Groupe D2 P1 B – Automne 2024

FEGHOUL Rayan - GAJAN Antoine

11/12/2024

Introduction

Dans le cadre de l'UV d'Apprentissage Automatique (SY19) de l'Université de Technologie de Compiègne, les étudiants sont amenés à travailler en groupe sur un projet de machine learning pour mettre en pratique les connaissances acquises tout au long du semestre.

Nous mettrons en place les techniques étudiées sur 3 datasets : deux simulés pour la régression et la classification et un dataset de notre choix. Dans ce rapport, nous présenterons et détaillerons les études menées.

1 Datasets simulés

1.1 Régression

1.1.1 Présentation du dataset

Le jeu de données simulé contient $n = 500$ observations de $p = 101$ variables. Les variables X_i avec $i \in 1, ..., 100$ sont les variables prédictives et y est la variable à prédire. On remarque que le jeu de données contient un nombre important de variables par rapport au nombre de données observées.

La variable y à prédire est distribuée selon une loi pouvant s'apparenter à une loi de moyenne $\mu = 60$. Quant aux variables prédictives, elles semblent indépendantes entre elles et suivre une loi uniforme entre 0 et 10 :

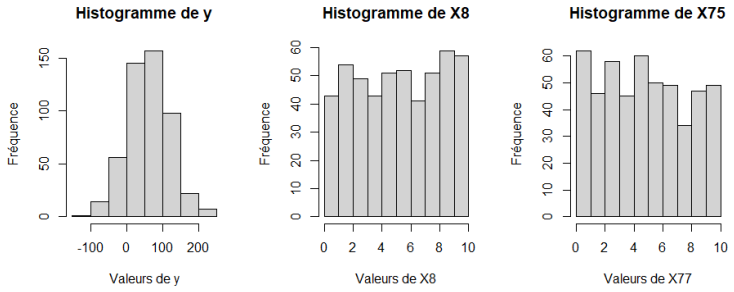

FIGURE 1 – Distribution des variables

1.1.2 Etude des corrélations entre les variables

Le nombre de variables étant relativement important par rapport au nombre d'observations, il est nécessaire d'étudier les corrélations entre les variables. Celle-ci étant difficile à visualiser de part le nombre important de variables, nous allons uniquement considérer les variables les plus corrélées :

Var1	Var2	Freq	abs_value
X92	y	0.3775479	0.3775479
X43	y	0.3232837	0.3232837
X57	y	-0.2718024	0.2718024
X8	y	0.2676847	0.2676847
X76	y	-0.2495047	0.2495047

FIGURE 2 – Les 5 paires de variables les plus corrélées

Sur cette matrice, seules 4 paires de variables ont une corrélation significative (supérieures à 25%). Nous remarquons de plus que ce lien est systématiquement avec la variable y. Dans le cadre d'une régression linéaire par exemple, nous aurons alors des coefficients significativement non nuls pour ces variables.

1.1.3 Réduction de la dimensionnalité

Pour répondre à la problématique initiale, nous avons mis en place différentes techniques d'apprentissage supervisé pour prédire la variable y. Les performances des modèles ont été évaluées à l'aide du **RMSE (Root Mean Squared Error)**, une métrique qui mesure la racine de la moyenne des erreurs quadratiques.

Le jeu de données comporte un nombre de variables important par rapport au nombre d'observations. Nous avons étudié plusieurs techniques pour réduire la dimensionnalité des données et potentiellement améliorer les performances des modèles : l' **analyse en composantes principales (ACP)**, la sélection des variables les plus importantes par Random Forest et la sélection de variable en minimisant le critère **BIC (Bayesian Information Criterion)**.

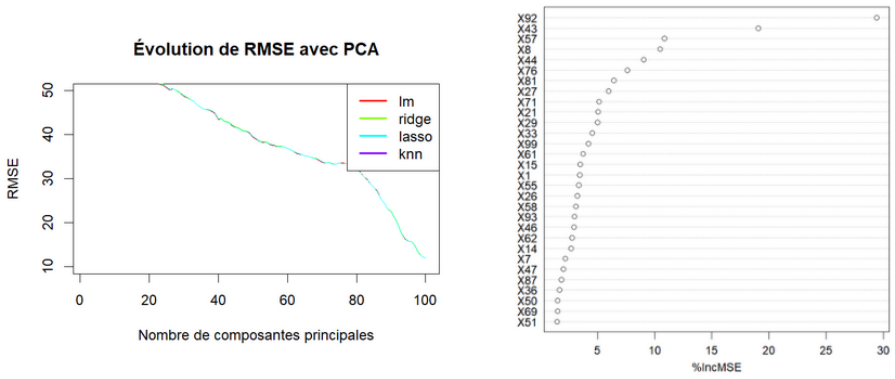

FIGURE 3 – Utilité de l'ACP (à gauche) et importance des variables du Random Forest (à droite)

