Rapport Online Shoppers

AMATO Grégoire, PHILIPPE Camille, SANCHEZ Albane

30 avril 2020

Sommaire

1 Introduction

Ce jeu de données contient 12330 observations, chacune représentant une session [1] indépendantes entre elles, ainsi chaque session appartient à un utilisateur différent, les sessions ont été collectées sur une période d'un an afin d'éviter toute tendance dûe à des campagnes (publicité, promotions...), des jours de fêtes etc. Le but de cette étude est de déterminer le comportement général d'un visiteur, particulièrement les acheteurs, sur le site et ainsi d'en déduire une tendance afin d'améliorer les ventes.

Le jeu de données contient 18 variables :

Des variables quantitatives relatives à la visite de pages et au temps passé dessus (en secondes) :

- **Administrative, Administrative duration** : gestion de compte
- **Informational, Informational duration** : informations, communication du site
- **Product related, Product related duration** : produits en vente

D'autres variables quantitatives :

- **Bounce rate** : moyenne du "taux de rebond", c'est-à-dire le pourcentage de visiteurs qui entrent sur une page du site et la quitte immédiatement sans effectuer d'autres requêtes sur la page.
- **Exit rate** : moyenne du "taux de sortie", c'est-à-dire le nombre de sorties par page, ainsi tous les "bounces" sont des "exits" mais pas l'inverse. Ainsi, dès qu'un visiteur quitte le site, on regarde à partir de quelle page il l'a quitté.
- **Page value** : moyenne de la valeur de la page représente la valeur moyenne d'une page qui, lors d'une session, a amené à une transaction. Par exemple, un visiteur arrive sur la page A du site et décide alors d'aller sur la page D pour effectuer un achat. La page B a alors une valeur de 0 euros car il n'y est pas passé mais la page A aura une valeur de x euros selon le montant de la transaction.
- **Special day** : "jour de fête" indique la proximité entre la date de visite et le jour d'une fête. La valeur est déterminée selon la dynamique du e-commerce comme la durée entre la commande et la livraison. Par exemple, pour la Saint-Valentin, la valeur est différente de 0 entre le 2 et le 12 Février, est nulle avant et après le 14 février, sauf si une autre fête est proche, et prend pour valeur maximale 1 le 8 février [2].

Des variables qualitatives :

- **Operating systems** : système d'exploitation du visiteur.
- **Browser** : navigateur internet du visiteur.
- **Region** : région géographique à partir de laquelle la session a été lancée par le visiteur.
- **Traffic type** : source de trafic par laquelle le visiteur est entré sur le site (par exemple bannière publicitaire, accès direct, sms...)
- **Visitor type** : type de visiteurs, tel que "Nouveau visiteur", "Visiteur régulier", "Autre"
- **Month** : mois de la visite.

Des booléens :

- **Weekend** : indique si la date de la visite est un week-end ou non.
- **Revenue** : indique si la visite s'est terminée par une transaction ou non.

Notre objectif sera ici de prédire le comportement des clients fréquentant un site, afin de prévoir s'il vont conclure leur visite par un achat ou non.

1. Entrer sur le site, (potentiellement) y effectuer des actions, puis quitter le site.
2. Le jeu de données provient de Turquie, les fêtes sont donc des fêtes Turques.

2 Régression Logistique

Ce modèle logit décrit la relation entre la variable Revenue, décrivant l'issue de la visite du client sur le site, achat ou non.
En l'occurence, ce modèle économétrique n'est pas le meilleur, en effet, certains coefficients ne sont pas représentatifs. Nous allons donc nous servir de la fonction `step` afin de trouver un modèle le plus précis possible au loyen de la minimisation de l'AIC.

Table 1

	Dependent variable :
	Revenue
ExitRates	-20.621^{***}
	(1.837)
PageValues	0.077^{***}
	(0.003)
MonthDec	-0.604^{***}
	(0.198)
MonthFeb	-1.708^{***}
	(0.641)
MonthJul	-0.036
	(0.242)
MonthJune	-0.238
	(0.299)
MonthMar	-0.677^{***}
	(0.197)
MonthMay	-0.633^{***}
	(0.185)
MonthNov	0.565^{***}
	(0.178)
MonthOct	-0.161
	(0.223)
MonthSep	-0.030
	(0.232)
TrafficType	0.002
	(0.009)
VisitorTypeOther	-0.695
	(0.583)
VisitorTypeReturning_Visitor	-0.189^{**}
	(0.090)

Weekend	0.097
	(0.078)
Constant	−1.451***
	(0.190)
Observations	9,865
Log Likelihood	−2,950.370
Akaike Inf. Crit.	5,932.740
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Le modèle ci-dessus est le modèle le plus précis, grâce à la technique de minimisation de l'AIC.
Les coefficients de la regression logistique n'étant pas interprétables de la sorte, nous allons étudier les odds ratio correspondants.

