Rapport

Approche Analytique et Prédictive de données de banque : Exploration de Données et Modélisation avec machine Learning

Redigé par : Abderrahmane BOUFARES etJean Christophe DEBEVE

Sous la supervision de monsieur Olivier PERON

aboufares@etud.univ-pau.fr

Table des matières

Introduction

Ce rapport présente une analyse approfondie des données d'une banque X, il vise à découvrir lequel des profils des clients de la banque sont susceptible d'avoir un prêt bancaire. Le profil d'un client sera identifié grâce à plusieurs caractéristiques tel que l'âge, les années d'expériences, le revenu mensuel par millier, le nombre de personnes par famille, les dépenses mensuelles, le niveau d'étude et l'acquisition d'un compte d'épargne logement. Pour le faire, nous allons effectuer différentes analyses et utiliser différents modèles afin de déterminer lequel des modèles est optimal.

Analyse Approfondie

Exploration des données

Caractéristiques de l'echantillon

Le choix des variables sera à la base de leur relation avec l'obtention du prêt, différents tests statistiques seront utilisés.

Le tableau suivant est la présentation de toutes nos variables explicatives, il contient les paramètres statistiques les plus importants. Il s'agit de la moyenne, l'écart-type et les quantiles.

	ID	Age	Exp	RMM	F	\
count	4894.000000	4894.000000	4894.000000	4894.000000	4894.000000	
mean	2499.444013	45.333265	20.088680	73.684716	2.405394	
std	1445.437454	11.474275	11.475238	46.045208	1.149174	
min	1.000000	23.000000	-3.000000	8.000000	1.000000	
25%	1242.250000	35.000000	10.000000	39.000000	1.000000	
50%	2501.500000	45.000000	20.000000	63.000000	2.000000	
75%	3750.750000	55.000000	30.000000	98.000000	3.000000	
max	5000.000000	67.000000	43.000000	224.000000	4.000000	

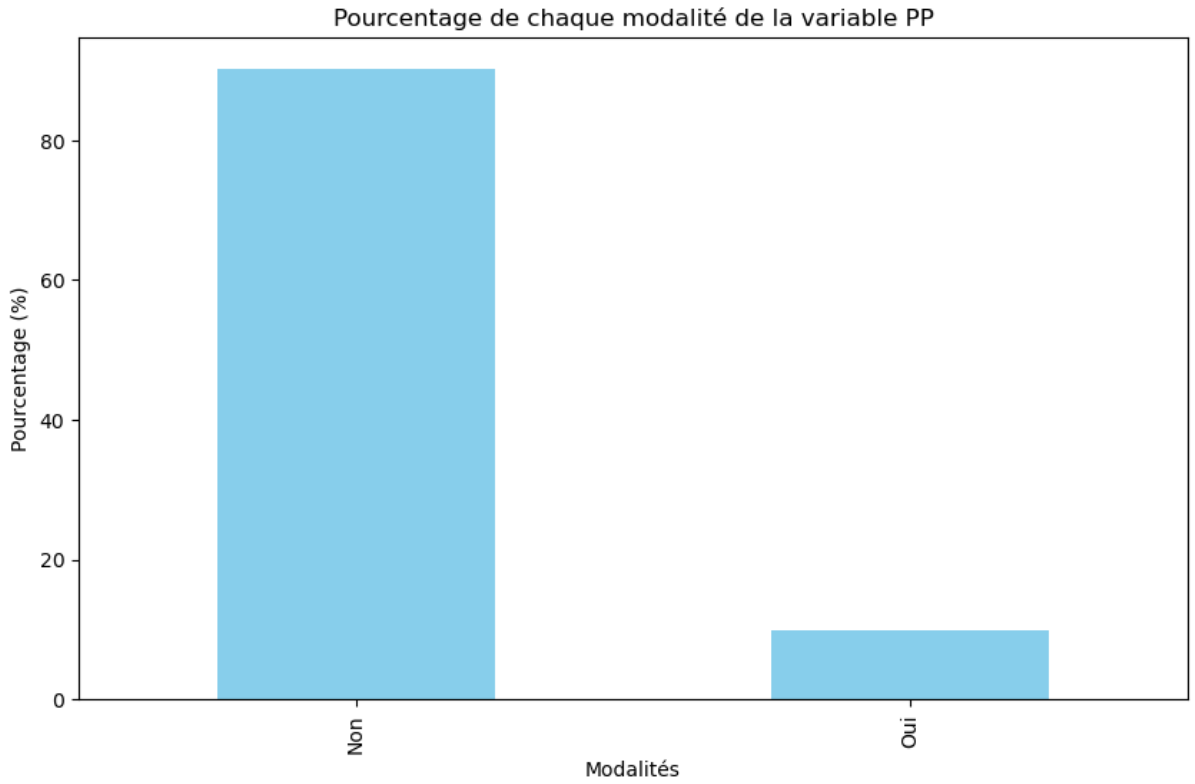
	DM	Educ
count	4894.000000	4894.000000
mean	1979.912137	1.885574
std	1742.801543	0.839199
min	100.000000	1.000000
25%	700.000000	1.000000
50%	1600.000000	2.000000
75%	2600.000000	3.000000
max	10000.000000	3.000000

Comme vous pouvez le constater :

- L’âge moyen de nos clients est de 45,33 allant de 23ans a 67ans.
- Le niveau d’expérience moyen est de 20ans allant de -3ans pour les étudiants jusqu’à 30ans.
- La moyenne du revenu mensuel par millier est de 73,68, elle va de 8 000 à 98 000.
- La moyenne des membres de familles est de 2,4 avec 4 modalités 1 personnes par famille, 2,3 ou 4.
- En moyenne les dépenses mensuelles pour nos clients sont d’équivalent de 1980, allant de 100 à 10 000.
- Le niveau d’éducation moyen de notre échantillon est de 1,88, avec 3 modalités : 1,2 et 3, le niveau 3 est le meilleur.