Le graphique ci-dessous montre que le RMSE reste minimal lorsqu'on conserve toutes les composantes principales. Cela suggère que l'ACP ne permet pas de réduire efficacement la dimensionnalité tout en conservant l'information pertinente pour notre problématique. **La sélection de variables par le criètre BIC** nous a permis de réduire le jeu de données à **45 variables**, jugées suffisantes pour prédire y. Les variables sélectionnées par le critère BIC coïncident avec les plus importantes détectées par le Random Forest, ce qui valide leur pertinence.

1.1.4 Résultats obtenus

Après la normalisation des données et l'entraînement des modèles, les performances en termes de RMSE ont été estimées par validation croisée (10 plis). Les résultats obtenus sont présentés dans le tableau ci-dessous. Le meilleur modèle a été utilisé sur Maggle pour avoir un MSE de 138.72 (RMSE = 11.77).

Modèle	**100 variables et normalisation**	**Sélection de 45 variables et normalisation**
Régression linéaire	12.03	11.03
Elastic net ($\alpha = 0.6$)	11.59	11.07
Régression Lasso	11.74	11.18
GAM	12.03	11.68
SVM à noyau linéaire	12.17	11.19
Régression Ridge	12.20	11.27
SVM à noyau gaussien	28.26	19.04
Random Forest	44.92	42.68
KNN ($K = 22$)	48.26	48.03
Arbre de décision	59.14	58.82

TABLE 1 – Performance des modèles de régression en termes de RMSE

1.1.5 Résumé des tests effectués

Nous avons réalisé six tests en utilisant des modèles et techniques différentes afin d'obtenir le meilleur MSE possible sur les données de test de Maggle :

Test	Modèle	Méthodologie	MSE
Test 1	Régression Lasso	-	147.59
Test 2	Régression linéaire	Sélection des 45 variables les plus pertinentes	138.72
Test 3	SVM	Sélection des 45 variables les plus pertinentes	159.67
Test 4	Ensemble learning	10 régressions linéaires avec 10 sélections différentes de variables	143.64
Test 5	Régression linéaire	Sélection des 40 variables issues du Random Forest	148.99
Test 6	Régression linéaire	Sélection des 30 variables issues du Random Forest	145.88

TABLE 2 – Résumé des résultats des différents tests sur le jeu de régression

1.2 Classification

1.2.1 Présentation du Dataset

Le dataset utilisé dans cette étude est conçu pour résoudre un problème de classification. Il contient **500 observations**, chacune décrite par **50 features** et une variable cible (Y) à prédire. Ces 50 features se divisent en deux types :

— **45 variables numériques** : (X_1 à X_{45})
— **5 variables catégoriques** : (X_{46} à X_{50})

Ce mélange de types de données reflète souvent des phénomènes complexes, où des caractéristiques continues et discrètes interagissent.

La variable cible (Y) comporte **3 classes distinctes** représentant les catégories à prédire. Voici un résumé de la distribution des classes cibles :

