Etude de l'abonnement d'une banque

Notre étude se base sur les données liées aux campagnes de marketing d'une institution bancaire portugaise. Les campagnes marketings étaient des appels téléphoniques. Notre jeu de données possède 45 211 lignes soit 45 211 appels. Ces appels ont pour but d'augmenter la souscription à un produit. Nous avons donc une variable « y » qui est notre variable d'intérêt. Elle est égale à « oui » ou « non » si le destinataire a souscrit pour un produit ou non. Notre objectif est de créer un modèle capable d'expliquer cette variable « y » afin de pouvoir prédire si un client pourrait souscrire pour un produit selon d'autres informations.

Tout d'abord, notre variable « **y** » possède 5 289 « yes » soit **11,7%** et 39 922 « non » soit **88,3%.** Nous avons essayé de croiser cette variable avec d'autres variables pour observer sa relation avec les autres de manière descriptive. Nous avons d'abord croisé notre variable « y » avec plusieurs variables catégorielles :

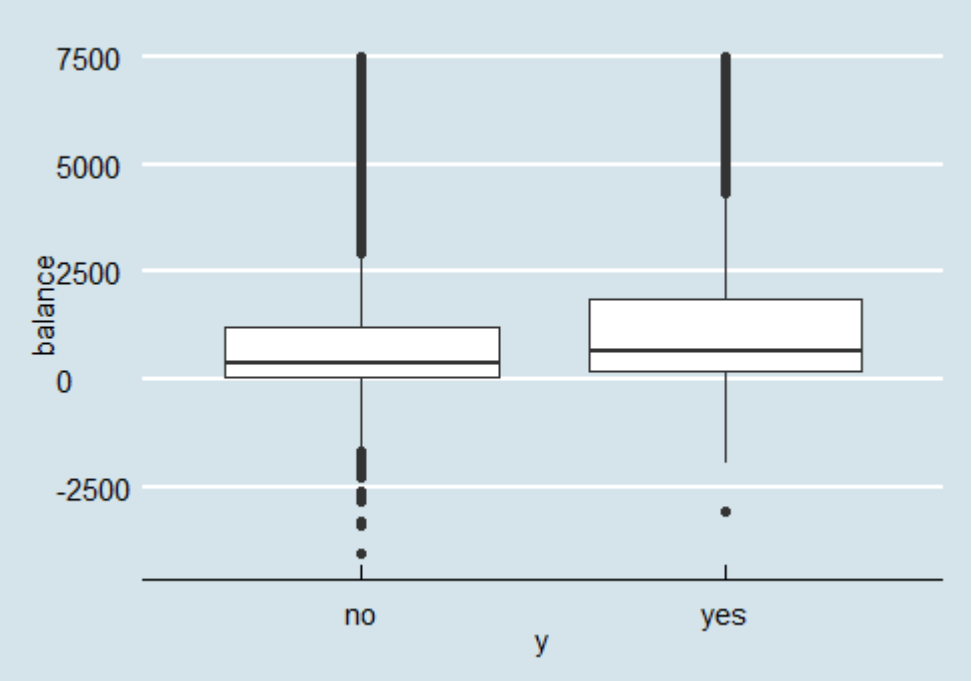
	Y		
	No	**Yes**	**Total**
Default = Non	88,2%	11,8%	**100,0%**
Default = Yes	93,6%	6,4%	**100,0%**
Housing = Non	83,3%	16,7%	**100,0%**
Housing = Yes	92,3%	7,7%	**100,0%**
Poutcome = Failure	87,4%	12,6%	**100,0%**
Poutcome = Other	83,3%	16,7%	**100,0%**
Poutcome = Success	35,3%	64,7%	**100,0%**
Poutcome = Unknown	90,8%	9,2%	**100,0%**
Contact = Cellulaire	85,0%	15,0%	**100,0%**
Contact = Telephone	86,6%	13,4%	**100,0%**
Contact = Unknown	96,0%	4,0%	**100,0%**

Sur ce tableau, nous avons croisé avec la variable ***Default*** qui est égale à « oui » ou « non » si la personne a un crédit en default. Pour cette variable, on voit qu'une grande majorité des gens n'ont pas souscrit. Malgré tout, on voit qu'il y a près de **12%** des gens qui n'ont pas de crédit en default qui ont souscrit alors qu'il n'y a que **6%** des gens qui ont un crédit en default qui ont souscrit. La variable ***Housing*** est égale à « oui » ou « non » si la personne possède un prêt pour son logement ou non. On observe le même phénomène que pour *Default*. Les personnes qui