	OR	2.5 %	97.5 %
(Intercept)	0.2342455	0.1600517	0.3372076
ExitRates	0.0000000	0.0000000	0.0000000
PageValues	1.0803823	1.0750547	1.0858948
MonthDec	0.5467078	0.3730215	0.8107226
MonthFeb	0.1812012	0.0411244	0.5465373
MonthJul	0.9650039	0.5996258	1.5500901
MonthJune	0.7884591	0.4324848	1.4009105
MonthMar	0.5081822	0.3471347	0.7528847
MonthMay	0.5311614	0.3721442	0.7697714
MonthNov	1.7600087	1.2535107	2.5161105
MonthOct	0.8515309	0.5512405	1.3219238
MonthSep	0.9701890	0.6159676	1.5312915
TrafficType	1.0018514	0.9840352	1.0196043
VisitorTypeOther	0.4989235	0.1417236	1.4213160
VisitorTypeReturning_Visitor	0.8275361	0.6940803	0.9893943
WeekendTRUE	1.1016416	0.9448490	1.2819880

Ces odds ratio nous permettent de connaître le lien entre les variables explicatives et le fait de finir la visite par un achat.
Un odd ratio égal à 1 indique qu'il y a indépendance entre la variable concernée et le fait d'acheter. Si la valeur est supérieure, cela représente une corrélation positive, l'inverse une corrélation négative.

Dans notre cas, nous remarquons que le parmi les mois, c'est le mois de Novembre qui est le seul à avoir un impact positif sur le fait d'acheter ou non.
`PageValues` a une corrélation positive, relativement faible ainsi que `TrafficType` et `WeekendTRUE`. Les autres variables ont un effet négatif assez important et l'effet négatif le plus fort est celui de `ExitRate`. Cela est logique puisque cette variable désigne le fait de quitter le site, et ne tiens pas compte en l'occurrence du fait de quitter le site après l'achat.

	FALSE	TRUE
0	8165	980
1	173	547

	FALSE	TRUE
0	2052	234
1	32	147

Pour l'échantillon d'entrainement, la sensibilité est de 0.98, la spécificité de 0.36, l'accuracy de 0.88 et l'erreur globale est donc de 0.12. Sur l'échantillon test, les valeurs sont assez similaires, avec une sensibilité de 0.98 spécificité de
Ce modèle est donc un modèle qui prédit bien les personnes qui achètent ou non.

3 LDA

	LD1
Administrative	0.0270648
Informational	0.0563033
BounceRates	3.5885302
ExitRates	-7.5187378
PageValues	0.0529099
SpecialDay	-0.0162251
MonthDec	-0.3359967
MonthFeb	-0.4606418
MonthJul	-0.0648612
MonthJune	-0.2011129
MonthMar	-0.3471192
MonthMay	-0.3200929
MonthNov	0.3973516
MonthOct	-0.1178372
MonthSep	-0.0618632
OperatingSystems	-0.0742318
Browser	0.0183782
Region	-0.0157249
TrafficType	0.0001872
VisitorTypeOther	-0.3639597
VisitorTypeReturning_Visitor	-0.2360458
WeekendTRUE	0.0534261

Ce modèle de LDA nous permet de connaitre les influences des variables sur le fait d'acheter ou non, de même que le modèle logit vu précédemment.
La LDA nous indique tout d'abord qu'il y a une disparité dans la variables `Revenue`. En effet, il y a 85% des individus qui n'achètent pas, et donc la variable prend la modalité FALSE (par la suite la modalité 0 de Revenue indique le fait de ne pas acheter).

Les coefficients nous permettent de savoir quelles variables ont le plus d'influences sur la variable `Revenue`. Ici c'est la variables `ExitRates` qui a la plus grande influence, influence négative. Tous les autres coefficients sont proches de 0.

Le fait d'avoir un gros déséquilibre entre les classes rend l'interprétation des coefficients inutile. Nous allons corriger l'équilibre des données grâce à l'oversampling et l'undersampling.

		Réalité	
		~~Achat~~	Achat
Prédictions	~~Achat~~	8174	1043
	Achat	164	484

Table 2: Matrice de confusion

	Indicateurs
Accuracy	0.878
95% IC	[0.871, 0.8841]
No Information Rate	0.8452
P-value [Acc>NIR]	2.2e-16
Kappa	0.3887
Sensitivity	0.98
Specificity	0.317

Table 3

D'après notre première estimation, la LDA est un bon outil de prédiction. En effet, l'accuracy est bonne et est supérieure au No Information Rate. L'indice de Kappa est supérieur à 0.2 et la p-value est faible. Mais voyons maitenant si nous pouvons améliorer notre modèle en rééquilibrant les classes.

Il est nécessaire de rééquilibrer les données de ce jeu de données car il y a un gros décalage entre le nombre de personnes qui achètent et celui des personnes qui n'achètent pas.

Cela va se faire au moyen de trois techniques ici, l'Oversampling, l'Undersampling puis les deux techniques combinées.
L'Oversampling consiste à rajouter des individu dans la classe minoritaire, en suivant la tendance des individus déjà présents dans la classe.
L'Undersampling consiste quant à lui à couper la classe majoritaire afin d'équilibrer les deux classes.

3.1 Oversampling

Grâce à cet algorithme, les deux classes sont équilibrées.

