PROJET SCORING

Marketing Bancaire

NDOYE Alioune
GADJI Mame Astou
BALDE Mariama
Kindiman

Table des matières

Introduction

Les données sur lesquelles on réalise notre étude sont liées aux campagnes de marketing direct (appels téléphoniques) d'une institution bancaire portugaise.
Ces campagnes de marketing étaient basées sur des appels téléphoniques. La plupart du temps, plus d'un contact avec le même client est nécessaire, afin de savoir si oui ou non, ce dernier va souscrire à un dépôt à terme.

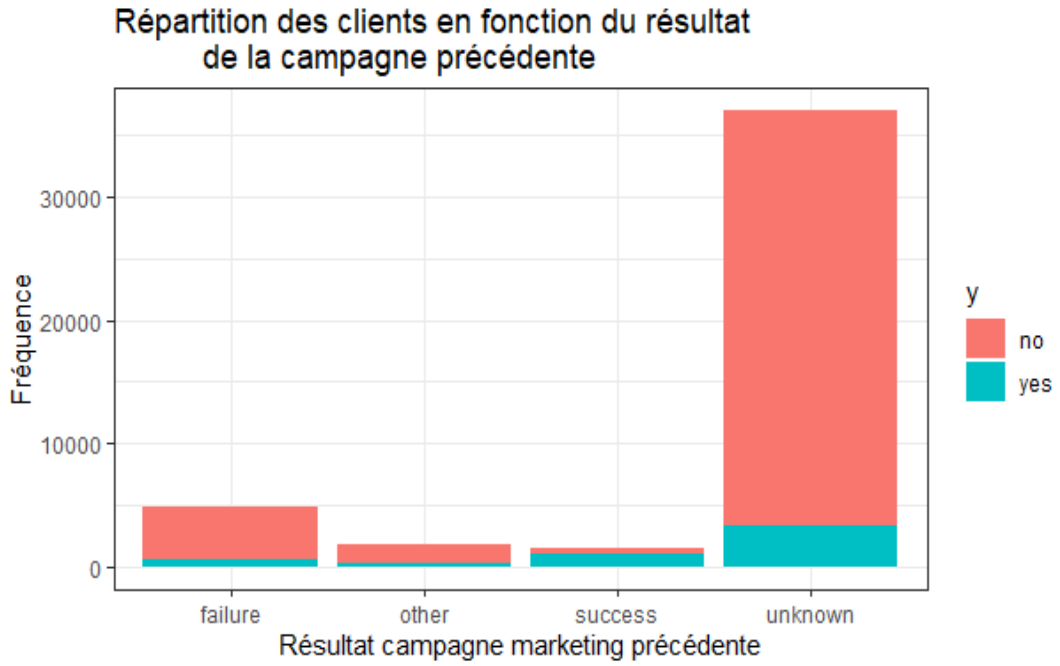
Notre base de données est composée de 17 variables comme l'âge du client, le type d'emploi, la situation matrimoniale, le niveau d'éducation, la situation de crédit du client. Elle contient aussi des données liées à la dernière campagne elle-même : le canal de contact, le mois et jour du contact ainsi que le résultat de la dernière campagne marketing.
La variable cible (y) est binaire et nous dit si le client accepte ou non de souscrire un dépôt à terme.

Nous allons pour réaliser notre étude passer par plusieurs étapes. D'abord, nous allons faire une analyse descriptive de la base de données afin de savoir comment les variables explicatives sont liées avec la variable cible. Ensuite, nous allons faire trois modèles de classifications suivant ces variables. Enfin, nous comparerons les résultats obtenus par ces modèles.

I – Analyse descriptive de la base de données

Les données sont principalement composées de variables catégorielles : job, marital, education, default, housing, loan, contact, month, day, campaign, poutcome. Il y a également des variables numériques comme duration, previous ou encore pdays. Vous trouverez en annexe la description de ses variables.

Voici la répartition de la variable cible (y). On constate qu'il y a beaucoup plus de réponses négatives.