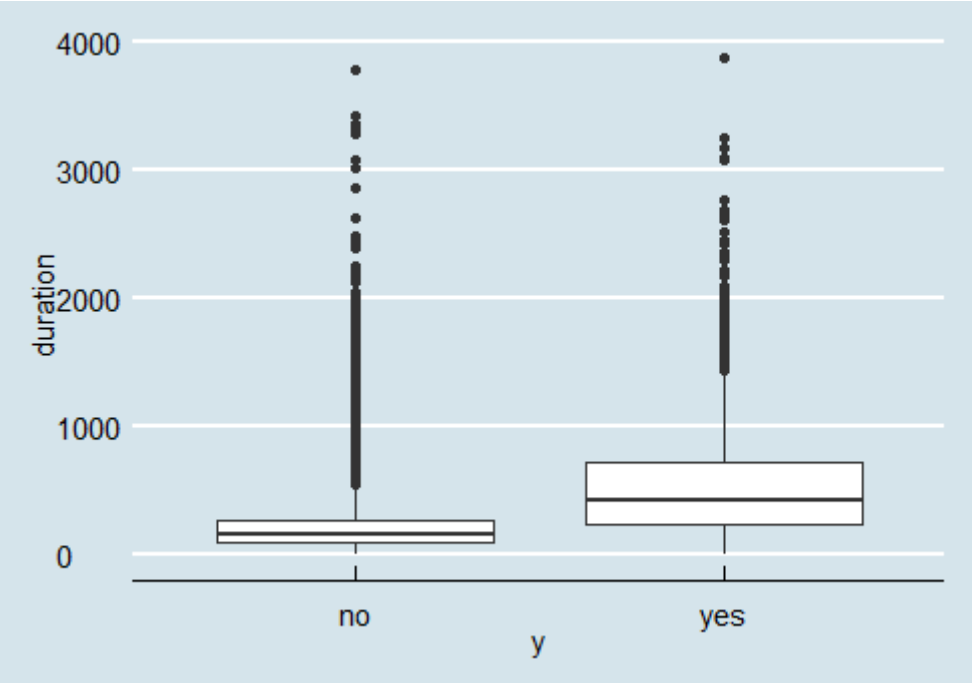
n'ont pas de prêt logement ont plus tendance à se souscrire contrairement aux personnes qui ont un prêt logement. Ensuite la variable ***Poutcome*** est le résultat de la campagne Marketing précédente : Failure, Other, Success ou Unknown. Ici, on distingue facilement que si la campagne précédente a été un succès pour la personne elle va se souscrire à un produit. Hormis la modalité « Success » qui a **64,7%** de « oui », aucune autres modalité atteint plus de **17%** de « yes » avec la variable « y ». Enfin, il reste la variable ***Contact*** qui définit le moyen de contact de la personne. Peu importe le moyen de contact, il y a une grande majorité de « non » mais les moyens de contacts Cellulaire et Telephone arrive à atteindre plus de **10%** de « oui » contrairement au moyen « Unknown ». Nous nous sommes également intéressés aux mois des appels avec la variable ***Month*** :

	Y		
	No	**Yes**	**Total**
Janvier	90,0%	10,0%	**100,0%**
Fevrier	83,4%	16,6%	**100,0%**
Mars	48,0%	52,0%	**100,0%**
Avril	80,3%	19,7%	**100,0%**
Mai	93,0%	7,0%	**100,0%**
Juin	89,0%	11,0%	**100,0%**
Juillet	91,0%	9,0%	**100,0%**
Août	89,0%	11,0%	**100,0%**
Septembre	53,5%	46,5%	**100,0%**
Octobre	56,2%	43,8%	**100,0%**
Novembre	90,0%	10,0%	**100,0%**
Décembre	53,3%	46,7%	**100,0%**

Lorsque l'on voit la répartition des appels par mois selon la variable d'intérêt, on peut remarquer plusieurs mois qui sortent du lot et possède une souscription plus grande les autres. Les mois ***Mars, Septembre, Octobre et décembre*** possèdent une souscription à un produit avec minimum **43%** de souscription jusqu'à plus de **52%** pour le mois de ***Mars.***



Nous avons ensuite réalisé quelques graphiques pour les variables quantitatives. Ci-dessous, voici un graphique qui croise la variable ***Balance*** qui est le solde de son compte moyen annuelle. Après avoir exclus certains points influents de ***Balance***, on observe que si les personnes qui souscrivent à un produit ont plus souvent un solde de son compte moyen plus que grand ceux qui ne se souscrivent pas. La médiane des personnes qui se souscrivent est un petit supérieur. De plus, les personnes qui ne souscrivent pas ont souvent un solde de compte négatif.