Parmi nos 4894 clients, seulement 9,78% ont obtenu un prêt.

```
Pourcentages de chaque modalité de la variable PP:
PP
Non    90.212505
Oui     9.787495
Name: count, dtype: float64
```



Etude des relations

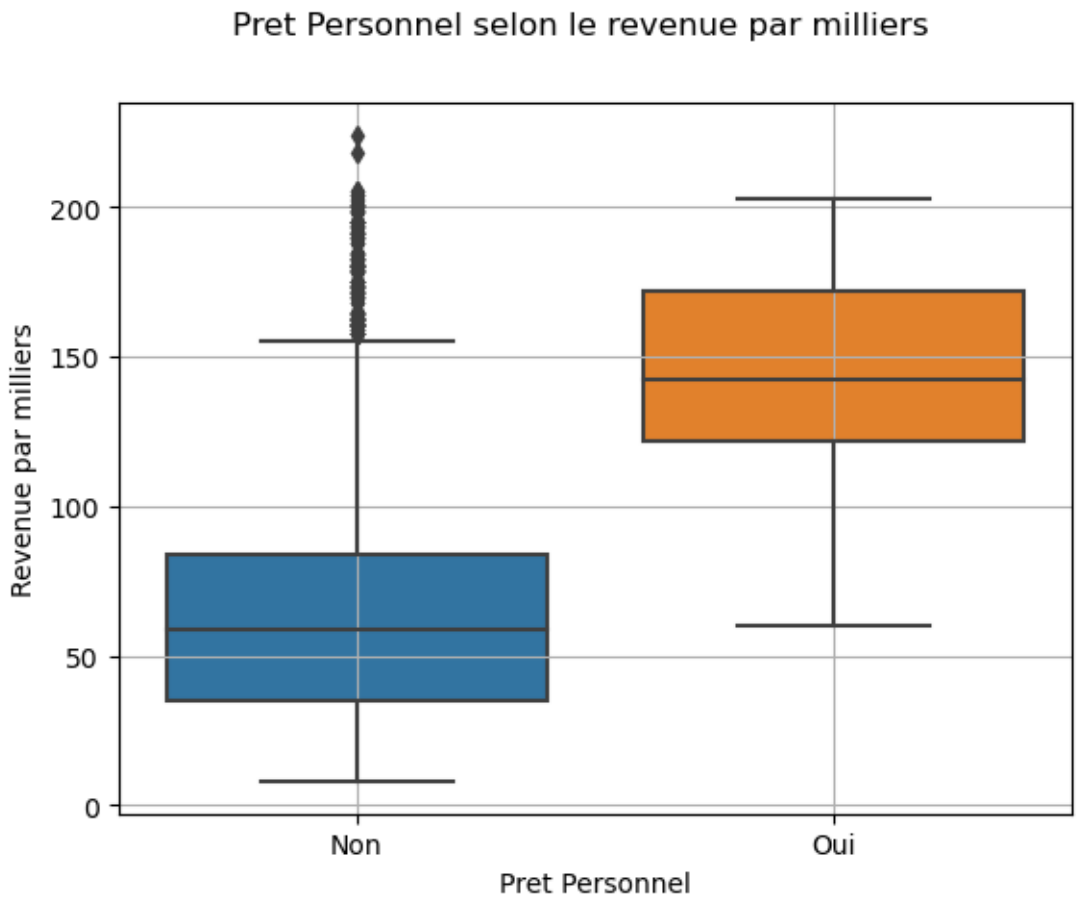
Dans cette partie nous allons étudier s’il existe une relation entre notre variable d’intérêt et nos variables explicatives.

Relation entre l'acquisition d'un prêt personnel et le revenu mensuel par milliers.

Nous constatons qu'il existe une relation entre ses deux variables en se basant sur les tests statistiques appropriés.

La boxplot et le tableau ci-dessous, nous montre que généralement les clients qui ont obtenu un prêt personnel ont un revenu plus élevé par rapport aux clients qui n'ont pas obtenu un prêt personnel.

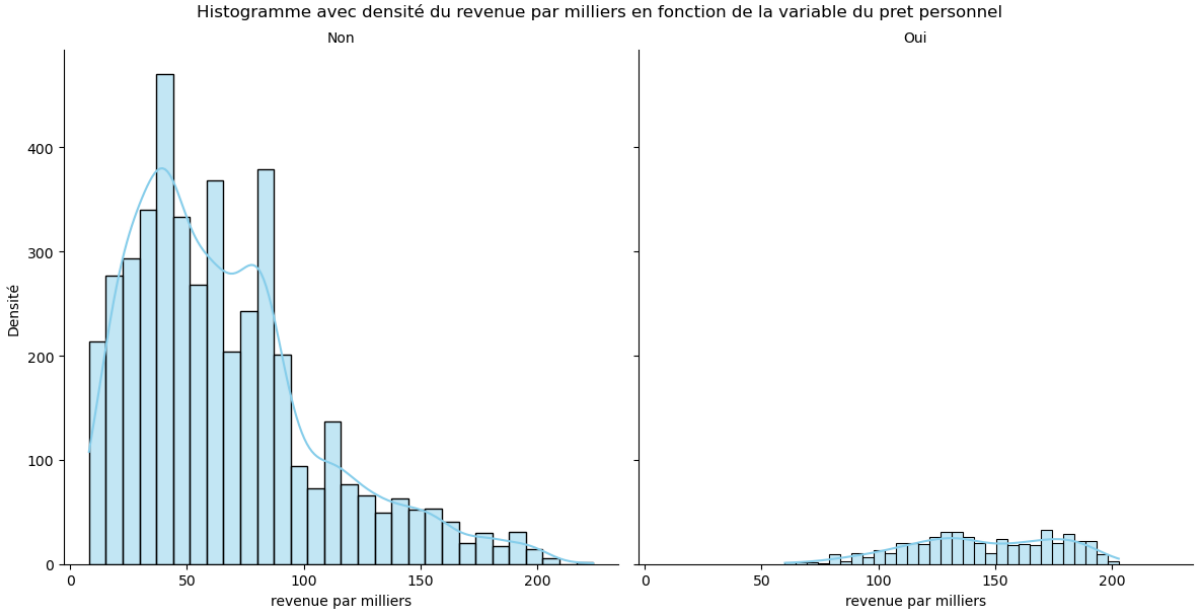
	count	mean	std	min	25%	50%	75%	max
PP								
Non	4415.0	65.985730	40.456626	8.0	35.0	59.0	84.0	224.0
Oui	479.0	144.647182	31.543335	60.0	122.0	142.0	172.0	203.0



