Prédire l'admissibilité au prêt

Dream Housing Finance Co Ltd.

YOUR ROAD TO A DREAM HOME

D.A.
Marc Sellam

Projet 8

Décembre 2020

TABLE DES MATIÈRES

1. INTRODUCTION

Les banques en lignes font une concurrence tarifaire féroce aux banques classiques, celles-ci se doivent donc d'être très réactives face aux demandes de leurs clients, notamment dans le cadre des demandes de crédit.

La création d'un outil permettant à un conseiller clientèle de fournir rapidement un accord de principe à une demande de crédit, lui permettra de garder son client et dans un second temps de justifier les tarifs un peu plus élevés de la banque, par son efficacité.

L'objectif étant de fournir une réponse binaire à cette demande de crédit en fonction des informations concernant le client, un algorithme de classification binaire sera utilisé.

2. LES DONNÉES FOURNIES

Le jeu de données fourni est un tableau de 614 lignes, chacune représentant une demande de prêt. Chaque ligne est composée de 13 colonnes :

-1 colonne contenant l'identification unique de la demande de prêt.

-11 colonnes numériques, contenant les caractéristiques des candidatures.

-1 colonne contenant la décision de la banque Y/N (accordé/refusé).

2.1 Origine des données

Les données ont été récupérées sur le site internet analyticsvidhya.com, un site internet de compétition en data science, les données datent du 26-05-2016 à 12 :01. :

https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/#ProblemStatement

2.2 Choix des données

Mesurer mes connaissances par la compétition avec d'autres et m'exercer au travers ce type d'exercice d'analyse typique de ce qui est demandé en entreprise, ont motivé mon choix pour cet échantillon de données.

2.3 Définitions des variables

Définissons dans un premier temps chacune des variables fournies par la banque afin de savoir de quels types d'informations nous disposons.

Variable	Description	Valeurs
Loan_ID	identification unique du prêt	LPXXXXXX
Gender	genre de l'emprunteur	['Male' 'Female']
Married	statut matrimonial de l'emprunteur	['No' 'Yes']
Dependents	nombre d'enfants de l'emprunteur	['0' '1' '2' '3+']
Education	niveau scolaire de l'emprunteur	['Graduate' 'Not Graduate']
Self_Employed	indépendant	['No' 'Yes']
ApplicantIncome	revenus de l'emprunteur	valeurs continues
CoapplicantIncome	revenus du co-emprunteur	valeurs continues
LoanAmount	montant du prêt	valeurs continues (en milliers)
Loan_Amount_Term	durée du prêt (en mois)	10 choix possibles (12 à 360)
Credit_History*	« cote » emprunteur	[1. 0.]
Property_Area	situation du bien	['Urban' 'Rural' 'Semiurban']
Loan_Status	(Target) état du prêt	['Y' 'N']

*Le "Credit History" ou historique de crédit :
Les américains utilisent énormément les cartes de crédit (plastic money) et le crédit. Pour bénéficier de cette possibilité, offerte par les banques ou certains magasins, il faut posséder ou construire son historique de crédit (credit history), sorte de cote, qui peut se comparer, dans une certaine mesure, au fichier de la Banque de France. Le principe est que plus l'on emprunte, à condition de rembourser régulièrement, plus la cote augmente et donc plus on obtient facilement un prêt (avec un taux intéressant), et plus on est sollicité par les organismes de crédit. Un américain débute le processus à seize ans.
source : https://sites.google.com/site/nlrfrance/1/affecte/a-l-arrivee/banque-us/credit-history

2.4 Types de variables

Les variables fournies se décomposent en 3 types :

Les variables qualitatives : 'Loan_ID', 'Gender', 'Married', 'Dependents', 'Education', 'Self_Employed', 'Property_Area', 'Loan_Status' (variables categorielles).

Les variables quantitatives continues :'CoapplicantIncome', 'LoanAmount', 'Loan_Amount_Term','Credit_History' .