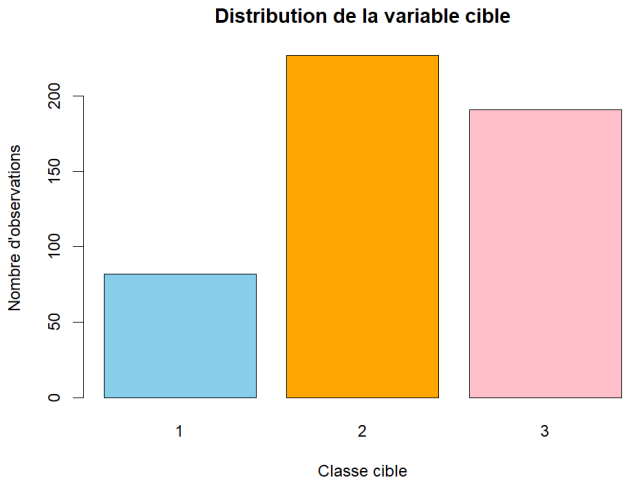

FIGURE 4 – Distribution de la variable cible

On remarque ici une **distribution déséquilibrée**, avec une forte domination de la classe 2, suivie par la classe 3, tandis que la classe 1 est nettement sous-représentée. Ce déséquilibre est une caractéristique importante du dataset, car il peut potentiellement biaiser les modèles d'apprentissage supervisé en faveur des classes majoritaires, rendant plus difficile la prédiction correcte des observations de la classe minoritaire.

Par ailleurs, une étude des **corrélations entre les variables numériques** a été réalisée. Les résultats montrent que le rang de la matrice de corrélation est égal au nombre total de variables (50), ce qui indique aucune variable n'est fortement corrélée avec une autre.

1.2.2 Distribution des variables prédictives

Nous avons remarqué que les différentes variables du jeu de données suivent différentes lois. En effet, nous avons remarqué que les variables X1 à X20 semblent suivent des lois uniformes, tandis que les variables X21 à X45 peuvent être assimilées à des variables gaussiennes. Enfin, les variables X46 à X50 sont entières.

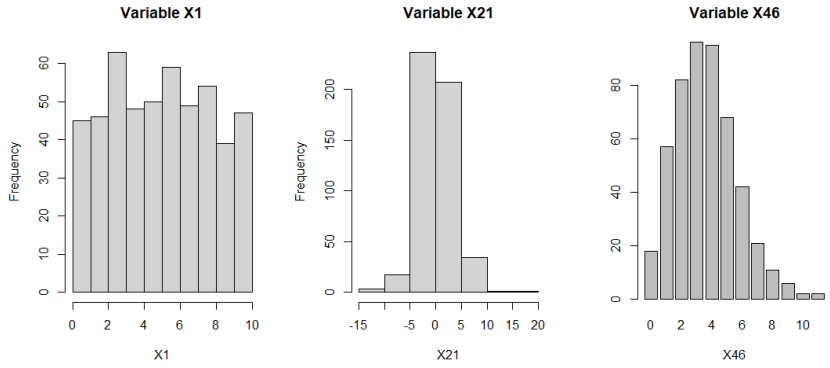

FIGURE 5 – Distribution des variables prédictives

1.2.3 Exploration préliminaire de modèles prédictifs

Tout d'abord, pour voir quel est le meilleur modèle prédictif, nous allons lancer plusieurs modèles avec pleins d'hyperparamètres différents à l'aide d'un grid-search.

La validation croisée à 10 plis (ou 10-fold cross-validation) est utilisée pour évaluer la performance du modèle. Le tableau résumant les résultats des modèles utilisés est proposé ci-dessous. L'accuracy indiqué correspond à la moyenne des accuracy obtenu sur les 10 plis de la validation croisée.

Modèle	Accuracy (en %)
Naive Bayes	65.19
QDA	64.77
SVM radial	63.60
LDA	60.43
Random Forest	60.23
Régression Logistique Multinomiale	59.00
KNN	50.20

TABLE 3 – Performance des modèles de classification en termes d'accuracy

1.2.4 Ensemble learning

L'ensemble learning est une technique consistant à combiner plusieurs modèles pour améliorer les performances prédictives par rapport à un modèle individuel. Cette approche permet de réduire à la fois le biais et la variance, produisant ainsi des prédictions plus robustes et fiables. Nous avons ici mis en œuvre une méthode d'ensemble en utilisant cinq modèles distincts : régression logistique multinomiale, QDA, Naive Bayes, SVM, et Random Forest.

Chaque modèle a été entraîné sur un ensemble spécifique de variables sélectionnées en fonction de leurs caractéristiques pertinentes. Pour QDA et Naive Bayes, seules les variables suivant une distribution gaussienne ont été utilisées. Pour le Random Forest, toutes les variables du dataset ont été utilisées. Pour la régression logistique multinomiale et le SVM, nous avons retenu les variables jugées importantes déterminées par les scores d'importance fournis par le modèle Random Forest.