La boxplot en dessus à gauche qui présente la non-obtention du prêt personnel montre une moyenne faible au niveau du revenu par millier égale à 66 par rapport à la boxplot à droite égal 144, la même chose pour les médianes, il suggère que plus le revenu est élevé plus la probabilité

d’obtenir un prêt est élevée et dans aucun cas un revenu par millier inferieur a 60 ne permet d’obtenir un prêt personnel.

Les deux histogrammes suivant nous donneront une vue plus claire sur la distribution des clients qui ont ou non obtenu un prêt par rapport à leurs revenus par milliers. La distribution du revenu est multimodale pour les deux histogrammes. Pour l’histogramme à gauche qui présente la non-obtention du prêt, il existe 3 pics avec une asymétrie étalée à droite dont la majorité des observations sont entre 10 et 100, ce qui veut exprimer que plus votre revenu est faible, plus votre probabilité d’avoir un prêt est faible. Pour l’histogramme à droite qui présente l’obtention du prêt personnel, il est multimodal aussi, avec une distribution cyclique, étalé vers la gauche et présente plusieurs pics. Vous allez bien constater qu’il n’existe aucune personne qui a obtenu un prêt avec un revenue inferieur a 50 000, ce qui suggère que plus les revenus sont élevés, plus les chances d'obtenir un prêt sont élevées.

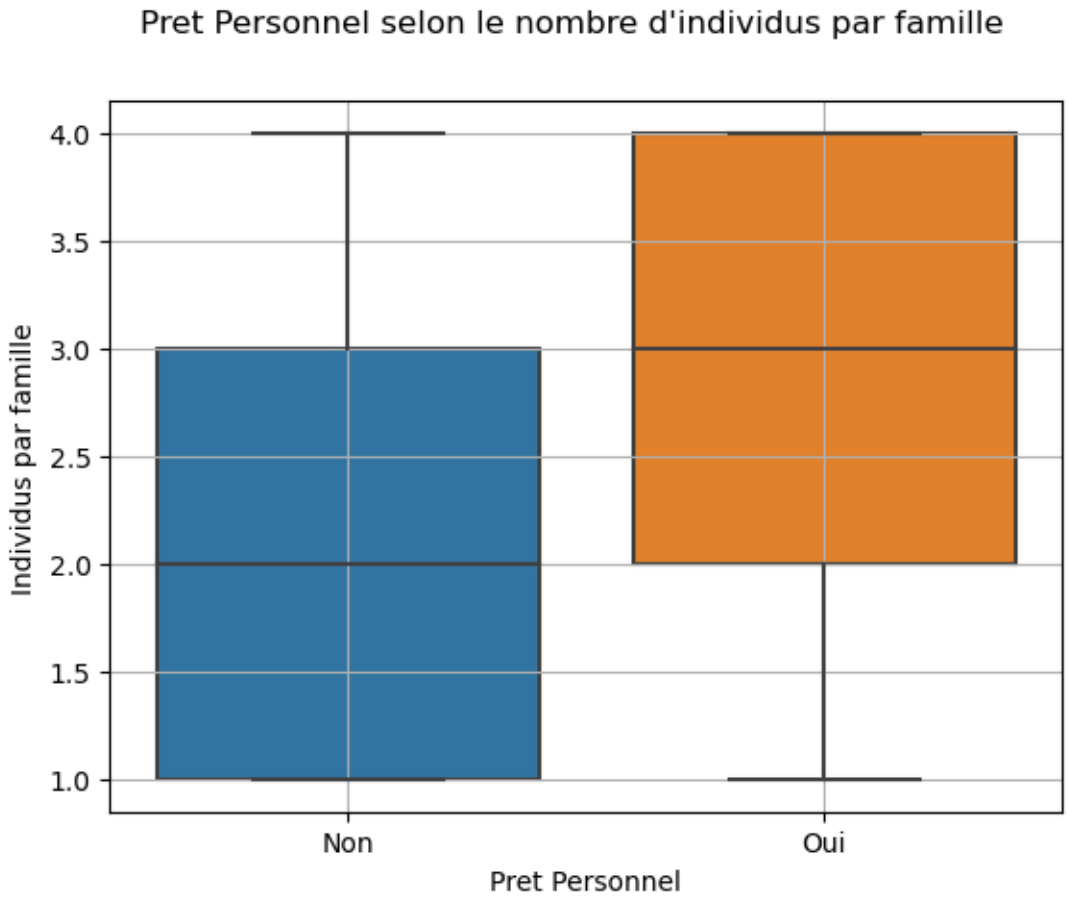


Relation entre l’acquisition d’un prêt personnel et le nombre de personnes par famille.

	count	mean	std	min	25%	50%	75%	max
PP								
Non	4415.0	2.382559	1.150723	1.0	1.0	2.0	3.0	4.0
Oui	479.0	2.615866	1.114116	1.0	2.0	3.0	4.0	4.0

Les tests statistiques appropriés nous ont montré qu'il existe une relation entre l'obtention d'un prêt personnel et le nombre d'individu par famille.