Les variables quantitatives discrètes : 'ApplicantIncome' qui est un entier.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Loan_ID            614 non-null    object 
 1   Gender             601 non-null    object 
 2   Married            611 non-null    object 
 3   Dependents         599 non-null    object 
 4   Education          614 non-null    object 
 5   Self_Employed      582 non-null    object 
 6   ApplicantIncome    614 non-null    int64  
 7   CoapplicantIncome  614 non-null    float64
 8   LoanAmount         592 non-null    float64
 9   Loan_Amount_Term   600 non-null    float64
 10  Credit_History     564 non-null    float64
 11  Property_Area      614 non-null    object 
 12  Loan_Status        614 non-null    object 
dtypes: float64(4), int64(1), object(8)
memory usage: 62.5+ KB
```

2.5 Descriptions des variables

Maintenant que nous savons de quelle d'information nous renseigne chacune des variables, décrivons-les.

2.5.1 Les variables quantitatives

Les variables quantitatives 'ApplicantIncome', 'CoapplicantIncome' et 'LoanAmount' semblent contenir des données aberrantes.

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.000000	564.000000
mean	5403.459283	1621.245798	146.412162	342.000000	0.842199
std	6109.041673	2926.248369	85.587325	65.120410	0.364878
min	150.000000	0.000000	9.000000	12.000000	0.000000
25%	2877.500000	0.000000	100.000000	360.000000	1.000000
50%	3812.500000	1188.500000	128.000000	360.000000	1.000000
75%	5795.000000	2297.250000	168.000000	360.000000	1.000000
max	81000.000000	41667.000000	700.000000	480.000000	1.000000

La variable 'Loan_Amount_Term' n'a que 10 modalités :
Loan_Amount_Term : [360. 120. 240. 180. 60. 300. 480. 36. 84. 12.]

2.5.2 Les variables qualitatives

•'Loan_ID' correspondant à l'identifiant unique de chaque demande de crédit, il est normal qu'il soit composé d'autant de lignes que d'identifiants soit 614.

•'Gender' nous indique que la grande majorité des demandes sont faites par des hommes soit 489 sur 601, modalités : ['Male' 'Female'].

•'Married' nous indique que la majorité des demandes sont faites par des personnes mariées soit 398 sur 611, modalités : ['No' 'Yes'].

•'Dependents' nous indique que la majorité des demandes sont faites par des personnes sans enfant soit 345 sur 599, modalités : ['0' '1' '2' '3+'] ('3+'pour 3 enfants ou +).

•'Education' nous indique que la majorité des demandes sont faites par des personnes diplômées soit 480 sur 614, modalités : ['Graduate' 'Not Graduate'].

•'Self_Employed' nous indique que la grande majorité des demandes sont faites par des salariés soit 500 sur 582, modalités : ['No' 'Yes'].

•'Property_Area' nous indique que la majorité des demandes sont faites pour un bien situé en zone semi-urbaine soit 233 sur 614, modalités : ['Urban' 'Rural' 'Semiurban'].

•'Loan_Status' , notre variable cible (target), nous indique que la majorité des demandes de crédit ont été acceptées soit 422 sur 614, modalités : ['Y' 'N']

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	Property_Area	Loan_Status
count	614	601	611	599	614	582	614	614
unique	614	2	2	4	2	2	3	2
top	LP001482	Male	Yes	0	Graduate	No	Semiurban	Y
freq	1	489	398	345	480	500	233	422

3. PRÉPARATION DES DONNÉES

3.1 Optimisation des variables

Afin d'obtenir un bon fonctionnement des algorithmes que nous allons utiliser, il est nécessaire de remplacer les variables qualitatives par des valeurs numériques comme suit :

•'Loan_ID' correspondant à un identifiant unique, il restera inchangé.

•'Gender' dont les modalités sont ['Male', 'Female'], seront respectivement remplacés par les entiers 1 et 0.

•'Married' dont les modalités sont ['No', 'Yes'], seront respectivement remplacés par les entiers 0 et 1.

•'Dependents' dont les modalités sont ['0', '1', '2' ,'3+'] , seront respectivement remplacés par les entiers 0, 1, 2 et 3 (pour 3 enfants ou +).

•'Education' dont les modalités sont ['Graduate', 'Not Graduate'], seront respectivement remplacées par 1 et 0.

•'Self_Employed' dont les modalités sont ['No', 'Yes'], seront respectivement remplacées par 0 et 1.

•'Property_Area' dont les modalités sont ['Urban','Rural', 'Semiurban'], seront respectivement remplacées par 2, 0 et 1.

•'Loan_Status', notre variable cible (target dont les modalités sont ['Y', 'N'], seront respectivement remplacées par 1 et 0.

•'Credit_History', dont les modalités sont [1.0, 2.0], seront respectivement remplacées par les entiers 1 et 0.

3.2 Traitement des valeurs manquantes

Il existe des valeurs manquantes sur des variables qualitatives et quantitatives (surlignées en cyan), l'utilisation de l'algorithme KNNImputer avec k = 1 permettra de remplacer les valeurs manquantes par leur plus proche voisin, cet algorithme calculera la distance euclidienne de la valeur la plus proche afin d'affecter cette valeur à la valeur manquante.

Cet algorithme pend en charge les données quantitatives continues, quantitatives discrètes, ordinales et catégorielles, ce qui le rend particulièrement utile pour traiter tous les types de données manquantes.

Lorsque k=1, l'algorithme recherchera une observation similaire dont la valeur manquante que nous recherchons ne manque pas, et utilisera cette donnée.