Afin d'évaluer la performance de chaque modèle, nous avons appliqué une validation croisée à 10 plis. Les prédictions finales ont été obtenues grâce à un mécanisme de vote majoritaire : pour chaque échantillon, chaque modèle prédit une classe, et la classe recueillant le plus grand nombre de votes est sélectionnée comme prédiction finale.

Pour identifier la combinaison optimale de modèles permettant de maximiser l'accuracy, nous avons testé toutes les combinaisons possibles des cinq modèles. La meilleure performance, avec une accuracy de 71,2%, a été atteinte en combinant les prédictions de la régression logistique multinomiale, du QDA, et du SVM.

1.2.5 Résumé des techniques expérimentées

Nous avons ainsi réalisé six tests en utilisant des modèles et techniques différentes pour obtenir la meilleure accuracy possible sur les données de test disponibles sur Maggle :

Test	**Modèle**	**Méthodologie**	**Accuracy**
Test 1	Naive Bayes	-	0.63
Test 2	LDA	Appliqué aux données catégorielles	0.62
Test 3	FDA + Naive Bayes	FDA comme prétraitement puis Naive Bayes	0.61
Test 4	QDA	Appliqué uniquement sur les variables suivant une distribution gaussienne	0.68
Test 5	Ensemble Learning	Combinaison de plusieurs modèles pour améliorer la performance globale	0.69
Test 6	QDA	Entraîné uniquement sur les données des classes 2 et 3 pour pallier le problème de classe minoritaire	0.63

Table 4 – Résumé des performances des différents modèles sur Moodle

2 Dataset Bank Marketing

2.1 Présentation du Dataset

Le dataset **Bank Marketing** contient des données issues de campagnes de marketing menées par une institution bancaire portugaise. Ces campagnes, réalisées par appels téléphoniques, visaient à promouvoir la souscription à un dépôt à terme. Plusieurs contacts avec les clients ont pu être nécessaires pour évaluer leur décision.

L'objectif principal de l'analyse est de prédire si un client souscrira à un dépôt à terme, ce qui est représenté par la variable cible **y**. Le dataset comporte **17 variables (catégoriques et numériques)** et **45211 observations**. Les variables sont décrites dans le tableau ci-dessous :

Variable	**Type**	**Informations**
Données bancaires relatives au client		
`age`	Numérique	Âge du client.
`job`	Catégorique	Type d'emploi parmi 12 catégories (ex. : "admin.", "technician", "student").
`marital`	Catégorique	Statut marital du client : "married", "single", "divorced" (inclut les veufs).
`education`	Catégorique	Niveau d'éducation : primaire, secondaire, tertiaire ou inconnu.
`default`	Binaire	Le client a-t-il des crédits en défaut : "yes", "no".
`balance`	Numérique	Solde annuel moyen du client en euros.
`housing`	Binaire	Le client a-t-il un prêt immobilier : "yes", "no".
`loan`	Binaire	Le client a-t-il un prêt personnel : "yes", "no".
Informations sur le dernier contact de la campagne en cours		
`contact`	Catégorique	Type de communication : "telephone", "cellular", "unknown".
`day`	Numérique	Jour du mois où le dernier contact a eu lieu.
`month`	Catégorique	Mois du dernier contact (abréviations des mois, ex. : "jan").
`duration`	Numérique	Durée du dernier appel en secondes.
Informations liées aux campagnes marketing		
`campaign`	Numérique	Nombre total de contacts effectués pour ce client lors de la campagne actuelle.
`pdays`	Numérique	Nombre de jours écoulés depuis le dernier contact lors d'une campagne précédente (-1 signifie jamais contacté).
`previous`	Numérique	Nombre total de contacts lors de campagnes précédentes.
`poutcome`	Catégorique	Résultat de la dernière campagne précédente : "success", "failure", "other", "unknown".
Variable cible		
`y`	Binaire	Le client a-t-il souscrit un dépôt à terme : "yes", "no".

Table 5 – Description des variables du dataset Bank Marketing

2.2 Analyse exploratoire

L'analyse exploratoire des données (EDA) permet d'obtenir une compréhension globale des caractéristiques du dataset et des relations potentielles entre les variables. Voici les principales observations sur ce dataset, ainsi que leur interprétation.

1. Distribution de la variable cible y

La variable cible, représentant la souscription ou non à un dépôt à terme, est fortement déséquilibrée : la majorité des clients n'ont pas souscrit. Cela reflète le comportement classique dans les campagnes marketing, où seuls quelques clients répondent positivement.