La boxplot en dessous à gauche qui présente la non-obtention du prêt personnel montre une moyenne au niveau du nombre de personne par famille plus faible égal à 2.38 par rapport à 2.61 pour la boxplot a droite, une médiane plus faible égale à 2 par rapport à la boxplot à droite égal 3, il suggère que plus le nombre de personne par famille est élevé plus la probabilité d'obtenir un prêt est élevée.

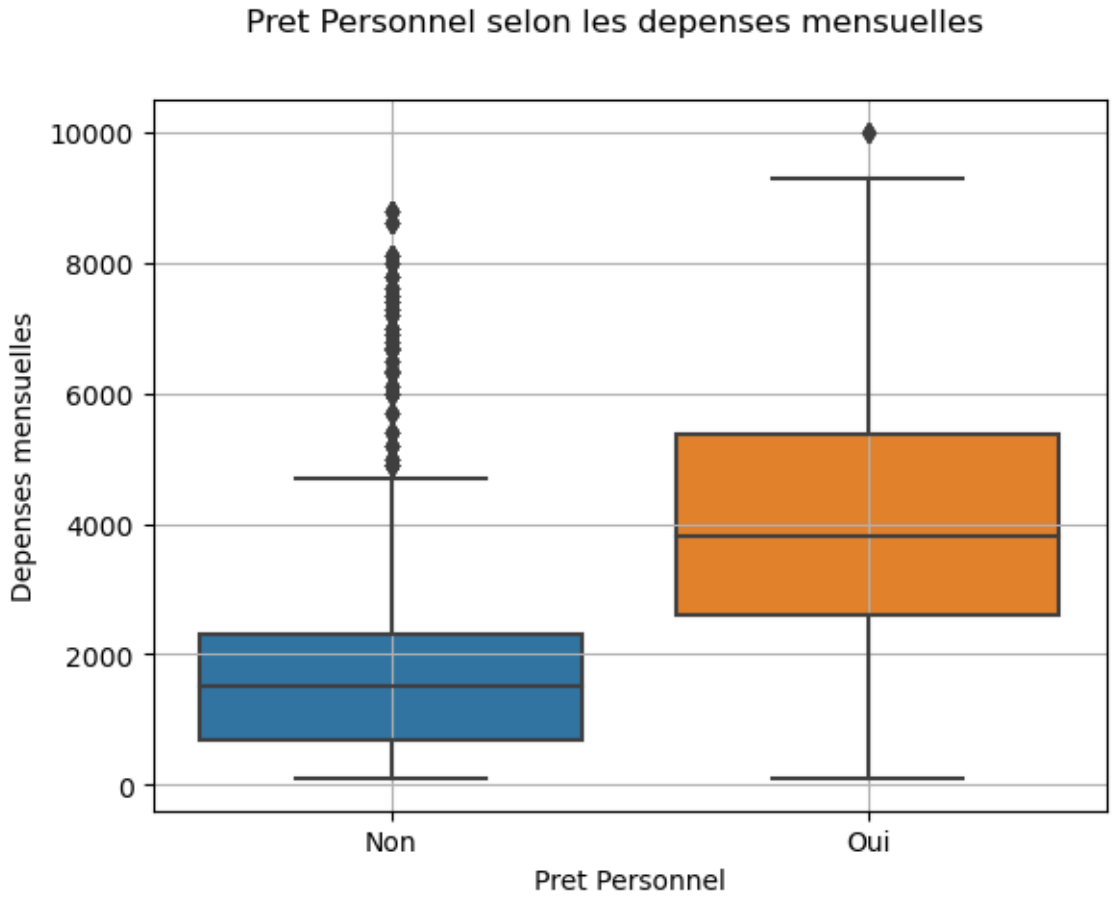


Relation entre l'acquisition d'un prêt personnel et les dépenses mensuelles.

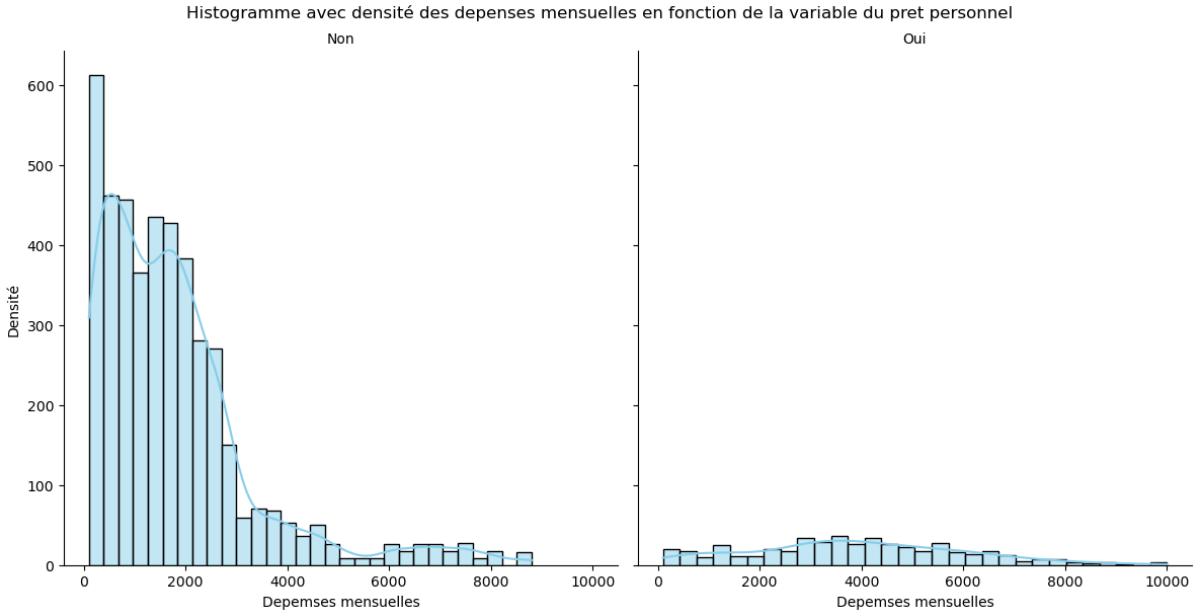
	count	mean	std	min	25%	50%	75%	max
PP								
Non	4415.0	1770.129105	1563.065215	100.0	700.0	1500.0	2300.0	8800.0
Oui	479.0	3913.507307	2092.247049	100.0	2600.0	3800.0	5365.0	10000.0

Les tests statistiques appropriés nous ont montré qu'il existe une relation entre l'obtention d'un prêt personnel et les dépenses mensuelles.