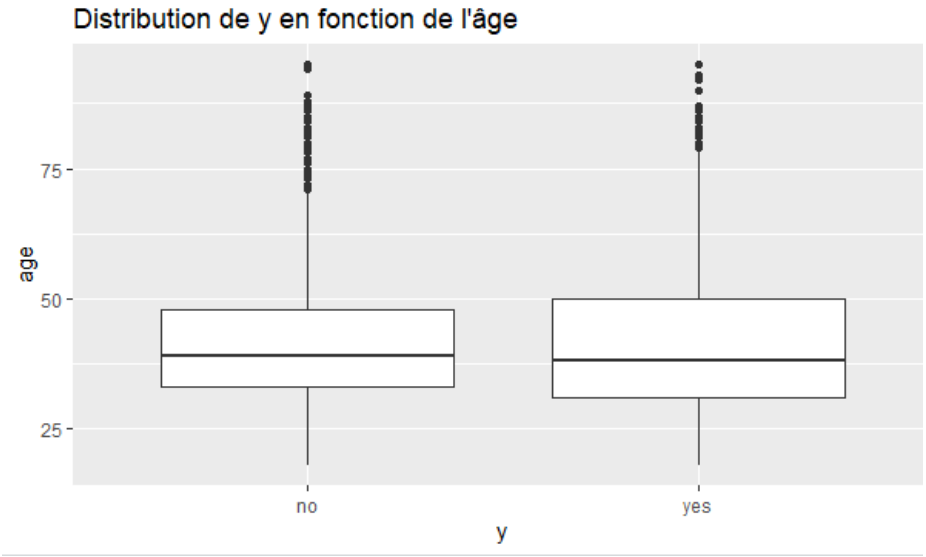


no	yes
39922	5289

Nous allons étudier le lien de quelques variables explicatives avec y en s'aidant de représentati on graphique.

1 – Répartition des clients en fonction du résultat de la campagne précédente

Grâce à ce graphique, nous pouvons voir que le résultat de la campagne précédente est pour la plupart représenté par les individus « unknown ». Parmi ces individus, la grande majorité à



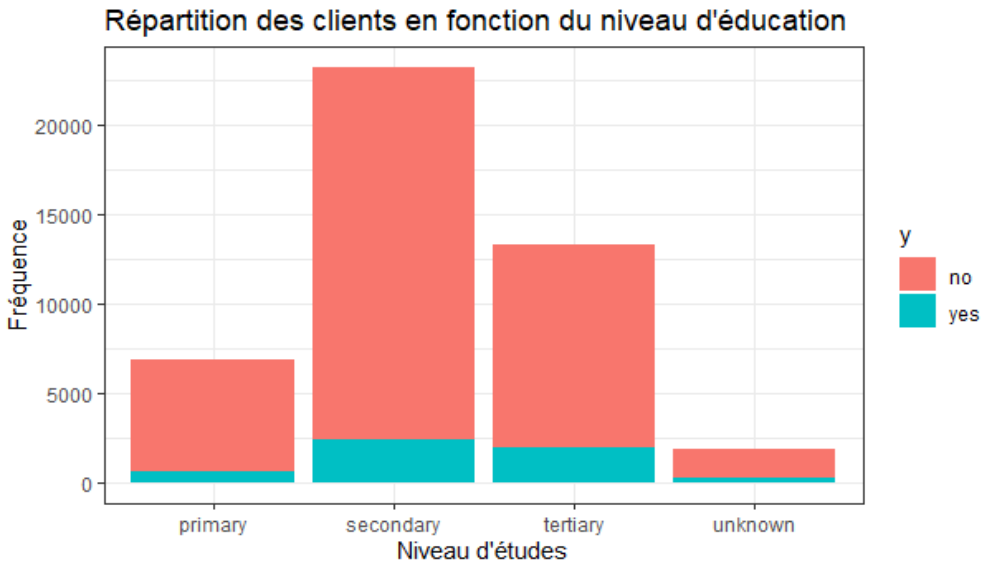
refuser de souscrire au dépôt à terme. C'est dans la catégorie « success » que l'on observe le plus faible taux de réponse négative et que l'on observe le plus grand taux de souscription.