```
Credit_History       50
Self_Employed        32
LoanAmount           22
Dependents           15
Loan_Amount_Term     14
Gender               13
Married               3
Loan_Status           0
Property_Area         0
CoapplicantIncome     0
ApplicantIncome       0
Education             0
Loan_ID               0
dtype: int64
```

3.2.1 Comparatif des valeurs quantitatives avant et après imputation

Apres remplacement des valeurs manquantes pour les variables 'LoanAmount' (22 sur 614) et 'Loan_Amount_Term' (14 sur 614), les valeurs moyennes, écart-types et médianes n'ont pas ou peu variés (surlignées en cyan).

Avant :

	LoanAmount	Loan_Amount_Term
count	592.000000	600.000000
mean	146.412162	342.000000
std	85.587325	65.120410
min	9.000000	12.000000
25%	100.000000	360.000000
50%	128.000000	360.000000
75%	168.000000	360.000000
max	700.000000	480.000000

Après :

	LoanAmount	Loan_Amount_Term
count	614.000000	614.000000
mean	146.084691	341.433225
std	85.730387	65.422442
min	9.000000	12.000000
25%	100.000000	360.000000
50%	127.000000	360.000000
75%	167.750000	360.000000
max	700.000000	480.000000

3.2.2 Comparatif des valeurs qualitatives avant et après imputation

Les comparatifs n'indiquent pas de changements importants après remplacement des données manquantes, l'algorithme a donc remplacé les valeurs manquantes avec succès.

Avant :

```
**********
Données d'origine :
Loan_Status        0        1
Gender
0.0          0.060261  0.122150
1.0          0.244300  0.552117
nan          0.008143  0.013029
**********
Données d'origine :
Loan_Status        0        1
Married
0.0          0.128664  0.218241
1.0          0.184039  0.464169
nan          0.000000  0.004886
**********
Données d'origine :
Loan_Status        0        1
Dependents
0            0.174267  0.387622
1            0.058632  0.107492
2            0.040717  0.123779
3            0.029316  0.053746
nan          0.009772  0.014658
**********
Données d'origine :
Loan_Status        0        1
Self_Employed
0.0          0.255700  0.558632
1.0          0.042345  0.091205
nan          0.014658  0.037459
**********
Données d'origine :
Loan_Status        0        1
Credit_History
0.0          0.133550  0.011401
1.0          0.157980  0.615635
nan          0.021173  0.060261
```

Après :