La boxplot en dessous à gauche qui présente la non-obtention du prêt personnel montre une moyenne plus faible au niveau des dépenses mensuelles égal à 1770.12 par rapport à 3913.5 pour la boxplot à droite, une médiane plus faible égale à environ 1750 par rapport à la boxplot à droite égal environ 4000, la même chose pour les médianes, il suggère que plus le revenu est élevé plus la probabilité d'obtenir un prêt est élevée et dans aucun cas un revenu par millier inferieur a 60 ne permet d'obtenir un prêt personnel.



L'histogramme à gauche ci-dessous qui présente la non-obtention du prêt, nous confirme cette constatation, il est unimodal et étalé à droite avec un pic majeur dans les trois premières barres à gauches, la majorité des individus qui n'ont pas obtenu des prêts personnels ont des dépenses mensuelles en dessous de 4000. L'histogramme à droite ci-dessous qui présente l'obtention du prêt personnel, il est multimodal et étalé à droite, on distingue une stagnation des densités avec un léger pic au milieu dans les 3 barres à gauche de 4000.



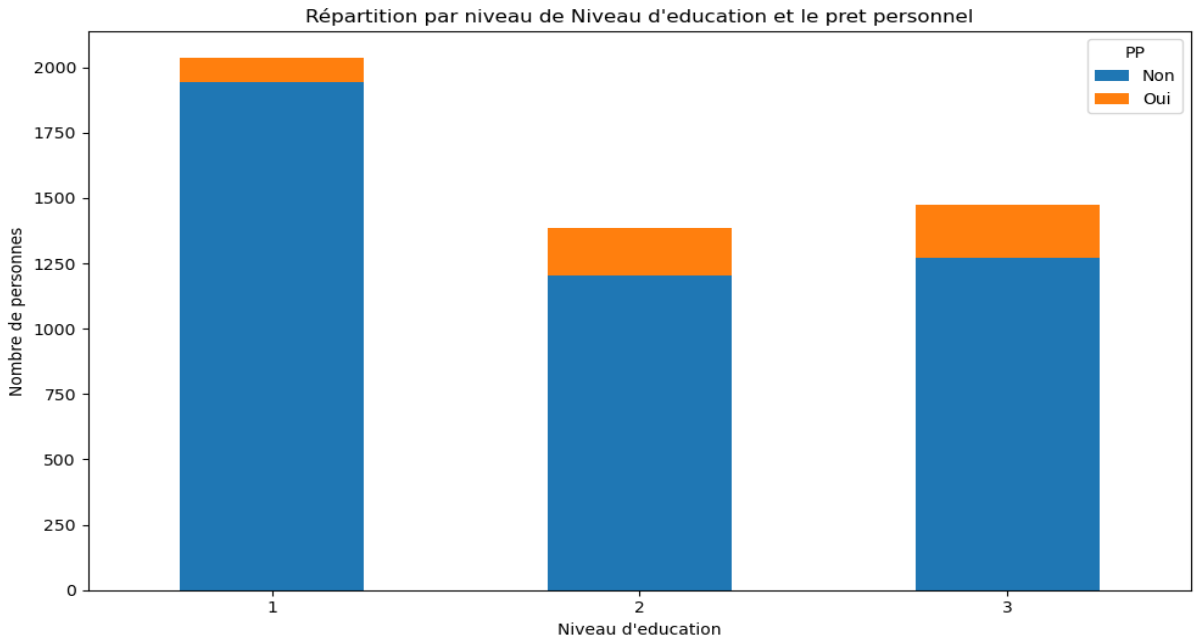
Relation entre l'acquisition d'un prêt personnel et le niveau d'éducation.

	count	mean	std	min	25%	50%	75%	max
PP								
Non	4415.0	1.847792	0.839352	1.0	1.0	2.0	3.0	3.0
Oui	479.0	2.233820	0.754085	1.0	2.0	2.0	3.0	3.0

Les tests statistiques appropriés nous ont montré qu'il existe une relation entre l'obtention d'un prêt personnel et les dépenses mensuelles.

Le tableau ci-dessus, nous montre que la moyenne au niveau d'éducation des non-détenteurs du prêt personnel est plus faible égal à 1.84 par rapport à 2.23 chez les détenteurs du prêt personnel.

Le diagramme empilé en dessous nous montre que plus votre niveau d'éducation est moyen (2) ou faible (3), la probabilité d'avoir un prêt est supérieur.