2– Distribution de y en fonction de l'âge

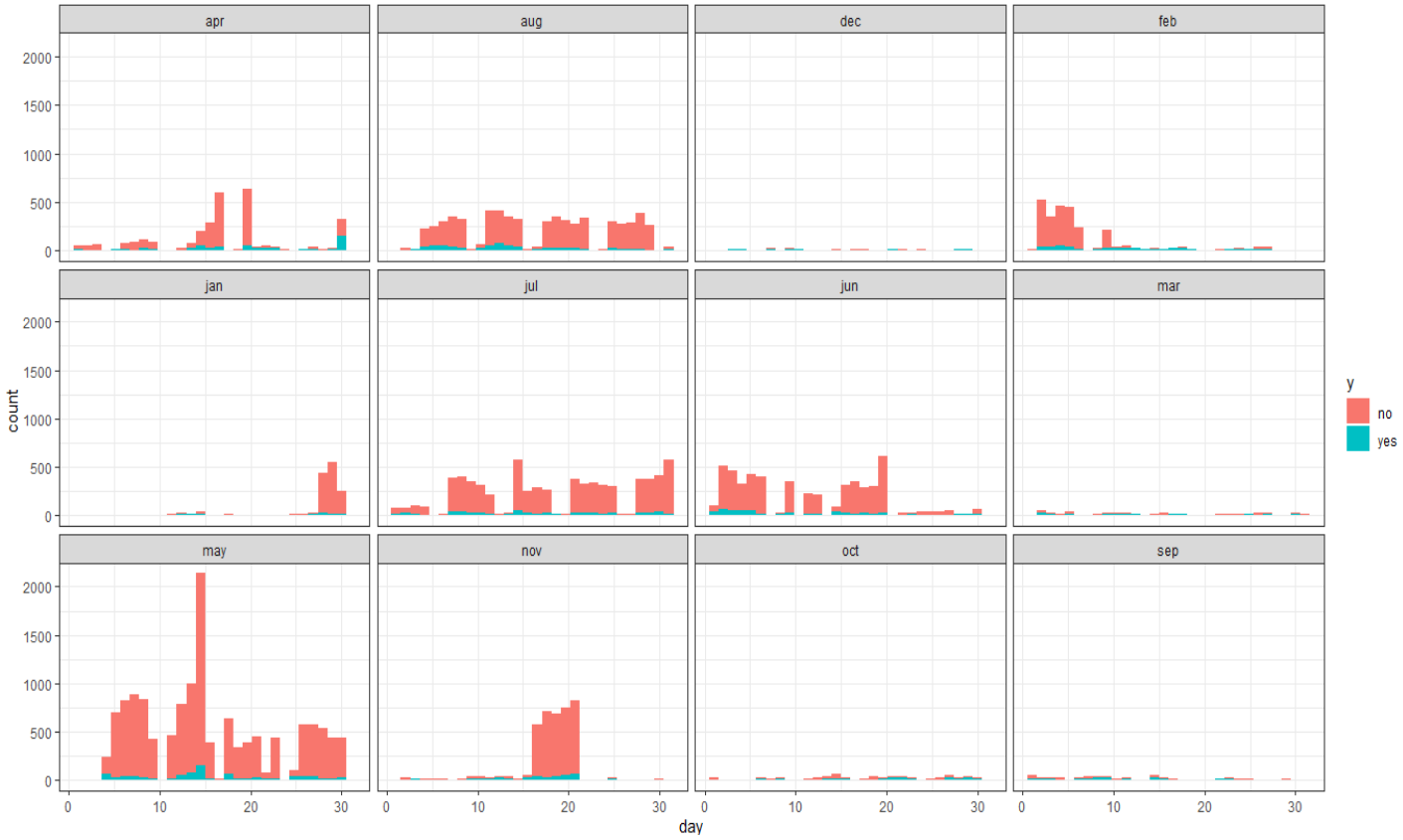
Nous pouvons dire que l'âge n'a pas l'air d'avoir d'effet sur la variable y. Les médianes pour les catégories de la variables cibles étant égales.

3 – Répartition des clients en fonction du niveau d'éducation

Nous pouvons constater que la catégorie « secondary » est la plus représentée suivi de « tertiary ». Le nombre de personne ayant accepté de souscrire au dépôt à terme suivant ces catégories est relativement bas.