```
**********
Données complétées :
Loan_Status        0        1
Gender
0            0.060261  0.123779
1            0.252443  0.563518
**********
Données complétées :
Loan_Status        0        1
Married
0            0.128664  0.219870
1            0.184039  0.467427
**********
Données complétées :
Loan_Status        0        1
Dependents
0            0.179153  0.395765
1            0.063518  0.110749
2            0.040717  0.125407
3            0.029316  0.055375
**********
Données complétées :
Loan_Status        0        1
Self_Employed
0            0.268730  0.591205
1            0.043974  0.096091
**********
Données complétées :
Loan_Status        0        1
Credit_History
0            0.136808  0.022801
1            0.175896  0.664495
```

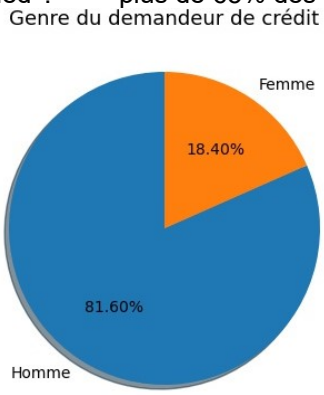
4. ANALYSES GRAPHIQUES

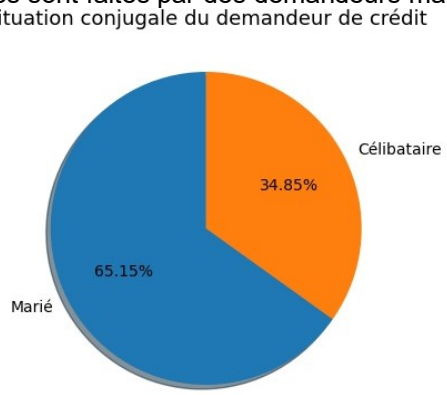
Maintenant que nous avons décrit nos données, passons aux analyses graphiques.

4.1 Analyses univariées

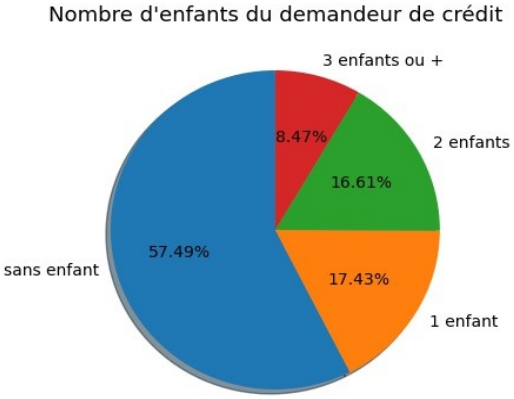
4.1.1 Les variables qualitatives

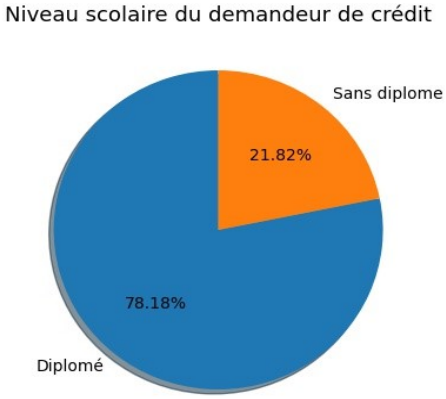