		Réalité	
		~~Achat~~	Achat
Prédictions	~~Achat~~	6531	2052
	Achat	1807	6286

Table 4: Matrice de confusion

	Indicateurs
Accuracy	0.769
95% IC	[0.7621, 0.775]
No Information Rate	0.5
P-value [Acc>NIR]	2.2e-16
Kappa	0.5372
Sensitivity	0.783
Specificity	0.754

Table 5

		Réalité	
		~~Achat~~	Achat
Prédictions	~~Achat~~	6531	2052
	Achat	1807	6286

Table 6: Matrice de confusion

	Indicateurs
Accuracy	0.769
95% IC	[0.7602, 0.7731]
No Information Rate	0.5
P-value [Acc>NIR]	<2.2e-16
Kappa	0.5333
Sensitivity	0.783
Specificity	0.754

Table 7

3.2 Undersampling

Avec l'algorithme d'Undersampling, les classes sont équilibrées mais les effectifs sont plus faibles que dans le cas de l'oversampling.

		Réalité	
		~~Achat~~	Achat
Prédictions	~~Achat~~	1218	379
	Achat	309	1148

Table 8: Matrice de confusion

	Indicateurs
Accuracy	0.775
95% IC	[0.7595, 0.7894)]
No Information Rate	0.5
P-value [Acc>NIR]	2.2e-16
Kappa	0.5494
Sensitivity	0.798
Specificity	0.752

Table 9

3.3 Oversampling et Undersampling

Les classes sont là encore mieux équilibrées et les effectifs sont compris entre ceux de l'Oversampling et l'Undersampling.

		Réalité	
		~~Achat~~	Achat
Prédictions	~~Achat~~	1218	379
	Achat	309	1148

Table 10: Matrice de confusion

	Indicateurs
Accuracy	0.775
95% IC	[0.7516, 0.7686)]
No Information Rate	0.5039
P-value [Acc>NIR]	2.2e-16
Kappa	0.5201
Sensitivity	0.798
Specificity	0.752

Table 11

3.4 Comparaison et nouveau modèle

Après comparaison des trois méthodes, on constate qu'elles rapportent des résultats similaires. Les AIC sont très proches et sont toujours supérieures au No Information Rate. L'indice de Kappa est supérieur à 0.2 et la p-value faible.
Il serait intinctif de conserver le modèle de l'Undersampling puisque c'est celui qui a l'AIC la plus faible. Cependant il est plus intéressant de travailler sur le modèle corrigé par l'Oversamplig puisque les effectifs sont les plus élevés et le modèle sera donc plus représentatif de la réalité.

Sur cette nouvelle LDA, on voit que les fréquences des personnes qui achètent et n'achètent pas sont parfaitement équilibrées.

	VisitorTypeOther	VisitorTypeReturning_Visitor	PageValues	BounceRates	Administrative
0	0.0069561	0.8699928	2.078889	0.0247744	2.105301
1	0.0077956	0.7678100	26.797849	0.0052518	3.489806

	Informational	MonthDec	MonthFeb	MonthJul	MonthJune	MonthMar	MonthMay
0	0.4507076	0.1454785	0.0165507	0.0364596	0.0242264	0.1685056	0.2858000
1	0.7836412	0.1187335	0.0023987	0.0309427	0.0158311	0.0961861	0.1911729

	MonthNov	MonthOct	MonthSep	OperatingSystems2	OperatingSystems3	OperatingSystems4
0	0.2126409	0.0429360	0.0338211	0.5177501	0.2236747	0.0390981
1	0.3963780	0.0593668	0.0474934	0.5983449	0.1370832	0.0491725

	OperatingSystems5	OperatingSystems6	OperatingSystems7	OperatingSystems8
0	0.0005997	0.0015591	0.0005997	0.0063564
1	0.0000000	0.0020389	0.0010794	0.0080355

Les moyennes intra-groupes nous indiquent le centre de gravité des groupes. Cela nous permet de connaître les variables pour lesquelles il y a les plus grandes disparités entre les modalités 0 (ne pas acheter) et 1 (acheter).
On remarque que dans la variable `PageValues` il y a une très grosse disparité. La moyenne de cette variable est beaucoup puls élevée pour les personnes qui achètent que pour celles n'achetant pas. Cela indique que la plupart des clients qui achètent passent par les mêmes pages. C'est la variable qui est la plus discriminante.
Les variables `Administratives`, `Informational` et `MonthNovember` discrimine aussi le fait d'acheter et ne pas acheter. On remarque donc que les clients qui sont les plus probables d'acheter sont les clients qui passent le plus de temps sur les pages Administratives et d'Information, mais aussi que les clients sont plus portés à acheter au mois de Novembre que les autres mois, ce qui est cohérent puisqu'il y a une fête ce mois ci, qui induit l'achat de cadeaux.

Dans ce modèle les prédictions sont bonnes car l'accuracy est élevée et supérieure au No Information Rate. La p-value est faible et l'indice de Kappa est supérieur à 0.2. De plus la sensibilité et spécificité sont élevées.

Pour conclure, les informations que nous avons ici ne nous permettent pas de détacher un profil type du client qui achèterait ou n'achtèterait pas. En effet, ce qui ressort de cette analyse c'est que les clients qui achètent sont des clients qui en ont besoin et non des achats pour le plaisir. Ceci est cohérent quand on sait qu'en Turquie, les clients préfèrent acheter en face, dans les magasins, et que l'achat sur internet n'est pas encore démocratisé.
De plus nous n'avons aucune informations sur le type de ventes qui est effectué ici, ce qui ne nous permet pas de cibler un type de population.