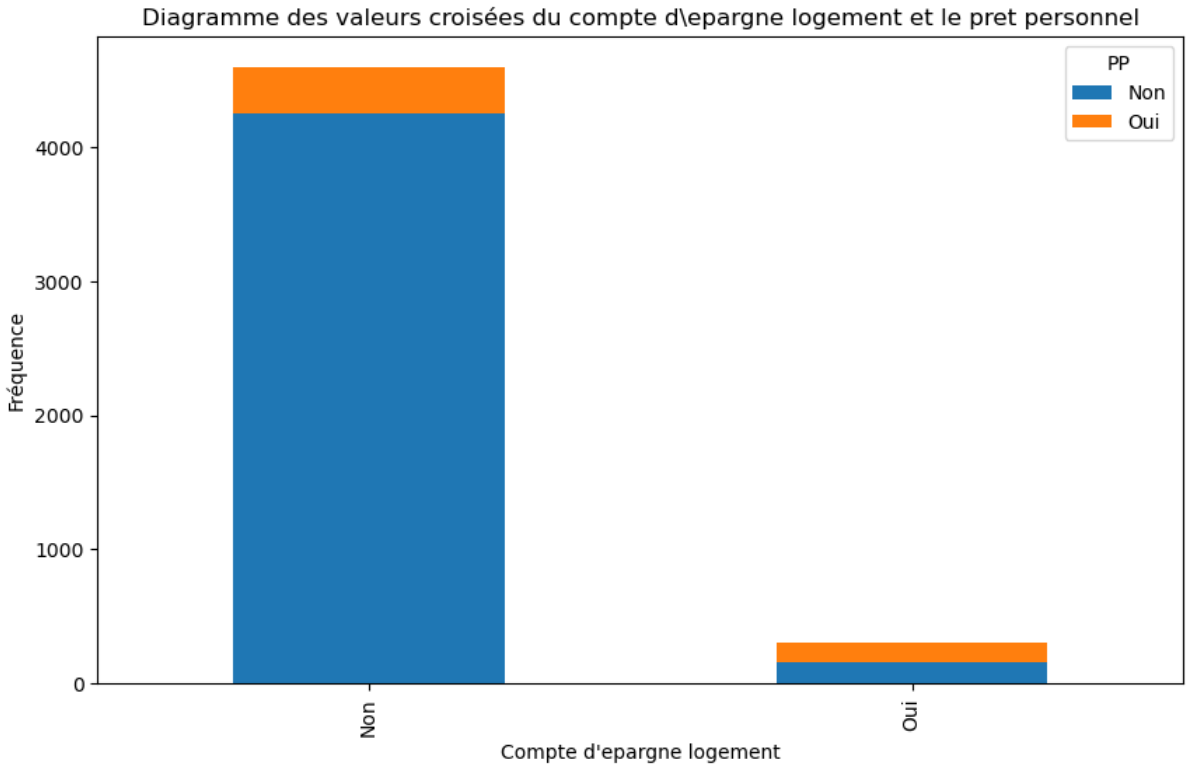
Relation entre l'acquisition d'un prêt personnel et la disposition d'un compte épargne logement.

Les tests statistiques appropriés nous ont montré qu'il existe une relation entre l'obtention d'un prêt personnel et les dépenses mensuelles.

Le tableau ci-dessous, nous montre que l'acquisition d'un compte d'épargne logement augmente le pourcentage d'obtention d'un prêt personnel de 7,9% à 46,6% au cas où le client a un compte d'épargne logement.

PP	Non	Oui
CEL		
Non	4255	339
	92%	7,9%
Oui	160	140
	53,3%	46,6%

Le diagramme empilé en dessous nous montre que la probabilité d'avoir un prêt personnel augmente quand le client a déjà un compte d'épargne logement.



Relation entre l'acquisition d'un prêt personnel et l'âge.

Il n'existe aucune relation entre l'obtention de prêt personnel et l'âge et ça été validé avec le test approprié.

Relation entre l'acquisition d'un prêt personnel et les années d'expérience.

Il n'existe aucune relation entre l'obtention de prêt personnel et les années d'expérience et ça été validé avec le test approprié.

Conclusion

Les variables influençant l'obtention d'un prêt personnel sont :

- Le revenu par milliers
- Le nombre d'individu par famille
- Les dépenses mensuelles

- Le niveau d'étude
- L'acquisition d'un compte d'épargne logement

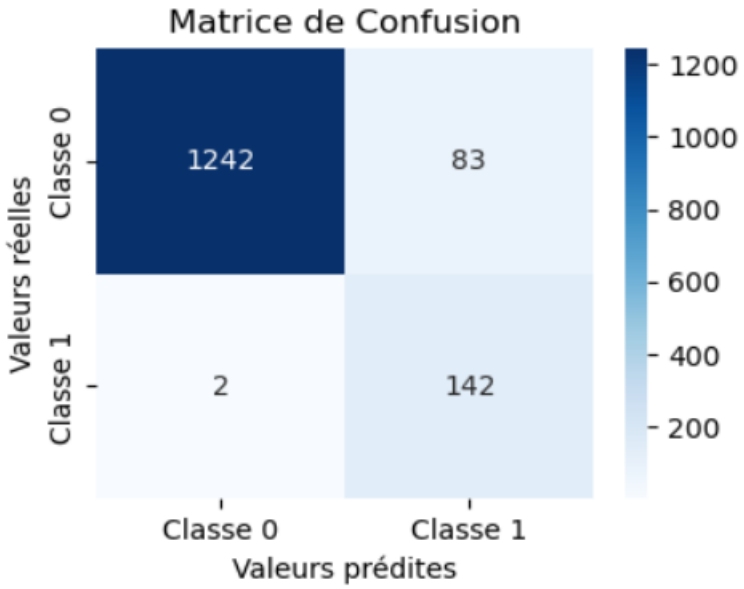
Choix de modèle

	Exactitude	Précision	Rappel	AUC
Logit	0.8965	0.485	0.916	0.9657
Probit	0.8965	0.485	0.916	0.9658
Log Log	0.9128	0.534	0.875	0.9604
Arbres	0.94213	0.631	0.9861	0.9811
SVM	0.8892	0.466	0.906	0.9603

Pour choisir le modèle le mieux approprié, nous avons comparé entre les modèles ci-dessus. Notre base de comparaison était de savoir lequel de ses modèles aux paramètres (Exactitude, précision, rappel et AUC) les plus élevés surtout les deux derniers sachant que l'objectif de notre travail est de prédire l'obtention ou la non-obtention d'un prêt personnel grâce à nos variables explicatives.