Variable 'Gender' : plus de 80% des demandes sont faites par des hommes.
Variable 'Married' : plus de 65% des demandes sont faites par des demandeurs mariés.



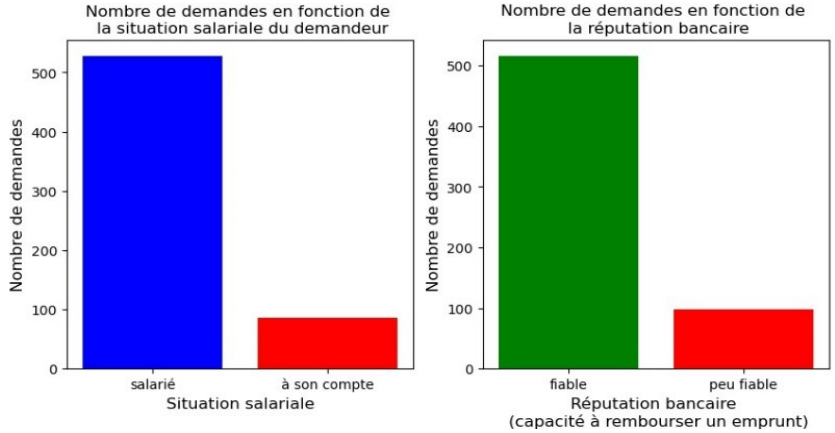


Variable 'Dependents' : la majorité des demandes sont effectuées par des demandeurs sans enfants.
Variable 'Education' : la grande majorité des demandeurs sont diplômés.

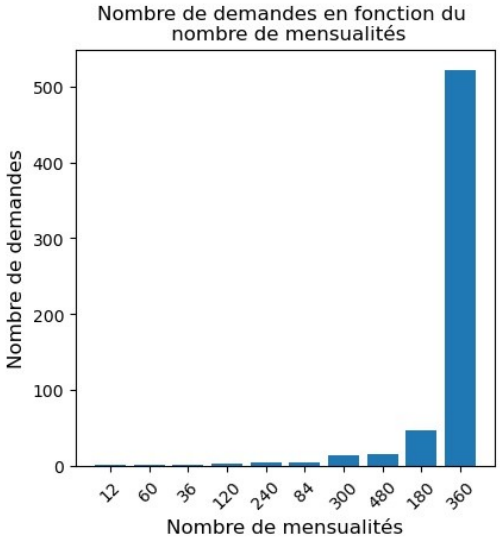


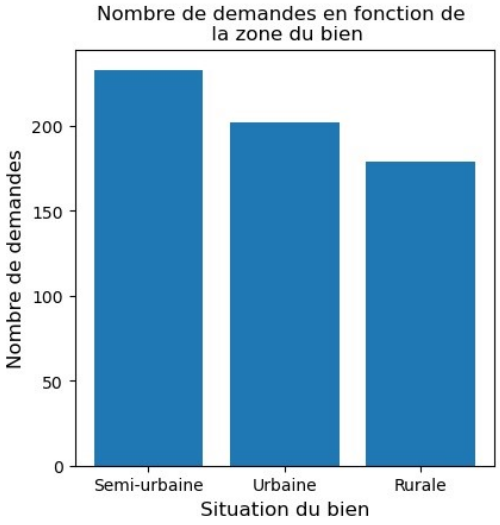


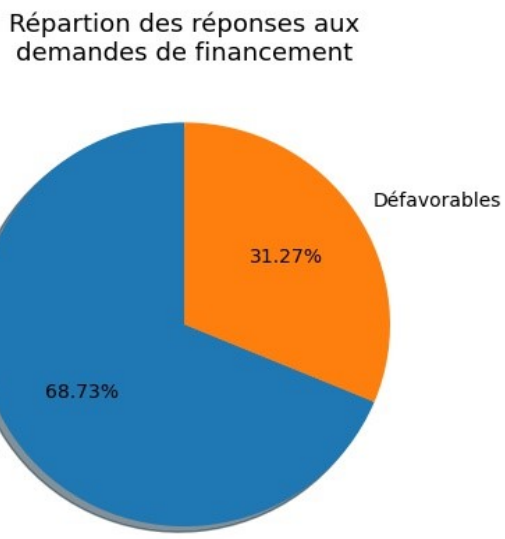
Variable 'Self_Employed': La grande majorité des demandeurs sont des salariés.
Variable 'Credit_History' : la grande majorité des demandeurs ont une réputation bancaire favorable.



Variable 'Loan_Amount_Term' : la grande majorité du nombre mensualités demandées est de 360 mois soit 30 ans.
Variable 'Property_Area' : un nombre légèrement plus important des demandes sont situés en zones semi-urbaines.





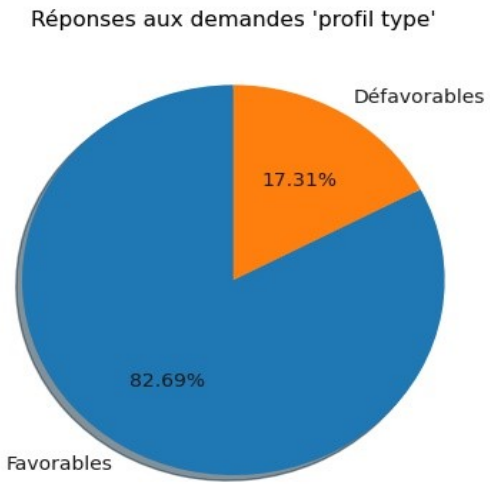


Variable 'Loan_Status':
La majorité des demandes obtiennent une réponse favorable.

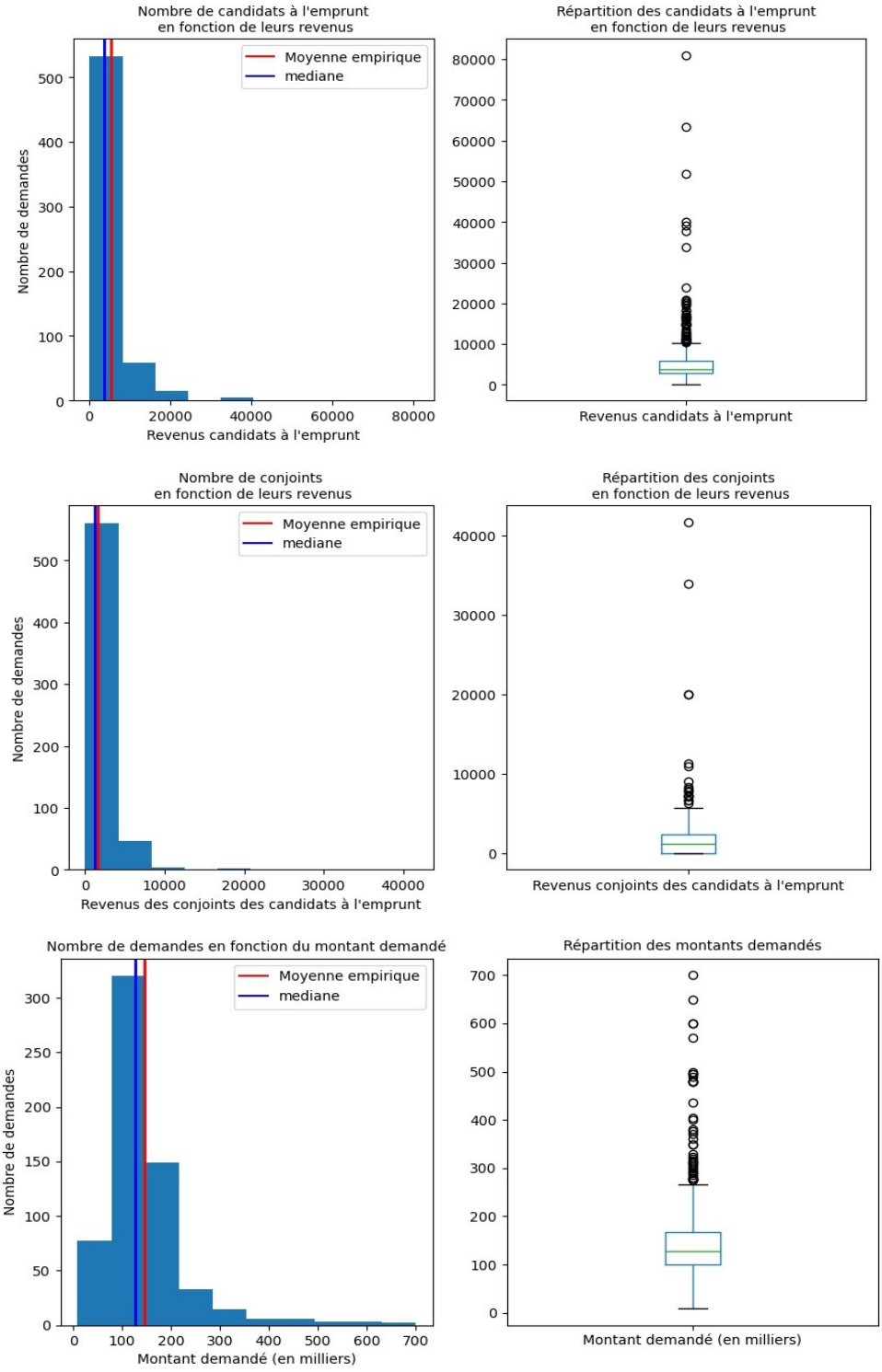