4 KNN

4.1 Premier modèle

Commençons par une construction d'un modèle KNN classique sans effectuer de modifications particulière sur les données [3].

		Réalité	
		~~Achat~~	Achat
Prédictions	~~Achat~~	2051	289
	Achat	33	92

Table 12: Matrice de confusion

Les prédictions sont très bonnes lorsqu'il s'agit d'un client non acheteur, mais assez mauvaises pour les acheteurs, ce qui est problématique pour notre étude, approfondissons le diagnostic de notre modèle :

3. Autre que la transformation en facteur et un one hot encoding.

	Indicateurs
Accuracy	0.869
95% IC	[0.8554, 0.8824]
No Information Rate	0.8454
P-value [Acc>NIR]	0.0004402
Kappa	0.311
Sensitivity	0.984
Specificity	0.241

Table 13

En effet, la précision est très bonne (presque 0.87), et est bien significativement différente du taux de non-informativité. Néanmoins, on se rend compte de deux choses :

- Premièrement, la sensibilité est excellente (0.98) mais la spécificité l'est beaucoup moins (0.24)
- Deuxièmement, le nombre d'observations présentes dans la catégorie des acheteurs est largement moindre que celui du nombre de non-acheteurs, ce qui est sûrement problématique pour les prédictions du modèle.

Pour répondre à ces deux problématiques, nous allons comparer le modèle classique à d'autres modèles construits sur le jeu de données modifié, à partir d'un algorithme d'oversampling, d'undersampling et d'un algorithme rassemblant ces deux derniers.

4.2 Oversampling

		Réalité	
		~~Achat~~	Achat
Prédictions	~~Achat~~	1483	590
	Achat	601	1536

Table 14: Matrice de confusion

Les prédictions sont nettement meilleures, les acheteurs sont beaucoup mieux prédits, ce qui en fait un modèle plus intéressant à conserver, poursuivons avec un rapide diagnostic :

	Indicateurs
Accuracy	0.717
95% IC	[0.7154, 0.7426]
No Information Rate	0.5017
P-value [Acc>NIR]	2e-16
Kappa	0.4584
Sensitivity	0.712
Specificity	0.722

Table 15

La sensibilité reste bonne, le modèle est bien plus équilibré que notre modèle de base, ce qui le rend bien plus intéressant. Voyons maintenant si nous pouvons faire mieux.

4.3 Undersampling

		Réalité	
		~~Achat~~	Achat
Prédictions	~~Achat~~	268	109
	Achat	112	272

Table 16: Matrice de confusion

Les résultats de ce modèle ont l'air sensiblement équivalents à notre modèle "oversampling", voyons si les indicateurs sont corrects.

	Indicateurs
Accuracy	0.71
95% IC	[0.6818, 0.7463]
No Information Rate	0.5128
P-value [Acc>NIR]	2e-16
Kappa	0.4304
Sensitivity	0.705
Specificity	0.714

Table 17

En effet, les indicateurs sont très proches des résultats obtenus sur le modèle précédent, mais ils sont toutefois un peu moins bon essayons alors une dernière estimation sur un algorithme hybride.

4.4 Oversampling et Undersampling

		Réalité	
		~~Achat~~	Achat
Prédictions	~~Achat~~	815	393
	Achat	385	872

Table 18: Matrice de confusion

Les prédictions sont à première vue meilleures que celles du modèles précédent, vérifions les indicateurs.

	Indicateurs
Accuracy	0.684
95% IC	[0.7027, 0.7385]
No Information Rate	0.5006
P-value [Acc>NIR]	2e-16
Kappa	0.4418
Sensitivity	0.679
Specificity	0.689

Table 19

4.5 Sélection du modèle

Puisque le modèle classique ne permets pas une détection satisfaisante des vrai positifs, nous allons confronter les modèles construits à partir de méthodes d'oversampling et d'undersampling. Pour les différencier, nous

utiliserons entre autres la Kappa value. Tous les modèles ont une très bonne Kappa value, mais le modèle Oversampling arrive en tête avec 0.4584, suivi par 0.4418 pour la méthode hybride et enfin elle est de 0.4304 pour l'undersampling.