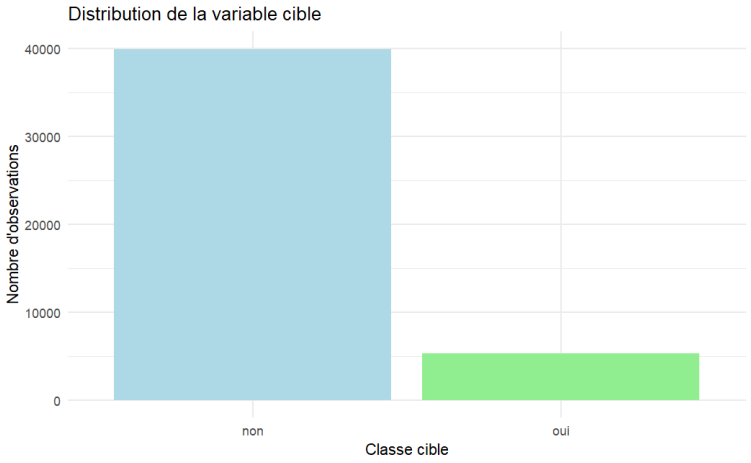

FIGURE 6 – Distribution de la variable cible

2. Relations entre les variables numériques

La matrice de corrélation entre les variables numériques montre que la corrélation la plus forte est de 0.45 entre les variables previous (nombre de contacts lors des campagnes précédentes) et pdays (nombre de jours depuis le dernier contact). Cela suggère que les clients ayant été contactés plus récemment sont susceptibles d'avoir eu un nombre plus élevé de contacts dans les campagnes précédentes.

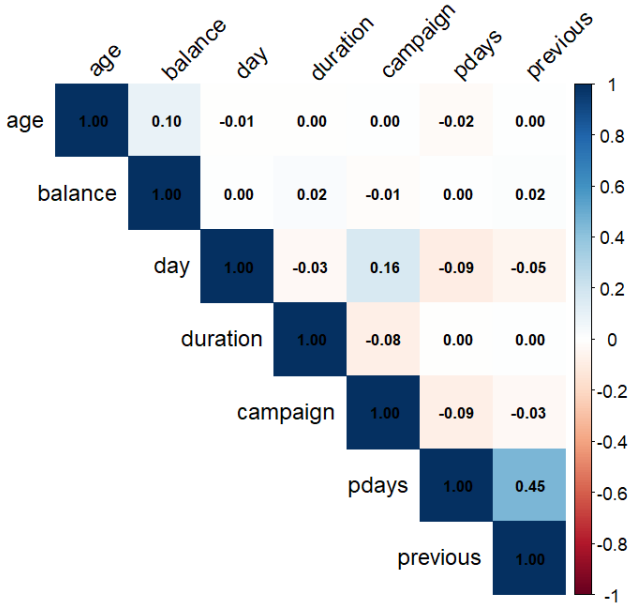

FIGURE 7 – Matrix de corrélation des variables numériques

3. Variables influentes :

La variable duration se distingue par sa forte corrélation avec la souscription. Les appels plus longs sont significativement associés à des souscriptions réussies, ce qui pourrait indiquer qu'un engagement plus important avec le client augmente les chances de succès de la campagne.

Les résultats des campagnes précédentes, reflétés par la variable poutcome, ont un impact sur la décision de souscrire à un dépôt à terme. Les clients ayant eu un résultat positif lors des campagnes précédentes (catégorie "success") sont plus susceptibles de souscrire que ceux qui ont eu des résultats moins favorables ("failure", "other", "unknown").