Le 'profil type' du candidat à l'emprunt est un homme diplômé, marié, sans enfant, salarié et qu'il a une « réputation bancaire » (cote) favorable à sa demande de crédit, ce profil a plus de 82% de chance d'obtenir une réponse favorable à sa demande.

Il demande dans plus de 88% des cas un rembour sement sur 30 ans.

Ce profile représente 104 cas sur 614 soit un peu moins de 17% des candidatures.



4.1.2 Les variables quantitatives :



Les graphiques ci-dessus à gauche montrent que les distributions de chacune des variables sont asymétriques à droite et que des valeurs extrêmement élevées se trouvent à droite de l'histogramme. Les boîtes à moustaches ci-dessus à droite affichent la distribution de chacune des variables, les cercles indiquent les valeurs aberrantes.

La variable ‘LoanAmount’ qui représente le montant du prêt demandé, semble présenter une boîte à moustaches qui n’est pas loin d’être équilibrée.

Valeurs d’asymétrie (skewness):

ApplicantIncome	6.539513
CoapplicantIncome	7.491530
LoanAmount	2.666462

Une des techniques d'identification des valeurs aberrantes est le Score IQR, cette technique utilise les scores IQR (IQR = Q3 - Q1) pour identifier les valeurs aberrantes. La règle de base est que tout ce qui n'est pas dans la plage de (Q1 - 1,5 IQR) et (Q3 + 1,5 IQR) est une valeur aberrante, et dans le cas de nos trois variables :

Variable ‘ApplicantIncome’ (revenus de l’emprunteur) :	nombre de valeurs aberrantes : 50
Variable ‘CoapplicantIncome’ (revenus du co-emprunteur) :	nombre de valeurs aberrantes : 18
Variable ‘LoanAmount’ (montant du prêt) :	nombre de valeurs aberrantes : 40

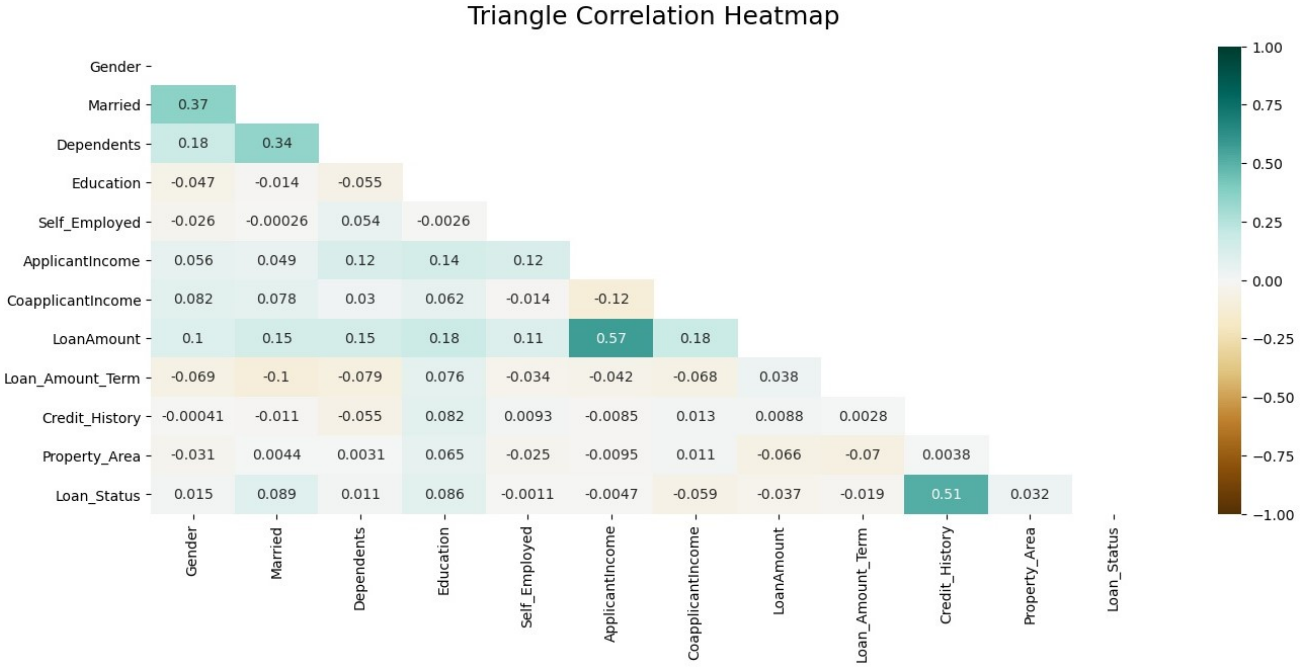
4.2 Analyses bivariées

Voyons graphiquement quelles variables ont une influence sur la décision de la banque à accorder un prêt, et dans quelle mesure.

4.2.1 Tableau des corrélations

Le tableau des corrélations nous indique une corrélation r² de 0.57 entre les revenus de l’emprunteur et le montant du prêt demandé, ainsi qu’une corrélation r² de 0.51 entre la « cote » bancaire de l’emprunteur (Credit_History) et la décision de la banque (Loan_Status), notre variable cible.

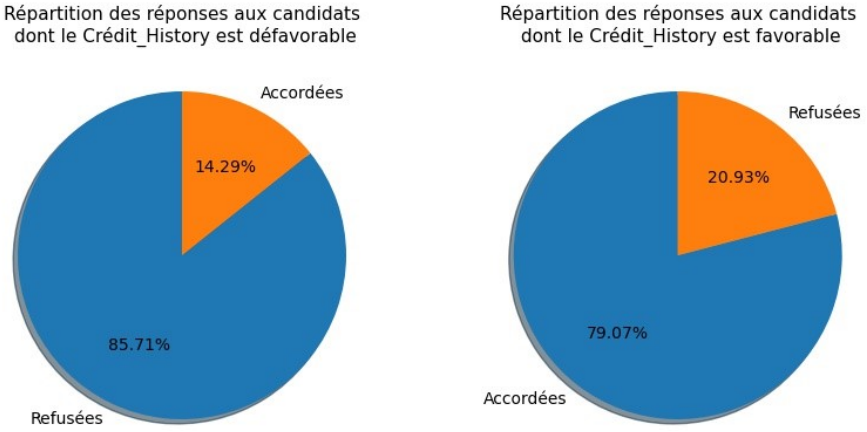
La corrélation r² entre la décision de la banque et la situation salariale du demandeur ‘Self_Employed’ est presque nulle (0.0011).