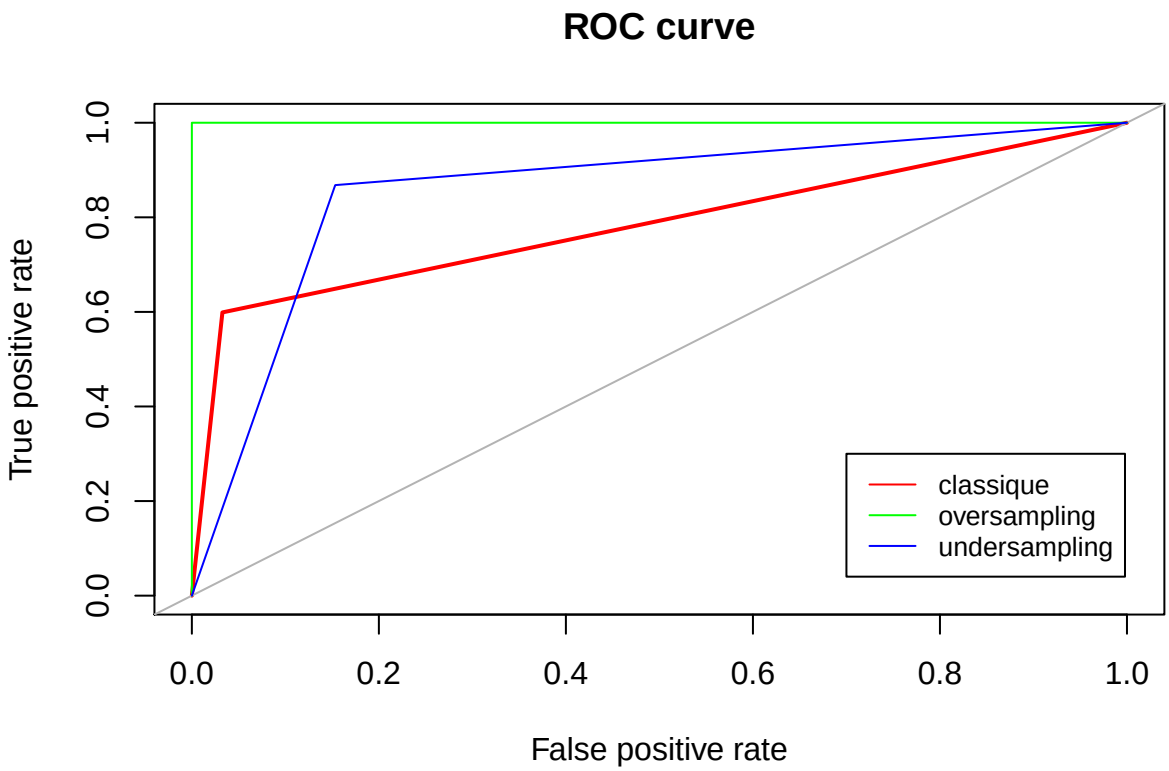
Ainsi le modèle sortant du lot est le modèle d'oversampling, c'est donc le modèle que nous retiendrons.

5 Arbres de décision

5.1 Arbre simple

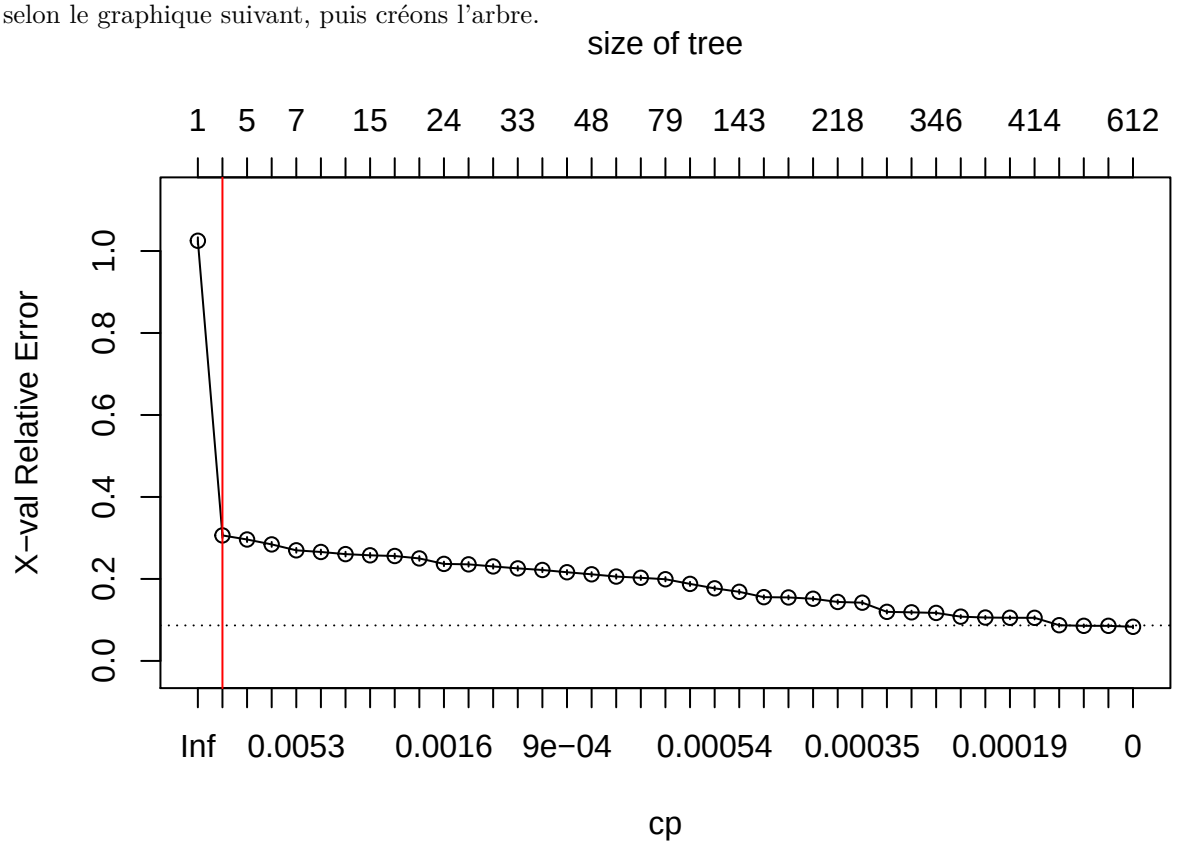
Afin de pouvoir créer le meilleur arbre de décision possible il convient de fixer des objectifs à atteindre. Ici il est de déterminer les facteurs faisant qu'une personne achète ou non. Nous nous intéressons donc principalement aux personnes qui achètent et notre objectif va être d'en détecter le plus possible.