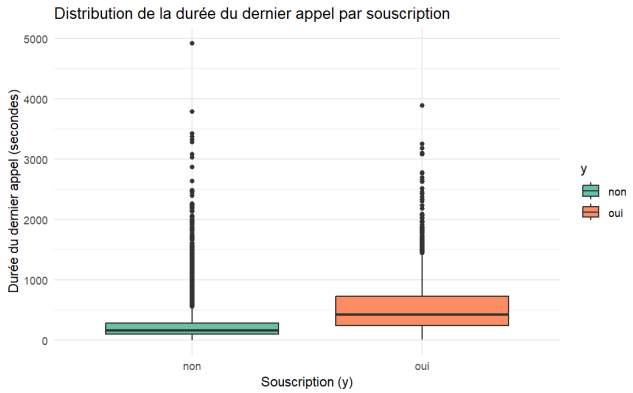

FIGURE 8 – Distribution de la durée des appels selon la souscription

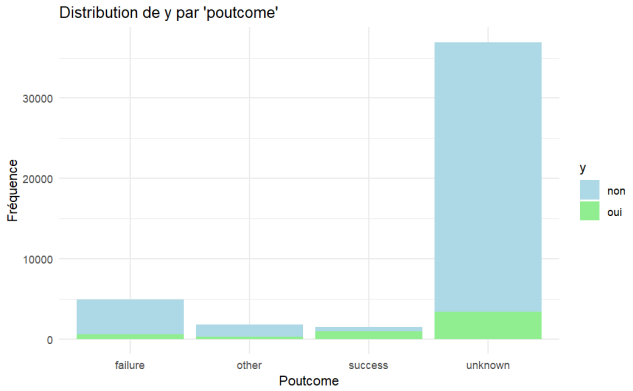

FIGURE 9 – Impact de poutcome sur la souscription

4. Variables peu pertinentes

Certaines variables présentent peu ou pas d'influence notable sur la souscription :

— Balance : Le solde annuel moyen n'affiche pas de tendance significative.
— Prêts immobiliers (housing) : Les clients ayant des prêts souscrivent légèrement moins souvent, sans tendance marquée.
— Statut marital (marital) : Aucune influence significative n'est observée.

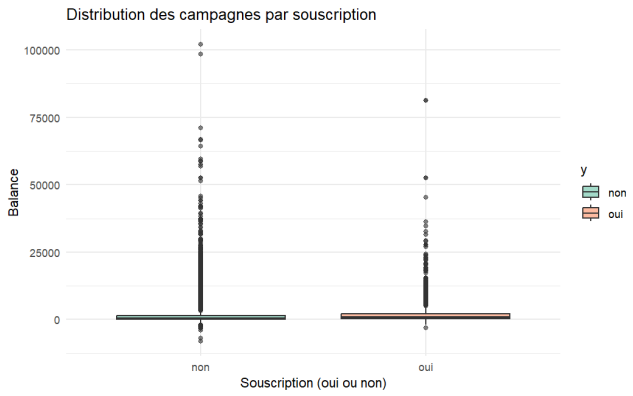

FIGURE 10 – Distribution des balances par souscription

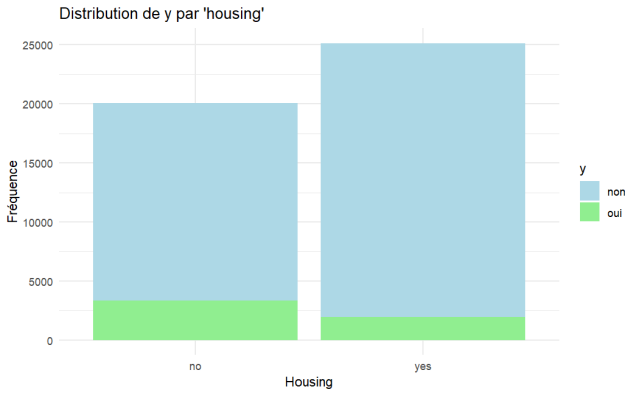

FIGURE 11 – Souscription et prêts immobiliers (*housing*)

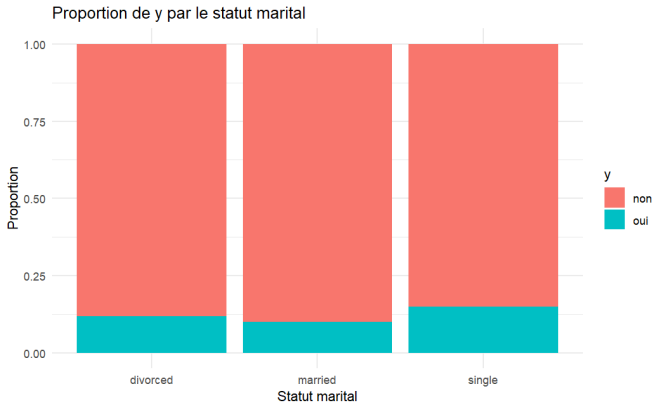

FIGURE 12 – Souscription et statut marital (*marital*)

5. Tendance mensuelle

Les souscriptions sont plus fréquentes en décembre, mars, octobre et septembre, ce qui suggère que ces périodes pourraient être optimisées pour des campagnes ciblées.