Pour le fait les arbres sont le modèle le plus optimale pour les résultats qui donne, et plus précisément son rappel qui permet de mesurer le nombre de prévisions positives correctes sur le nombre total de données positives et AUC qui nous permet de voir combien de toutes les classes positives et négatives, combien parmi elles ont été prédites correctement.

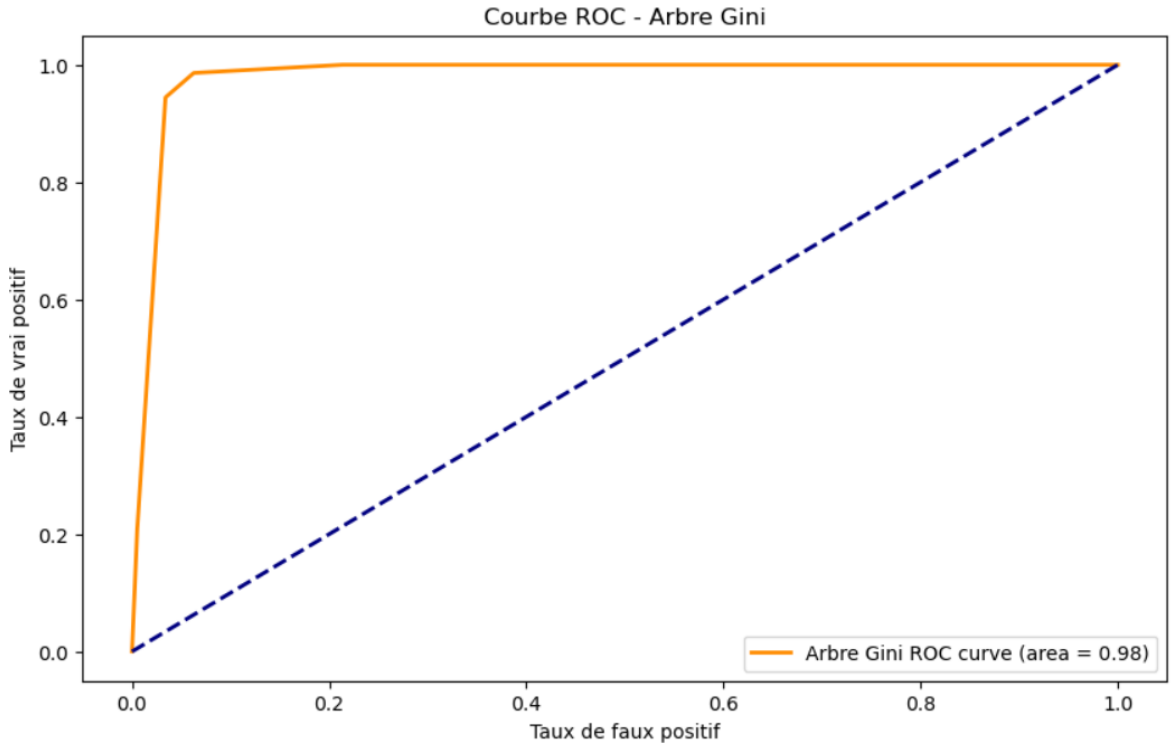
Matrice de confusion



La matrice de confusion ci-dessus nous donne une vision a quel point notre modèle est bien entrainé, il a pu prédire correctement 1242 observations, par rapport à 83 non correcte pour la classe 0 qui présente la non-obtention du prêt personnel, et qui montre que ses prédictions pour

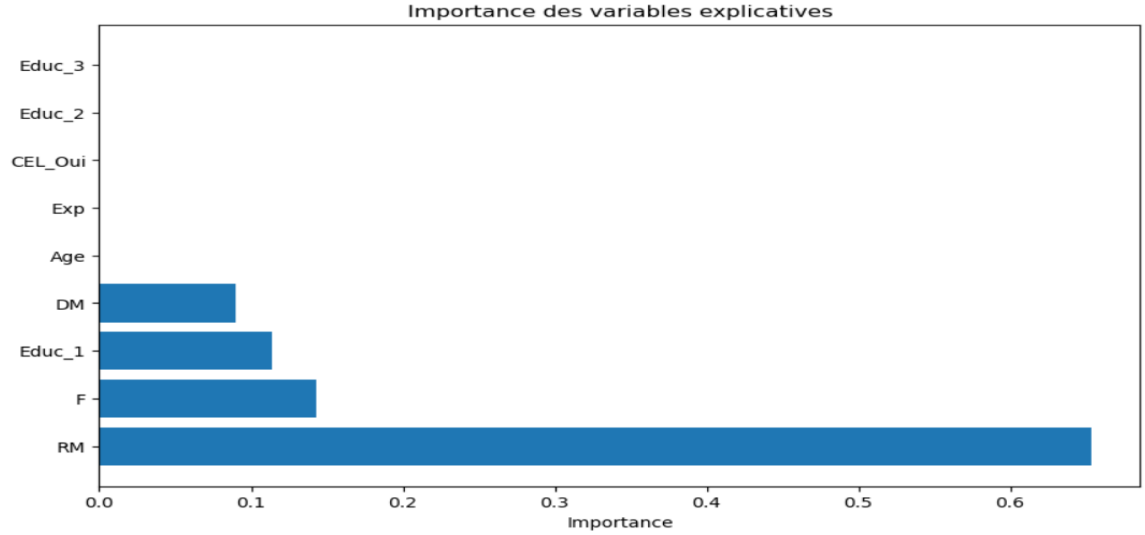
cette classe ont 94% juste. Pour la classe 1, elle a prédit correctement 142 observations contre 2 erreurs, qui donne un pourcentage très élevé égal 98,61% pour la classe 1 qui concerne les clients qui ont obtenu le prêt personnel.