Ensuite nous avons croisé notre variable d'intérêt avec la durée de l'appel. On voit une nette différence de durée entre les personnes qui se souscrivent ou non. En effet, ceux qui se souscrivent ont plus souvent eu un appel qui a été plus long que ceux qui ne souscrivent pas. La médiane est quasiment le double entre les deux types de personnes.

Afin d'appliquer plusieurs modèles à notre jeu de données, nous avons en amont divisé ce dernier en une base d'apprentissage et une base test. Etant donné que nous possédons beaucoup plus de « no » que de « yes », nous avons décidé de créer des bases d'apprentissage et test ayant autant de « no » et de « yes ». De plus, le nombre assez conséquent de « yes » permet de créer deux bases avec assez de lignes. Nous avons donc une base d'apprentissage avec 7086 lignes avec **50%** de « no » et **50%** de « yes » et une base test de 3038 lignes avec **50%** de « no » et **50%** de « yes ».

Concernant les modèles choisis, Nous avons appliqué une régression logistique et une random forest sur mes données. Ces deux modèles permettront d'appliquer une classification sur la variable d'intérêt.

Pour chaque type de modèle, Nous avons cherché le meilleur modèle qui pourrait être utilisé. Ainsi Nous avons d'abord créé un modèle à partir de variables que nous avons jugées utiles dans le modèle. Un autre modèle qui sera le modèle complet. Nous réaliserons un dernier modèle qui sera une sélection pas à pas à partir du critère d'AIC. La sélection pas à pas choisira

d’ajouter ou retirer une variable une par une afin de minimiser au maximum l’AIC. Voici les 3 modèles :

Modèle 1(arbitraire): ***y*** ~ ***balance*** + ***duration*** + ***month*** + ***poutcome*** + ***campaign*** + ***job*** + ***contact*** + ***default*** + ***housing***

Modèle 2(complet) : ***y*** ~ ***toutes les variables soient 16 variables***

Modèle 3(selection pas à pas): ***y*** ~ ***balance*** + ***duration*** + ***month*** + ***poutcome*** + ***campaign*** + ***job*** + ***contact*** + ***default*** + ***housing + campaign + loan + marital + education + day + previous***

Notre modèle 3 qui est le résultat de la sélection pas à pas a donc choisi 15 des 16 variables disponibles. C’est la variable ***pdays*** qui n’a pas été choisi.

Après avoir créé ses trois modèles, Nous avons comparé le résultat de McFadden qui est une mesure de qualité d’ajustement du modèle.

Modèle 1 : ***McFadden = 43,4%***

Modèle 2 : ***McFadden = 43,872%***

Modèle 3 : ***McFadden = 43,860%***

On voit donc que le modèle 2 possède le meilleur McFadden mais qui est aussi très proche de celui du modèle 3.

	Maximum de vraisemblance	**AIC**	**BIC**	**Déviance**
Modèle 1	-2 777	5 621	5 847	5 555
Modèle 2	-2 757	5 600	5 895	5 514
Modèle 3	-2 757	5 595	5 869	5 515

Ci-dessus, on observe différentes caractéristique qui mesure la qualité du modèle. Chacun de ses indicateurs doit être minimisé. On voit que le modèle 3 est celui minimisant presque tous les indicateurs sauf la déviance. Le modèle 2 possède une déviance plus petite de 1. En effet, globalement le modèle 3 possède des indicateurs assez proches du modèle 2. Pour essayer de les départager, nous avons réalisé la matrice de confusion sur la base d'apprentissage. Cela peut permettre de différencier leur qualité de prédiction. Avant de réaliser ces matrices de confusion, nous avons entrainé les modèles en réalisant une cross-validation. Voici donc les 3 matrices de confusion :

Modèle 1 avec une régression logistique :

		Reference	
		No	**Yes**
Prédiction	**No**	2 995	634
	Yes	548	2 909

Modèle 2 avec une régression logistique :

		Reference	
		No	**Yes**
Prédiction	**No**	2 987	618
	Yes	556	2 925

Modèle 3 avec une régression logistique :