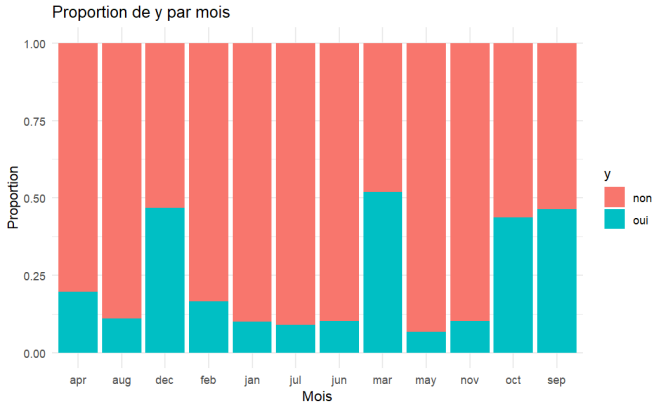

FIGURE 13 – Souscription par mois

L'analyse exploratoire révèle plusieurs points intéressants. Les variables duration et poutcome ressortent comme les plus influentes, indiquant des pistes prometteuses pour la modélisation. Les variables peu influentes (balance, housing, loan, marital) pourraient être de moindre importance pour les modèles prédictifs.

2.3 Méthodes d'apprentissage

À partir des données observées, nous avons cherché à comprendre l'impact de la campagne marketing et à prédire si un nouveau client allait souscrire ou non à un dépôt à terme en fonction des variables prédictives.

Pour cela, nous avons réalisé un apprentissage supervisé en utilisant les différents modèles étudiés au cours du semestre. Étant donné la taille importante de notre jeu de données, nous avons séparé celui-ci en un jeu d'entraînement (80%) et un jeu de test (20%).

2.3.1 Impact de la distribution des classes

Étant donné la répartition inégale des classes dans le jeu de données, il est crucial de tenir compte de cette particularité. Pour cela, nous adopterons diverses approches. En raison de la nature des variables et de la répartition des données, nous étudierons des modèles appris sur des données avec One Hot Encoding et sur des données issues d'un down-sampling de notre jeu d'entraînement afin d'obtenir une distribution des classes plus équilibrée.

Ensuite, nous analyserons les métriques "accuracy" et "F1-score". En présence d'une répartition déséquilibrée des observations, il est important de vérifier si le modèle se contente de prédire majoritairement la classe dominante (résultant en une accuracy élevée mais un F1-score faible) ou s'il parvient réellement à comprendre les subtilités du problème (ce qui se traduit par un F1-score élevé). Cette analyse du F1-score sera déterminante pour évaluer la véritable performance du modèle.

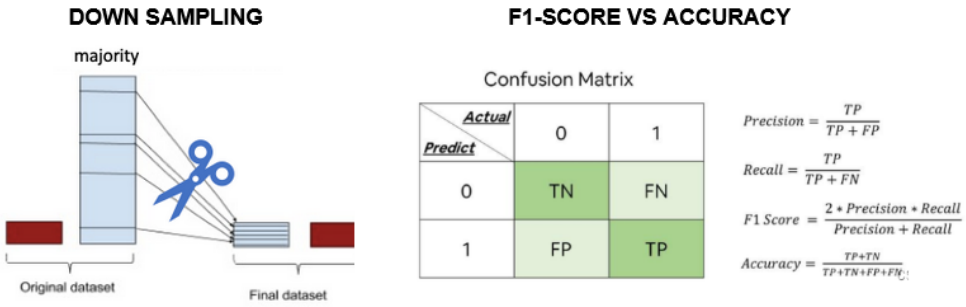

FIGURE 14 – Schémas explicatifs du down-sampling et du F1-Score

2.3.2 Apprentissage des modèles

Pour chaque modèle étudié en cours, nous avons mis en œuvre une validation croisée en 10 plis. Cette méthode nous a permis d'optimiser efficacement les hyperparamètres de chaque modèle.

En plus des modèles classiques étudiés tout au long du semestre, nous avons essayé une approche par réseaux de neurones avec la librairie Keras. Nous avons mis en place 3 couches cachées, comportant respectivement 64, 32, et 16 neurones avec une fonction d'activation ReLU. La sortie, de nature sigmoide, donne la probabilité pour un client de souscrire à un dépôt à terme.