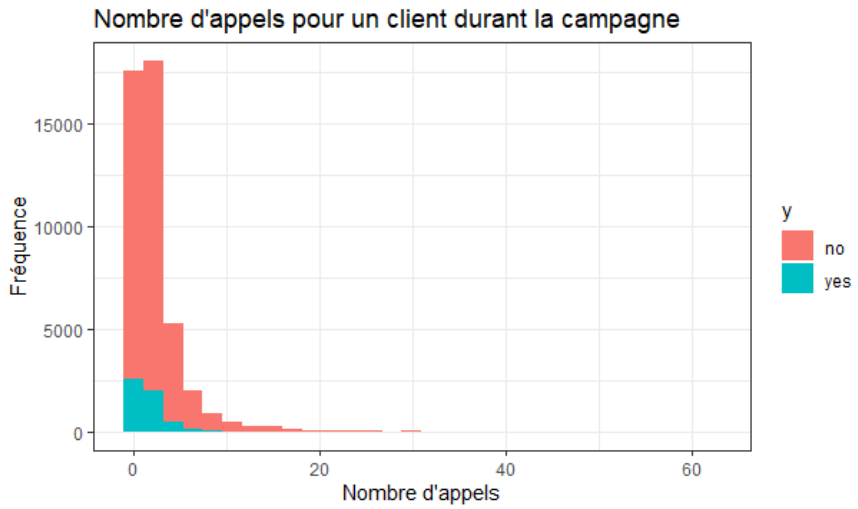


4 – Périodes des campagnes de marketing



On peut voir que les campagnes de marketing se font le plus à partir du mois d'avril jusqu'à août. Pour les autres mois il y a des appels même si le nombre est moindre. On a le plus grand nombre de souscription en fin avril.

5 - Le volume d'appels par client



Le nombre d'appels par client pour une campagne est en moyenne égale à 3. La médiane est égale à 2 mais on voit que pour certaines personnes, nous pouvons avoir beaucoup d'appels. Ainsi, le nombre maximum d'appels est 63.

II – Modèles de classification

L'objectif d'une méthode de classification est de diviser en un ou plusieurs sous-ensembles homogène un ensemble de données étudiées. Les membres d'une classe ressemblent plus aux autres membres de la même classe qu'aux membres d'une autre classe. Nous allons étudier trois modèles de classifications avant de comparer les résultats obtenus : l'analyse discriminante, la régression logistique et les arbres de décision.

1 – La régression logistique

La régression logistique est généralement définie comme un modèle statistique qui permet d'étudier les relations entre un ensemble de variables Xi et une variable qualitative Y. En effet c'est un modèle linéaire généralisé qui utilise une fonction logistique comme fonction de lien.

Elle permet également de déterminer la probabilité de survenance d'un évènement ou non à partir de l'optimisation des coefficients de régression. Elle est ainsi comprise entre 0 et 1.

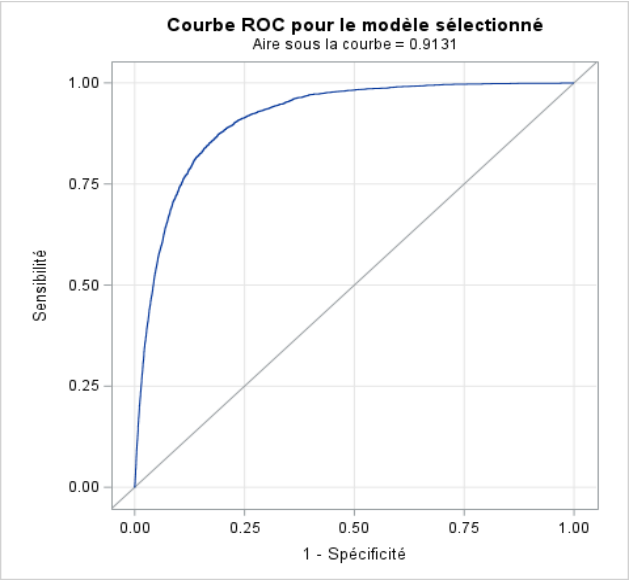
Si la valeur prédite est supérieure à un seuil, l'événement est susceptible de se produire, alors que lorsque cette valeur est inférieure au même seuil, il ne l'est pas.

Nous avons créé deux modèles de régression logistique : une, sans interaction entre les variables et une autre prenant en compte les interactions entre les différentes variables.

A – Modèle sans interaction entre les variables explicatives

Dans ce modèle, les variables sélectionnées par la méthode Backward sont : campaign, contact, day, duration, education, housing, job, marital, loan, month et poutcome. Ces variables sont significatives à 1%. Certaines variables, « age, default, pdays, previous et balance » ont été supprimés du modèle.