Courbe ROC



Dans le graphe ci-dessus, la droite jaune ponche vers le 1, elle présente un angle 90 degré presque. Cet état de courbe nous montre que notre test est presque parfait au niveau de la précision.

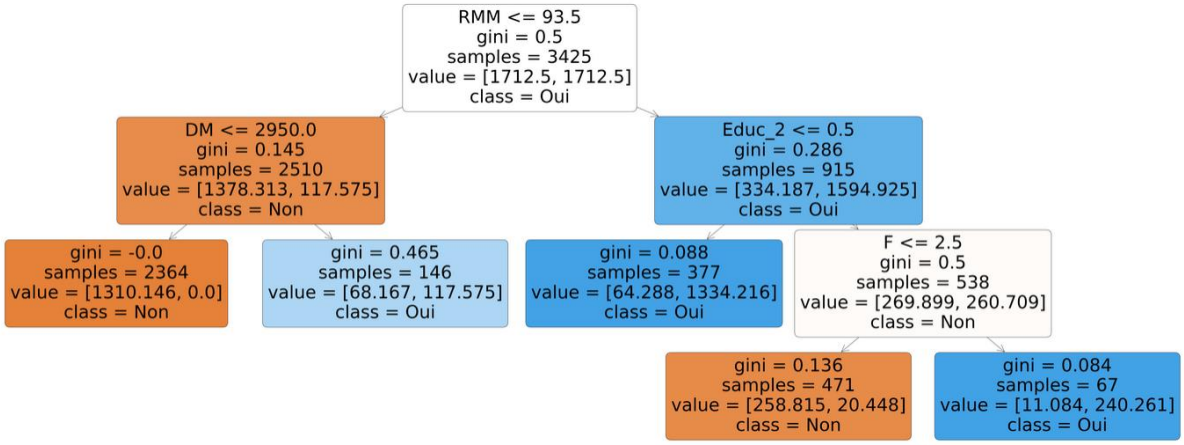
Importance des variables



Dans notre analyse, la variable de prédiction la plus importante est le revenu mensuel. Si la contribution de la principale variable de prédiction, revenu mensuel, est d'environ 70 %, alors vous pouvez la comparer aux autres variables afin de déterminer leur importance. Ainsi, vous pouvez vous concentrer sur les prédicteurs les plus importants. La liste suivante décrit les variables les plus importantes de cet arbre :

- Le nombre de personnes en famille est d'environ 15 %.
- Le niveau d'éducation élevé (Education = 1) est de 12 %.
- La dépense mensuelle est d'environ 10 %.

Arbre de décision



L'arbre de décision est un algorithme qui se base sur la discrimination. Les carrés bleus représentent l'obtention du prêt et les oranges représentent la non-obtention du prêt personnel. Quand la condition est respectée, nous basculant vers la condition en bas gauche et le contraire quand c'est faux.

Pour faire plus simple nous avons pris une capture d'écran des résultats du clients numéro 9, selon notre base de données, il n'a pas de prêt personnel.

ID	Age	Experience	Revenu Millie	Famille	Dépense men	Education	Compte Epar	Prêt Personnel
9	35	10	81	3	600	2	Non	Non

Soi-disant que nous ne connaissons pas les résultats du prêt personnel pour ses deux clients, nous allons donc essayer de suivre les conditions de l'arbre ci-dessous et de prédire les résultats du prêt personnel, et à la fin comparer si nos prédictions sont juste ou non.

La première question à poser toujours en se basant sur le graphique ci-dessus est si le revenu mensuel par millier du client 9 est inferieur a 93.5. Dans ce cas, la réponse est oui, donc nous allons procéder à la vérification de la condition à gauche pour vérifier si ses dépenses mensuelles sont inferieur a 2950. La réponse est toujours oui donc nous allons basculer au carré

orange qui nous informe que le client 9 n'aura pas de prêt personnel. Ce résultat de l'arbre doit être comparer avec le résultat de notre base de données dont nous vous avons mis une capture d'écran ci-dessus afin de savoir si notre arbre donne les bonnes prédictions. Tout à fait les résultats de l'arbre et de la base de données sont identiques, on peut dire que l'arbre donne de bonnes prédictions.

Prédiction de la nouvelle base de données

Nous allons essayer les résultats d'obtention du prêt personnel pour une nouvelle base de données dont on n'a pas les résultats de cette variable. Le résultat ci-dessous présente les probabilités de chaque client de la nouvelle base de données à obtenir un prêt personnel. Le client 3 par exemple, il a une probabilité de 95.40 % d'obtenir un prêt personnel ainsi de suite pour les autres clients.

	Non	Oui
3	0.045969	0.954031
13	0.045969	0.954031
10	0.045969	0.954031
6	0.045969	0.954031
0	0.367001	0.632999
8	0.926779	0.073221
21	0.926779	0.073221
18	0.926779	0.073221
17	0.926779	0.073221
25	1.000000	0.000000
26	1.000000	0.000000
23	1.000000	0.000000
22	1.000000	0.000000
27	1.000000	0.000000
28	1.000000	0.000000
20	1.000000	0.000000
19	1.000000	0.000000

Conclusion

En conclusion, notre analyse de variables en première partie, nous a permis de définir les variables qui ont impact sur l'obtention de prêt personnel, en deuxième partie grâce à l'algorithme de machine Learning, nous avons pu prédire les résultats de l'obtention de prêt personnel, d'identifier les variables les plus importantes et qui forme les conditions majeurs de l'obtention du prêt tel que le revenue mensuel par millier et le nombre d'individus par famille et créer un arbre de décision qui illustre tout cet algorithme. Les résultats de la matrice de confusion et du tableau de paramètre nous ont montré l'efficacité de l'algorithme et le modèle des arbres de décision, ce qui a été confirmé par la comparaison entre les résultats de l'arbre et les résultats de la base de données dans la section arbre de décision.