fait que les gens effectuent un don et le montant de ce don au regard de la tarification dynamique des trousses. Nous avons donc pu établir un indice de performance pour nos modèles calculé comme la somme des dons reçus moins les coûts dynamiques, autrement dit, les profits nets.

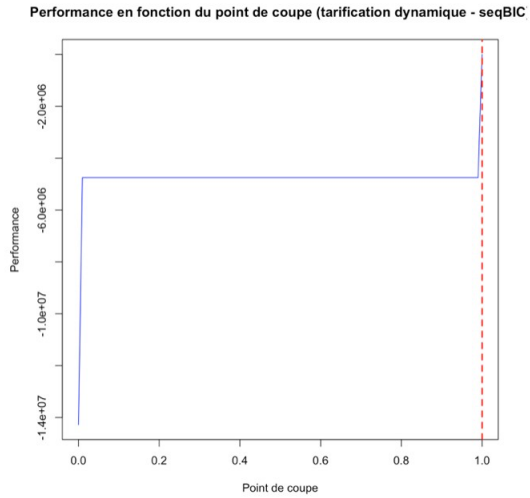
Cet indice de performance a ensuite été calculé pour différents seuils de décision, de 0 à 1, avec des pas de 0,01. Pour sélectionner le meilleur point de coupe (seuil de décision) nous avons ensuite employé une méthode d'optimisation pour tester quel seuil de décision nous permettait d'obtenir le plus grand indice de performance (i.e. les profits nets). Les graphiques suivants montrent les points de coupe optimaux des différents modèles que nous avons testés :

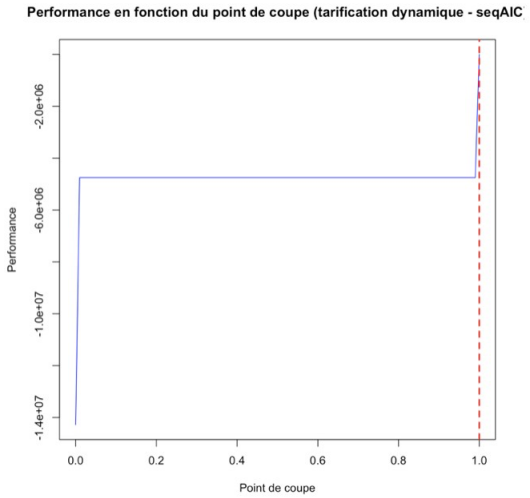
Tableaux 16 & 17



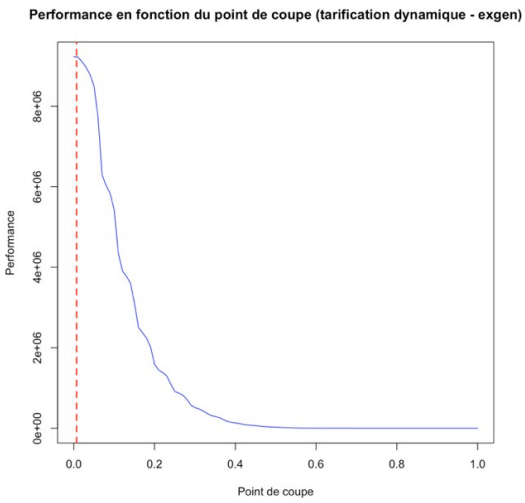


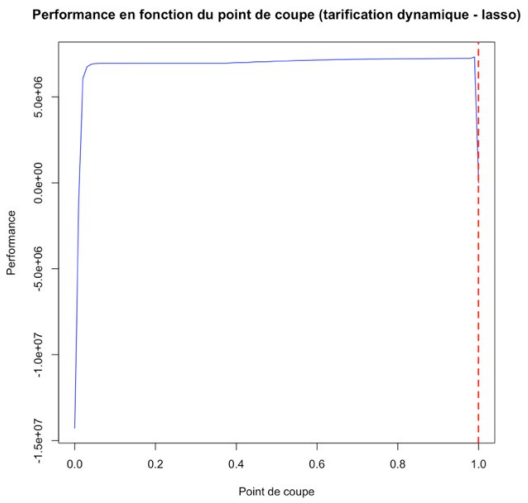
Tableaux 18 & 19





Tableaux 20 & 21





Le tableau X nous permet de voir de manière plus détaillée la performance des différents modèles. En temps normal, nous choisirions probablement un modèle le plus exact au regard d'un critère quelconque. Par exemple, si nous cherchions à connaître l'efficacité d'un test médical pour une maladie infectieuse, il serait probablement utile de ne laisser aucun malade indétecté puisqu'il pourrait infecter d'autres personnes. Dans ce cas, le choix du modèle aurait sûrement été fait sur la base de la sensibilité du test. Cependant, comme nous l'avons dit plus haut, l'indice de performance qui compte pour nous est le profit net.

Tableau 22

Résultats des modèles						
Modèle	No. variables	Pt. coupure	Taux bonne classif.	Sensibilité	Spécificité	Profit
Baseline	NA	NA	0.17	1.00	0.00	−15,043,730.00
base	11.00	0.31	0.85	0.41	0.88	1,308,985.00
complet	265.00	1.00	0.72	0.18	0.87	−1,660,000.00
seqAIC	261.00	1.00	0.57	0.17	0.88	−4,749,830.00
seqBIC	261.00	1.00	0.57	0.17	0.88	−4,749,830.00
exgen_modele	3.00	0.01	0.50	0.50	0.50	−14,276,635.00
lasso	15.00	1.00	0.93	1.00	0.93	7,410,535.00

Or, c'est le modèle basé sur la méthode Lasso qui nous permet de maximiser nos profits. En sélectionnant les individus pour lesquels nous avons une confiance de 99,95296% qu'ils feront un don (arrondi à 1.00 dans le tableau), nous parvenons à obtenir 7,410,535$ de profits nets. Pour parvenir à de tels profits, le modèle suggère d'envoyer des trousses à 73,093 personnes. Cela signifie qu'en moyenne, nous pouvons nous attendre à ce que les personnes à qui la trousse est envoyée fassent, en moyenne, un don de 101.39$.

En somme, cette méthode nous semble plutôt satisfaisante, car si la fondation avait envoyé des trousses à tout le monde, elle aurait accusé un déficit financier de 15,043,730$. De même, le modèle obtenu à partir des variables de base, qui est le deuxième meilleur modèle, ne permet de faire que 1,308,985$ de profits. Le modèle Lasso a donc une performance 7x meilleure à celle de la deuxième meilleure option.

Annexe:

Tableau 23: matrice de confusion des différents modèles:

Modèle	VN	FN	FP	VP
base	814077	35117	115924	24882
complet	672803	176391	101161	39645
seqAIC	495423	353771	69698	71108
seqBIC	495423	353771	69698	71108
exgen_modele	849194	140806	849194	140806
lasso	849194	0	67713	73093