En se basant sur le tableau d'analyse des effets de type 3, on peut dire que « duration » a le plus grand pouvoir explicatif avec un Khi-2 de Wald égale 2995, ensuite on a « poutcome » puis « month ».



Association des probabilités prédites et des réponses observées			
Pourcentage concordant	91.3	D de Somers	0.826
Pourcentage discordant	8.7	Gamma	0.826
Pourcentage lié	0.0	Tau-a	0.171
Paires	103484038	c	0.913

Matrice de confusion apprentissage Modèle 1

La procédure FREQ

Fréquence Pourcentage Pct de ligne Pct de col.	Table de y par I_y		
	I_y(Dans : y)		
y	no	yes	Total
no	27259 86.13 97.54 92.01	687 2.17 2.46 33.94	27946 88.30
yes	2366 7.48 63.89 7.99	1337 4.22 36.11 66.06	3703 11.70
Total	29625 93.60	2024 6.40	31649 100.00

Matrice de confusion test Modèle 1

La procédure FREQ

Fréquence Pourcentage Pct de ligne Pct de col.	Table de y par I_y		
	I_y(Dans : y)		
y	no	yes	Total
no	11653 85.92 97.30 91.88	323 2.38 2.70 36.75	11976 88.31
yes	1030 7.59 64.94 8.12	556 4.10 35.06 63.25	1586 11.69
Total	12683 93.52	879 6.48	13562 100.00

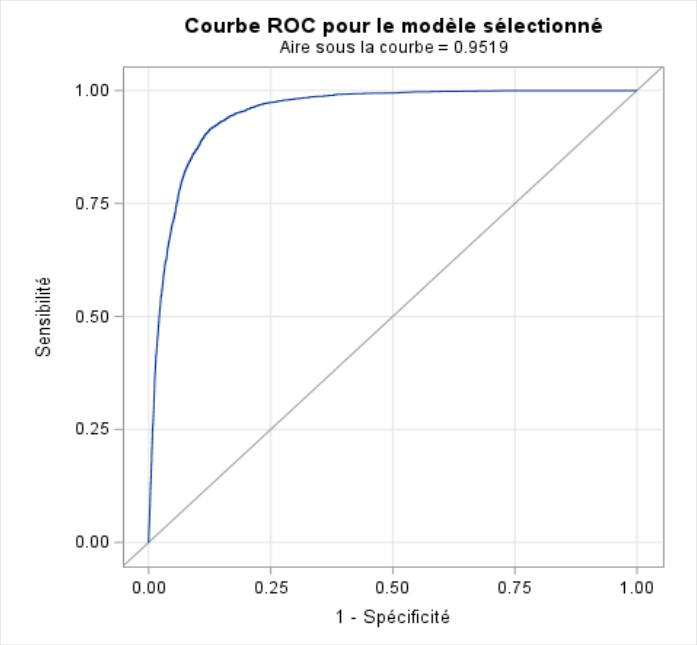
Le pourcentage de concordant pour ce modèle est de 91.3 ce qui est très bien. L'aire sous la courbe ROC est égal à 0.913. Plus cet indicateur est proche de 1, meilleur est le pouvoir discriminant du modèle.

Grâce aux deux matrices de confusions (sur la base d'apprentissage et sur la base test), on peut voir que le modèle est stable sachant que les pourcentages dans chaque partie de la matrice correspondent.

B – Modèle avec interactions entre les variables explicatives

Il peut arriver que certaines variables soient reliées entre elles donc il peut être bien de modéliser ces interactions. Nous avons gardé que les interactions fortes dans le modèle en fixant le degré de signifiance à 1%. Contrairement au modèle sans interaction, ici toutes les

variables sont significatives et on constate que les variables interagissent toutes entre elles ce qui augmente naturellement la qualité du modèle.