Nous allons commencer par réaliser plusieurs échantillons afin de déterminer lequel permet d'obtenir les meilleurs résultats, pour cela nous utilisons la courbe ROC qui permet d'estimer le nombre de vrais positifs en fonction du nombre de faux positifs.



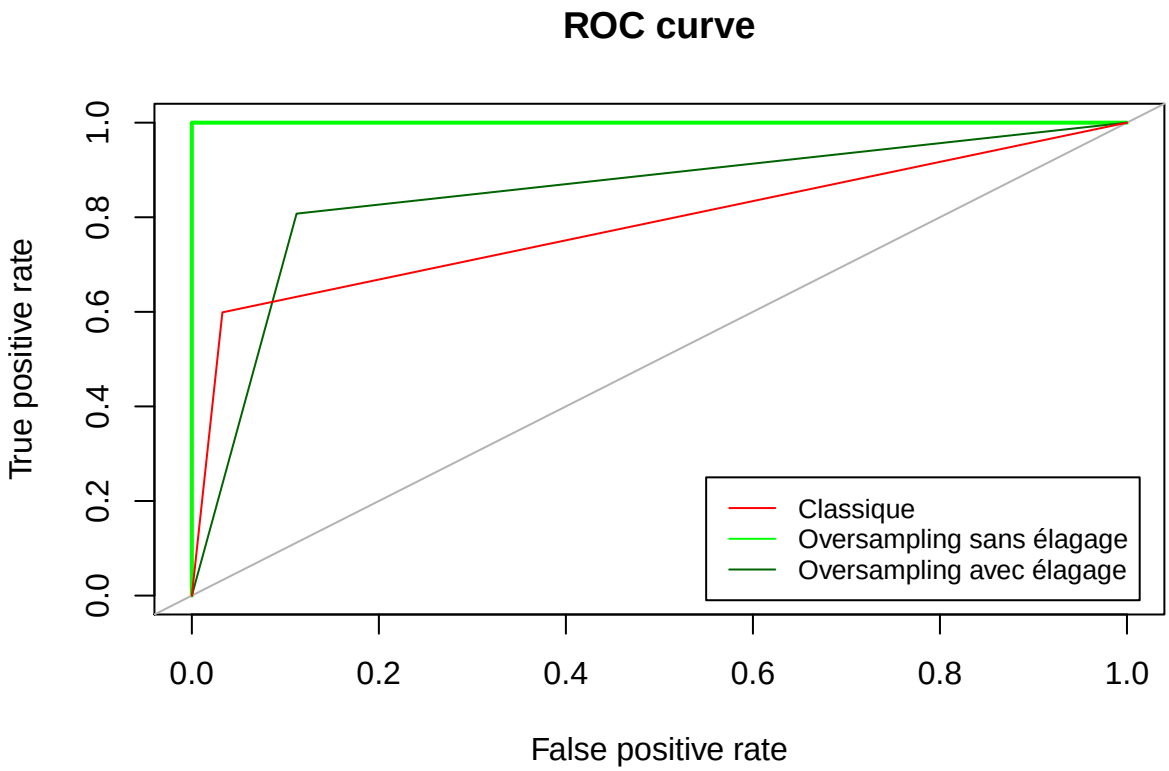
Ici, avec des arbres élagués automatiquement pour minimiser l'erreur l'échantillon le plus intéressant est celui en sur-échantillonnage. En effet celui-ci permet d'égaliser le nombre de visiteurs achetant à ceux n'achetant pas. Ainsi il leur donne le même poids et les rend plus facile à analyser. L'aire sous la courbe, presque égale à 1, indique très clairement que ce modèle sera celui qui permet d'effectuer les meilleurs prédictions, c'est donc celui que nous allons garder.

Nous pouvons donc lancer la création de l'arbre. Pour cela nous déterminons le nombre de noeuds optimal selon le graphique suivant, puis créons l'arbre.

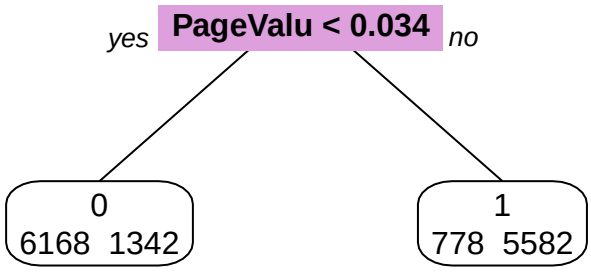


Lorsque l'on cherche à élaguer l'arbre afin de minimiser l'erreur on se rend compte que chaque branche additionelle permet de minimiser l'erreur. Cependant chaque nouvelle feuille n'apporte pas grand chose après la seconde et son grand nombre semble très lié au surapprentissage, ainsi nous ne conserverons que la première séparation.