		Reference	
		No	**Yes**
Prédiction	**No**	2 984	618
	Yes	559	2 925

En comparant les 3 matrices de confusion, les 3 modèles possèdent une ***Précision*** de **83%** minimum avec le modèle 2 qui a une ***Précision*** supérieur au modèle 3 de **0.4%**. Pour ce qui est de la sensibilité et la sensibilité, on peut voir que les indicateurs entre le modèle 2 et 3 sont très similaires. Le modèle 2 arrive à mieux prédire 3 personnes « no » de plus que le modèle 3. Le modèle 1 est celui qui performe le moins sur la base d'apprentissage. Nous appliquerons le

modèle 3 sur la base test car il réalise des performances proches même supérieures parfois en ayant une variable de moins.

Ensuite nous avons réalisé des randoms forest sur les mêmes modèles que pour la régression logistique. Voici les matrices de confusion de chaque random forest sur la base d'apprentissage :

Modèle 1 sur une random forest :

		Reference	
		No	**Yes**
Prédiction	**No**	2 510	1 033
	Yes	192	3 351

Modèle 2 avec une random forest :

		Reference	
		No	**Yes**
Prédiction	**No**	2 634	909
	Yes	147	3 396

Modèle 3 avec une random forest :

		Reference	
		No	**Yes**
Prédiction	**No**	2 589	954
	Yes	148	3395

Les performances sur la base d'apprentissage du random forest sont différentes de la régression logistique, on peut voir rapidement que la random forest arrive à mieux identifier les personnes « yes » que la régression logistique. En contrepartie, il y a moins de bonnes prédictions sur les personnes « no ». Le modèle 2 arrive mieux prédire les personnes. La plus grande ***Précision*** est atteinte par le modèle 2 avec plus de **85%** alors que le modèle 3 possède **84,5%** de ***Précision***. Les modèles 2 et 3 sont donc difficiles à départager car le modèle 3 concurrence le modèle 2 en

ayant certains indicateurs supérieurs au modèle 2 avec une variable de moins. Nous appliquerons les modèles 2 et 3 sur la base test pour observer encore leur différence.

Nous avons ensuite comparé leur qualité de prédiction sur la base test. Voici la matrice de confusion et l'AUC pour la régression logistique pour les modèles 2 et 3.

Modèle 2 avec une régression logistique :

		Reference	
		No	**Yes**
Prédiction	**No**	1 292	276
	Yes	227	1 243

AUC= 90,99%

Modèle 3 avec une régression logistique :

		Reference	
		No	**Yes**
Prédiction	**No**	1 293	275
	Yes	226	1 244

AUC= 90,98%

La matrice de confusion sur la base test avec la régression logistique est un peu différente de la matrice de confusion sur la base d'apprentissage. En effet, ici le modèle 3 est supérieur au modèle 2 de très peu sur la sensibilité et la spécificité. Les deux modèles possèdent une ***Précision*** et un ***AUC*** très similaire. L'***AUC*** du modèle 3 est supérieur au modèle 2 de **0.01%** mais sur la matrice de confusion le modèle 2 prédit une personne de plus que le 3 en « yes » et « no ». On peut donc dire que le modèle 3 est meilleur que le modèle 2 mais de très peu. Dans notre cas, nous avons très peu de personnes « yes » il est important de bien prédire les « yes », chaque personne compte. Ainsi le modèle 3 peut servir pour comparer avec la random forest.

Modèle 3 avec une random forest :

		Reference	
		No	**Yes**
Prédiction	**No**	1 169	350
	Yes	57	1 462

AUC= 93,27%

De la manière que sur la base d'apprentissage, la prédiction de personnes « yes » est nettement meilleure avec une random forest que la régression logistique mais le nombre de personnes « no » est un peu moins performant que la régression logistique. De plus ***l'AUC*** est nettement supérieur à la régression logistique avec plus de **2%** de plus pour la random forest. Enfin la ***Précision*** n'est pas mauvaise aussi car elle est de **86,6%.**

Ainsi, si nous deviens choisir entre la régression logistique et la random forest pour notre étude, nous choisissons la random forest avec le modèle 2. La random forest montre de meilleurs résultats pour prédire les personnes « yes » que la régression logistique ce qui est important dans notre cas car le nombre de personnes « yes » est beaucoup plus faible de base dans notre jeu de données initial. De plus, le modèle 2 performe légèrement mieux que le modèle 3 pour prédire les « yes ».