Association des probabilités prédites et des réponses observées			
Pourcentage concordant	95.2	D de Somers	0.904
Pourcentage discordant	4.8	Gamma	0.904
Pourcentage lié	0.0	Tau-a	0.187
Paires	103484038	c	0.952

Matrice de confusion apprentissage Modèle 2

La procédure FREQ

Fréquence Pourcentage Pct de ligne Pct de col.	Table de y par I_y		
	I_y(Dans : y)		
y	no	yes	Total
no	27109 85.66 97.00 94.62	837 2.64 3.00 27.91	27946 88.30
yes	1541 4.87 41.61 5.38	2162 6.83 58.39 72.09	3703 11.70
Total	28650 90.52	2999 9.48	31649 100.00

Matrice de confusion test Modèle 2

La procédure FREQ

Fréquence Pourcentage Pct de ligne Pct de col.	Table de y par I_y		
	I_y(Dans : y)		
y	no	yes	Total
no	11426 84.31 95.47 93.15	542 4.00 4.53 42.11	11968 88.31
yes	840 6.20 53.00 6.85	745 5.50 47.00 57.89	1585 11.69
Total	12266 90.50	1287 9.50	13553 100.00

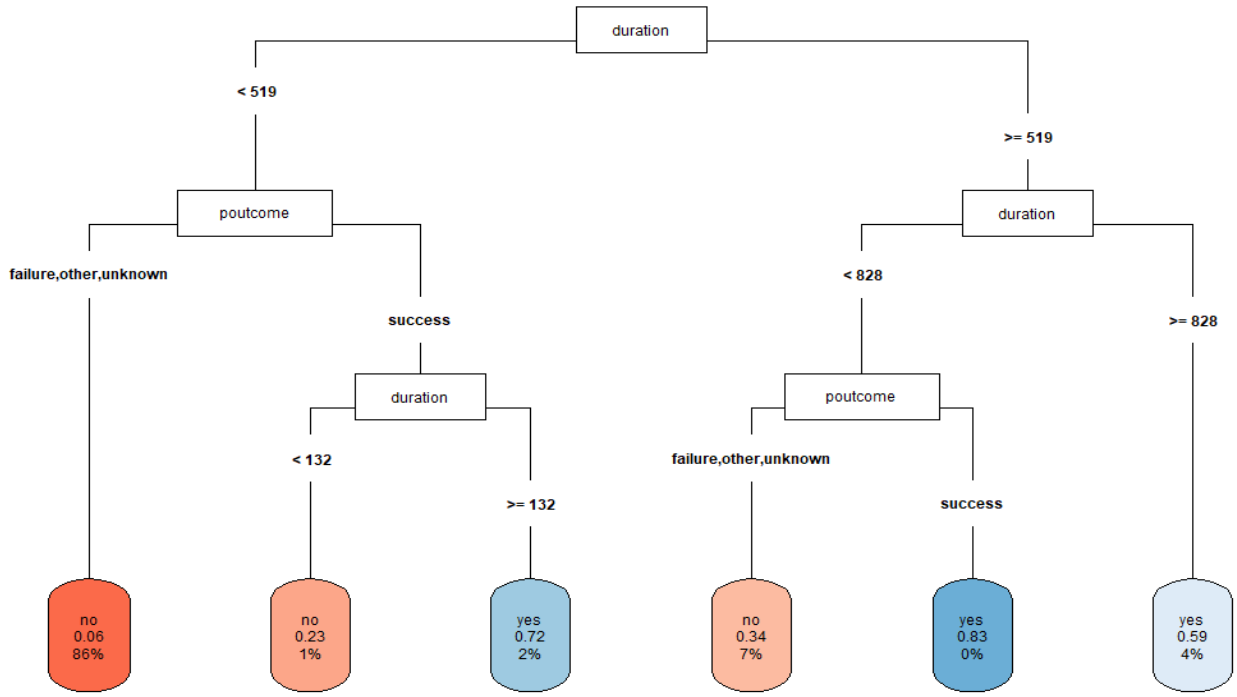
L'indice de Gini, représenté par le Somer's D indique la qualité du modèle. Dans le modèle sans interaction, il est à 0.82. Il passe à 0.90 dans le modèle avec interactions. Le pourcentage de concordant pour ce modèle est de 95.2. On observe ainsi une évolution entre les deux modèles. L'aire sous la courbe ROC est égal à 0.952.

Grâce aux deux matrices de confusions (sur la base d'apprentissage et sur la base test), on peut voir que le modèle est stable.