Cela donne donc la courbe ROC suivante :



Cela permet donc d'obtenir l'arbre (très) élagué ci-contre.



C'est donc la variable Page Value qui permet de déterminer avec le plus d'efficacité si les visiteurs achèteront ou non. Elle indique en effet la valeur moyenne attribuée à chaque page que le visiteur a consulté. Ainsi si celle ci est inférieure à 0.94 il y a de grandes chances pour que le visiteur n'achète rien.

Il ne nous reste plus qu'à tester cet arbre en conditions "réelles", ce qui nous permet d'obtenir les résultats suivant.

		Réalité	
		~~Achat~~	Achat
Prédictions	~~Achat~~	3191	259
	Achat	285	375

Table 20: Matrice de confusion

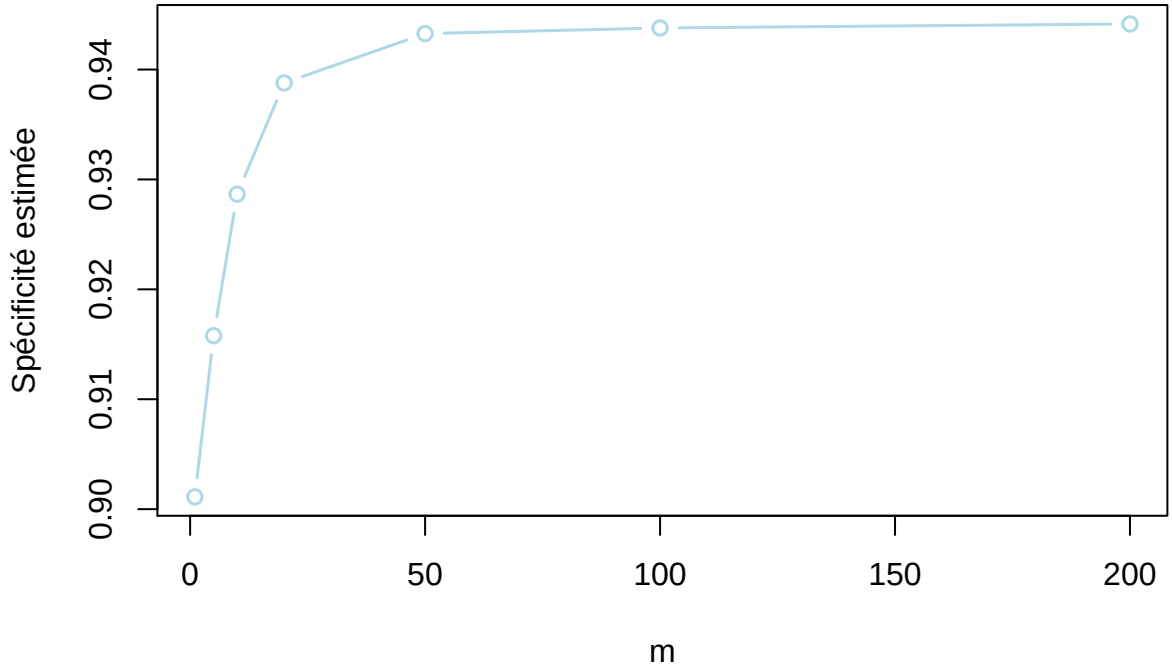
	Indicateur
Accuracy	0.868
95% IC	[0.8632, 0.8837]
No Information Rate	0.854
P-value [Acc>NIR]	0.0001
Kappa	0.5749
Sensitivity	0.918
Specificity	0.591

TABLE 21: Arbre de décision

Les résultats obtenus indiquent une forte précision. Les résultats obtenus sont significatifs et la spécificité est assez élevée. Bien qu'il y ait un nombre assez élevé de faux positifs ce modèle reste valide étant donné que ce que nous souhaitons est avant tout de déterminer le plus d'acheteurs possible.

5.2 Forêts aléatoires

Au vu du faible pourcentage de personnes achetant, les forêts aléatoire et le bagging ne sont pas optimaux pour estimer ce modèle à la suite des arbres et le boosting semble être une meilleure option, cependant pour des raisons techniques elle n'est pas réalisable car beaucoup trop longue en termes de calculs.
Cependant afin d'avoir tout de même un modèle plus robuste d'arbre nous allons utiliser une forêt aléatoire. Afin de réaliser une forêt optimisée on commence par déterminer le nombre d'arbres optimaux. On considère l'objectif de maximiser la spécificité, donc le nombre d'acheteurs détecté. On garde les données en oversampling étant donné que c'est ce qui est le plus efficace sur les arbres.



À partir du 100ème arbre, la spécificité estimée reste assez stable et n'augmente plus. Ainsi on crée une forêt de 100 arbres.

	Indicateurs
OOB error	0.0503394
Accuracy estimée	0.9496606
Sensibilité estimée	0.0699262

Table 22: Estimation des résultats

Avec ce modèle on peut s'attendre à avoir des prévisions assez bonnes, l'erreur Out Of Bag étant d'à peine plus 5% et la spécificité estimée à 72%. Ainsi c'est ce modèle que nous allons utiliser pour la prédiction en Random Forest.