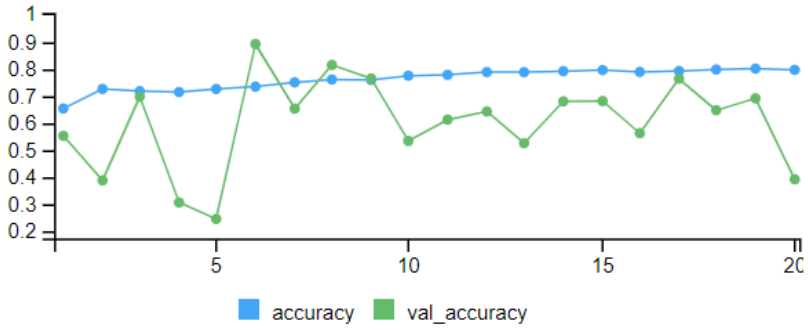

FIGURE 15 – Résultats du réseau de neurones

Le réseau de neurones n'est pas parvenu à généraliser l'apprentissage. En effet, en dépit de son accuracy très élevé, le réseau de neurone a overfitté le jeu d'entrainement, mais n'a pas été capable de prédire correctement sur le jeu de test, comme en témoigne sa matrice de confusion et son F1-Score de 0.43.

Classe réelle	**Classe prédite**	
	1	2
1	**7609**	399
2	638	**396**

TABLE 6 – Matrice de confusion sur le jeu de test avec le réseau de neurones

2.3.3 Résultats obtenus

Nous avons analysé nos résultats à l'aide des métriques accuracy et F1-Score sur le jeu de test. Nous avons d'abord entraîné nos données "brutes", puis avons réitéré l'expérience avec des données sous forme de One Hot Encoding et le down-sampling.

Modèle	Données brutes		One Hot		Downsampling	
	F1	Accuracy	F1	Accuracy	F1	Accuracy
Random Forest	0.907	0.980	0.524	0.902	0.610	0.861
Réseau de neurones			0.520	0.851	0.441	0.748
LDA	0.507	0.903	0.496	0.897	0.542	0.852
Arbre	0.455	0.903	0.455	0.900	0.555	0.837
QDA	0.446	0.866	0.461	0.864	0.458	0.851
Reg. Log.	0.442	0.903	0.442	0.899	0.529	0.858
SVM Radial	0.415	0.903			0.555	0.841
SVM Lineaire	0.392	0.899			0.536	0.840
KNN	0.256	0.888			0.424	0.765
Nayes Naïf	0.002	0.889			0.422	0.841

TABLE 7 – Performances des modèles avec données brutes, One Hot Encoding et downsampling sur le jeu de test

Globalement, la mise en place du One Hot Encoding n'a pas permis d'améliorer significativement les performances des modèles en termes de F1-Score. Cependant, il semblerait que le downsampling ait permis d'améliorer le F1-Score pour tous les modèles sauf le Random Forest.

La forêt d'arbres entraînée sur les données brutes est parvenue à un F1-Score de 91%, ce qui indique que le modèle est parvenu à tenir compte des spécificités du jeu de données (sous-représentation d'une classe, grand nombre de données,...). Ce meilleur modèle nous a donné la matrice de confusion suivante sur le jeu de test.

Classe réelle	**Classe prédite**	
	1	2
1	**7898**	75
2	176	**893**

TABLE 8 – Matrice de confusion sur le jeu de test avec le Random Forest

Ce modèle pourrait être utilisé par les organisateurs de la campagne marketing pour prédire, selon les caractéristiques d'un individu, s'il va souscrire à un dépôt à terme.

Conclusion

Ce projet a été une opportunité pour mettre en pratique les concepts théoriques que nous avons étudiés tout au long du semestre. À travers l'analyse exploratoire des données et la mise en œuvre de méthodes d'apprentissage, nous avons approfondi notre compréhension des données et développé des modèles prédictifs.

Les méthodes d'apprentissage supervisé se sont avérées efficaces pour prédire nos variables cibles. Ces résultats soulignent l'importance d'une sélection rigoureuse des caractéristiques et d'un ajustement méthodique des modèles pour améliorer les performances prédictives.