Pour finir avec la régression logistique, on pourrait retenir le modèle avec interactions sachant qu'on a plus de qualité. Le problème étant qu'il est assez complexe, on pourrait choisir celui sans interaction sachant qu'il a aussi une bonne qualité prédictive mais est aussi plus simple.

2 – Arbres de décisions

Les arbres de décision permettent de résoudre des problèmes de discrimination en segmentant de façon progressive un échantillon en vue de la prédiction d'un résultat. L'arbre obtenu avec la base d'apprentissage est le suivant :



	importance
duration	995.6580295
poutcome	610.2668281
contact	1.9314884
pdays	1.3716930
default	0.3862977
balance	0.1330343
campaign	0.1330343
previous	0.1330343

On constate que duration et poutcome sont les variables les plus importantes pour la construction de cet arbre avec des grandes valeurs (995 et 610) tandis que le reste des variables a des valeurs de l'importance aux alentours de 1 ou 0. La durée de l'appel est jugée comme plus importante ce qui parait logique car plus l'appel sera long plus on aura des chances de convaincre le potentiel client qu'on appelle. On a ainsi une première séparation de la base entre les observations avec une durée d'appel inférieure à 519 secondes ce qui correspond à 8 minutes environ, et les appels dont la durée est supérieure ou égale à 519 secondes. A partir de là, une deuxième séparation est faite à partir de poutcome d'un côté et duration de l'autre. Au final, nous obtenons 6 branches.

L'efficacité de cette méthode est appréciée en évaluant la matrice de confusion construite après avoir utilisée la base de test sur l'arbre. Le résultat obtenu est le suivant :

Y	Y_predict		Total
	No	Yes	
No	11636 97%	340 3%	**11976** **100%**
Yes	1036 65%	550 35%	**1586** **100%**
Total	**12672**	**890**	**13562**

3 – L'analyse discriminante

La méthode d'analyse discriminante est une méthode dite prédictive. En effet elle permet de faire des prédictions et d'expliquer l'appartenance d'un individu à une classe (groupe) prédéfinie sur la base de ses caractéristiques mesurées par des variables prédictives.

On cherche ainsi un ensemble d'axes qui résument au mieux la distance existante entre groupes d'observations. Dans ce nouveau repère, les points des groupes doivent être aussi distants les uns des autres que possibles, et aussi proches que possible les uns des autres au sein d'un même groupe. On parle ainsi de maximiser la variance inter-classe et de minimiser celle intra-classe.

Pour cet algorithme, nous avons, au préalable, créé des variables indicatrices pour les variables catégorielles sachant que l'instruction VAR de la PROC DISCRIM ne peut contenir que des

valeurs numériques. Cela est dû au fait que cette procédure se base sur l'Analyse en Composantes Principales (ACP) qui peut se réaliser que sur des variables quantitatives.

Suite à cela, nous avons utilisé la PROC STEPDISC qui permet de sélectionner les variables significatives qu'on mettra ensuite dans l'analyse discriminante. Suite à cette étape, nous avons 38 variables qui ont été supprimées.

Après, nous avons fait l'analyse discriminante avec les variables restantes à savoir: contact_cellular, contact_telephone, day_2, day_5, day_6, day_7, day_11, ay_17, day_18, day_19, day_20, day_27, day_30, duration, education_tertiary, housing_no, job_admin_, job_retired, job_student, loan_no, marital_married, month_apr, month_aug, month_feb, month_jan, month_jul, month_jun, month_mar, month_may, month_nov, poutcome_success et previous.

L'option MANOVA dans la PROC DISCRIM permet d'avoir une sortie sur les tests de significativité globale du modèle. Nous pouvons nous aider du Wilks' Lambda qui a comme hypothèse nulle la superposition des classes de la variable y. On rejette ainsi cette hypothèse car la p-value est très petite ce qui signifie que le modèle de classification est bon.

Une partie des résultats de la PROC DISCRIM se trouve ci-dessous. Ainsi, on peut constater que la fonction discriminante utilisée est plutôt stable entre la base d'apprentissage et celle de test. Nous pouvons voir cela grâce aux matrices de confusions obtenues sur ces deux bases de données. On remarque que le taux d'erreur est élevé pour les observations de la modalité « yes » de la variable y.