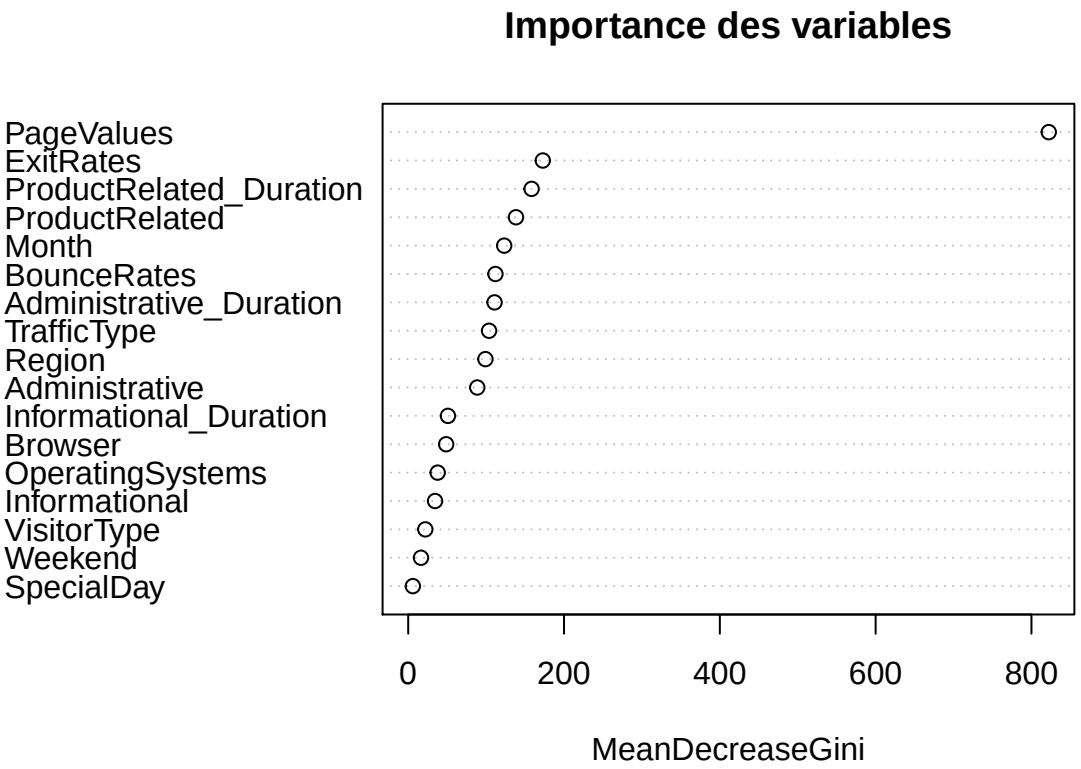
		Réalité	
		~~Achat~~	Achat
Prédictions	~~Achat~~	3318	259
	Achat	158	375

Table 23: Matrice de confusion

	Indicateurs
Accuracy	0.899
95% IC	[0.8927, 0.9111]
No Information Rate	0.854
P-value [Acc>NIR]	<2.2e-16
Kappa	0.5848
Sensitivity	0.955
Specificity	0.591

Table 24: Forêt aléatoire

Malgré des résultats moins bon que ceux attendus, ils restent toutefois assez bon. La précision du modèle est significativement supérieure à 90% et la spécificité est presque de 60%. Cela dit, cette spécificité reste assez faible étant donné que c'est le facteur que nous avons cherché à maximiser et cela implique que plus de 30% des acheteurs n'ont pas été détectés suite à cette prévision.



Une fois de plus c'est la variable Page Values qui a (de loin) le plus d'incidence sur le modèle. Le taux de sortie a aussi une forte influence. Ensuite ce sont les pages reliées au produit qui influent sur ce modèle. Si les visiteurs consultent beaucoup de pages de produit et passent du temps sur celles-ci on peut supposer qu'ils sont plus susceptibles d'acheter ensuite. Un facteur sur lequel il est facile d'influer est le mois, qui lui aussi est relativement important.

6 Conclusion

	LDA	KNN	Arb.Déc	Rand. For
Accuracy	0.7667	0.717	0.874	0.902
Spécificité	0.7510	0.722	0.695	0.598

Table 25: Comparaison des modèles

Le modèle à privilégier serait à première vue le modèle Random Forest car la précision des prédictions est de loin la plus élevée. De plus la spécificité est relativement bonne. Cependant nous ne pouvons pas exclure l'utilisation de la LDA en ce sens que la spécificité est la plus élevée et que la précision est bonne (et supérieure au No Information Rate).

Ce choix dépend donc de notre objectif, s'il est d'estimer le plus fidèlement possible le comportement des clients du site, il faut choisir la forêt aléatoire, et si on veut déterminer les critères d'achat sur le site, alors la LDA sera privilégiée.

Dans chaque modèle la variable de la valeur de la page est celle qui a le plus d'influence. Ainsi une Page Value élevée est associée à un visiteur qui achètera. Afin d'augmenter les achats passés sur le site il faudrait analyser les pages à valeur élevée afin de comprendre les mécanismes impliquant cela.