La procédure DISCRIM

Statistiques multivariées et statistique F exacte					
S=1 M=34.5 N=15787.5					
Statistique	Valeur	Valeur F	DDL num.	DDL den.	Pr > F
Wilks' Lambda	0.68863805	201.09	71	31577	<.0001
Pillai's Trace	0.31136195	201.09	71	31577	<.0001
Hotelling-Lawley Trace	0.45214166	201.09	71	31577	<.0001
Roy's Greatest Root	0.45214166	201.09	71	31577	<.0001

La procédure DISCRIM
Synthèse de classification pour données de calibration : B.TRAIN_DUMMY
Synthèse de validation croisée utilisant Fonction discriminante quadratique

Nombre d'observations et pourcentage classifiés dans y			
De y	no	yes	Total
no	25324 90.62	2622 9.38	27946 100.00
yes	1803 48.69	1900 51.31	3703 100.00
Total	27127 85.71	4522 14.29	31649 100.00
A priori	0.883	0.117	

La procédure DISCRIM
Synthèse de classification pour données de test : B.TEST_DUMMY
Synthèse de classification utilisant Fonction discriminante quadratique

Profil d'observation pour les données de test	
Nombre d'observations lues	13562
Nombre d'observations utilisées	13562

Nombre d'observations et pourcentage classifiés dans y			
De y	no	yes	Total
no	10833 90.46	1143 9.54	11976 100.00
yes	770 48.55	816 51.45	1586 100.00
Total	11603 85.56	1959 14.44	13562 100.00
A priori	0.883	0.117	

Conclusion

On voit que les modèles de classifications que l'on a utilisés prédisent mieux les observations avec la modalité « no » pour la variable y. Pour l'analyse discriminante, on a 91% de bonnes prédictions pour cette modalité, 97% avec l'arbre de décision et 85% avec la régression logistique. Pour ce qui est des prédictions de la modalité « yes », représentant le fait que l'individu va accepter de souscrire au dépôt à terme, les modèles sont moins précis. En effet, pour l'analyse discriminante, nous avons 51% de bonnes prédictions, 35% pour l'arbre de décision et la régression logistique (modèle sans interaction). Le modèle avec interaction de son côté a 47% de bonnes prédictions. Cette difficulté que les modèles ont pour prédire la modalité « yes » est peut-être dû au fait qu'elle est moins représentée comparée à « no ». En effet, la variable y est composée à 88% par cette dernière modalité.

Sachant que l'objectif de cette étude est de prédire la souscription au dépôt à terme c'est-à-dire la modalité « yes » de y, on peut dire que l'analyse discriminante répond mieux à notre besoin lorsqu'on se base sur les résultats obtenus à partir de la base de test des modèles. Cependant, il faut relever que le principal inconvénient de l'analyse discriminante est le fait qu'elle suppose que les prédicteurs suivent une loi normale multivariée. De ce fait, elle ne supporte pas les variables catégorielles (d'où la création des variables indicatrices pour y pallier) contrairement à la régression logistique / arbres de décisions qui sont des modèles de probabilité direct. Sur ce point, nous pouvons dire que l'analyse discriminante est en retard. En temps de calcul, la régression logistique bat les records lorsqu'on a une grande base de données.

Pour finir, on retiendra aussi que l'arbre de décision est une méthode efficace. Elle permet de traiter les grandes bases de données avec rapidité et sans difficulté. C'est une méthode non paramétrique et robuste face aux données aberrantes et manquantes. Cependant elle a un problème de stabilité sur les petites bases de données et à la difficulté concernant la détection des interactions entre les variables explicatives.

Bibliographie

Base de données et description des variables,

https://data.world/data-society/bank-marketing-data

La classification,
https://fr.wikipedia.org/wiki/Analyse_des_donn%C3%A9es#:~:text=Le%20but%20de%20la%20classification,membres%20d'une%20autre%20classe, Wikipédia.

La régression logistique,

https://datascientest.com/regression-logistique-quest-ce-que-cest, Datascientest.

L'analyse discriminante,
https://fr.wikipedia.org/wiki/Analyse_discriminante_lin%C3%A9aire, Wikipédia.

L'analyse discriminante, https://od-datamining.com/knwbase/lanalyse-discriminante-expliquee-a-ma-fille/, OD-Datamining.

ANNEXE

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Input variables:

bank client data:

1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')

related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

social and economic context attributes:

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')