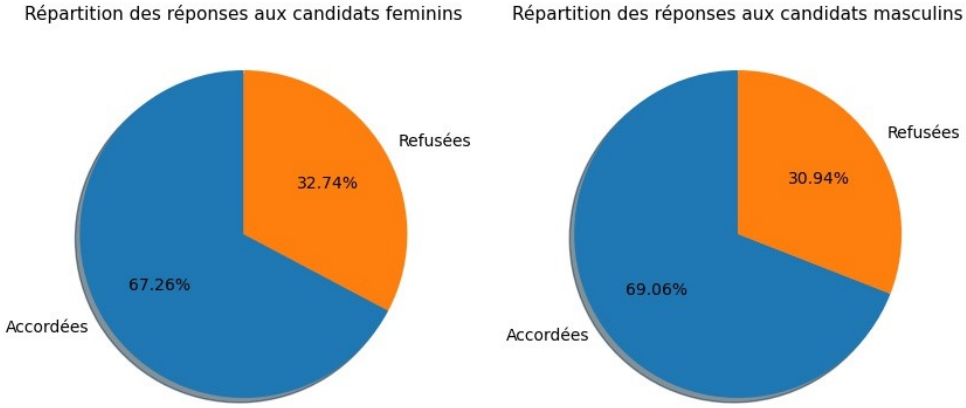
4.2.2 Les variables qualitatives

La corrélation r² entre ‘Credit_History’ et la réponse à la demande de prêt du tableau des corrélation est confirmé par ces deux graphiques, un ‘Credit_History’ favorable garanti à 79% l’obtention du prêt au candidat à l’emprunt alors que lorsque qu’il est défavorable, le candidat aura plus de 85% de chances que sa demande se solde par un refus.

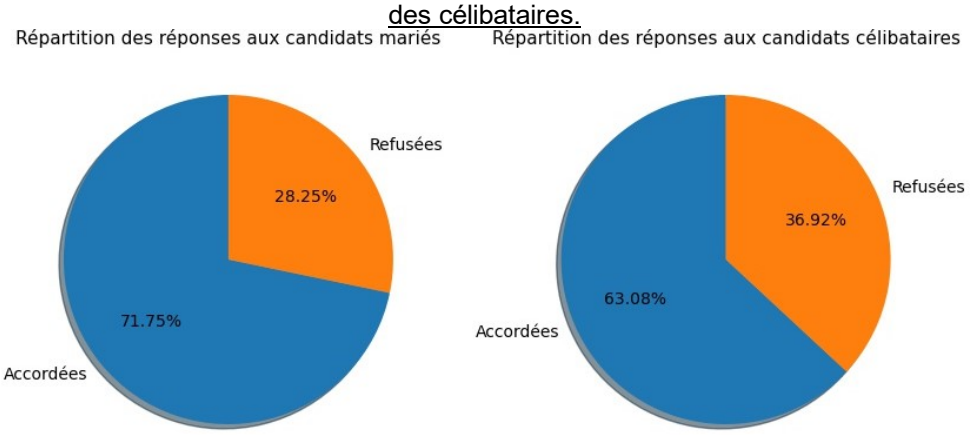
Le candidat aura un peu plus de 17% de chance de ne pas obtenir le prêt si le ‘Credit_History’ est favorable, et un peu plus de 2% de chance d'obtenir un prêt si son ‘Credit_History’ est défavorable.



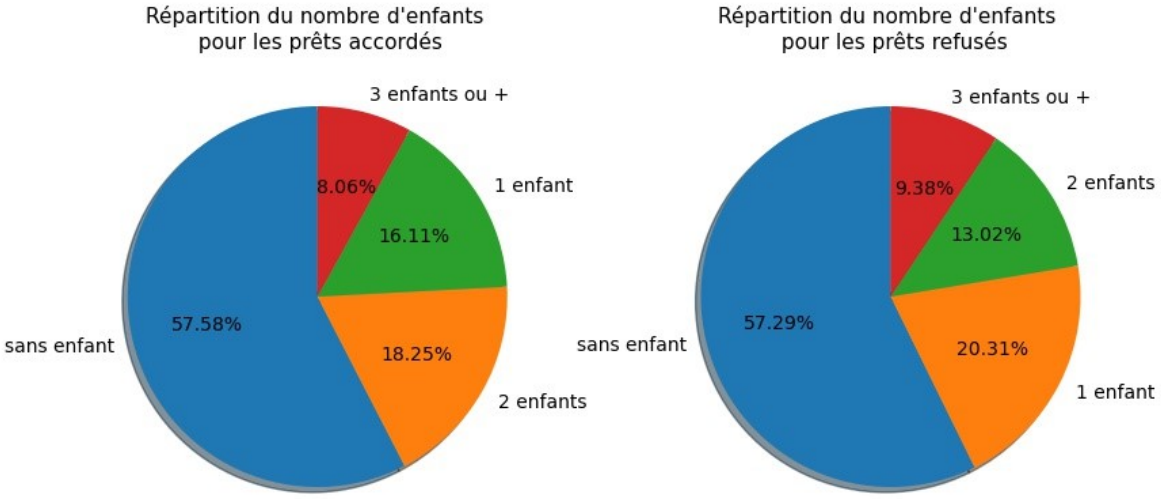
La répartition des réponses est similaire d’un genre à l’autre.



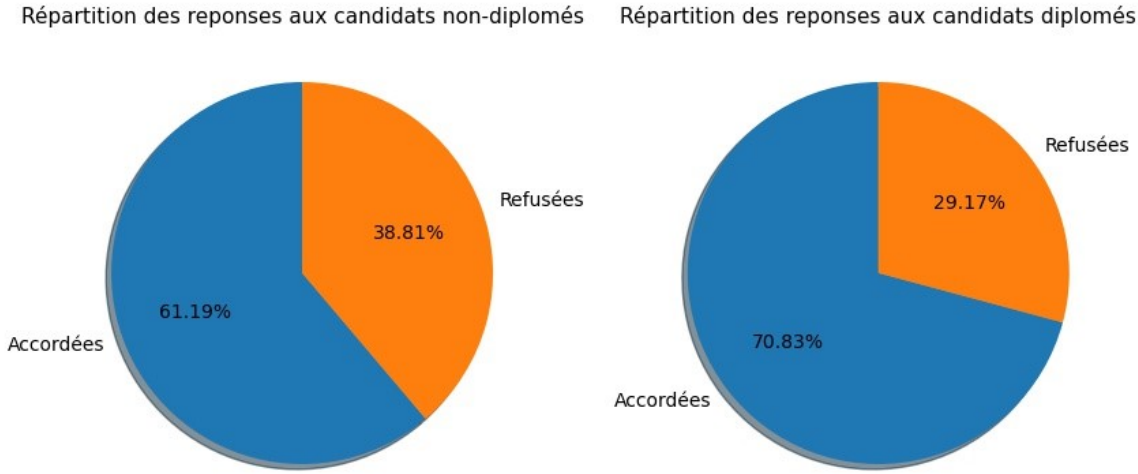
La répartition des réponses favorables aux candidats mariés est un peu plus importante que celle des célibataires.



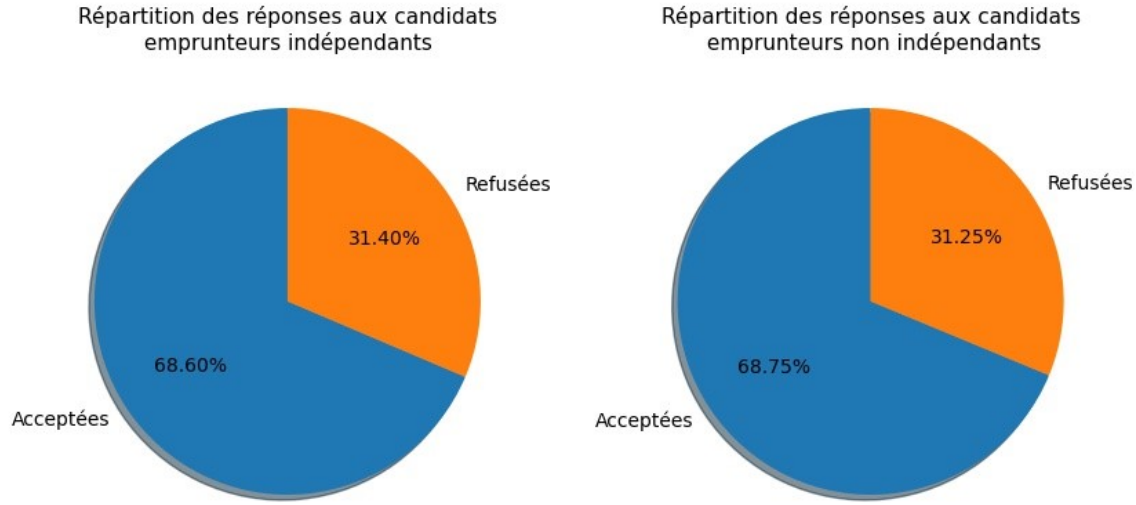
La répartition du nombre d'enfants est similaire selon que la réponse soit favorable ou non.



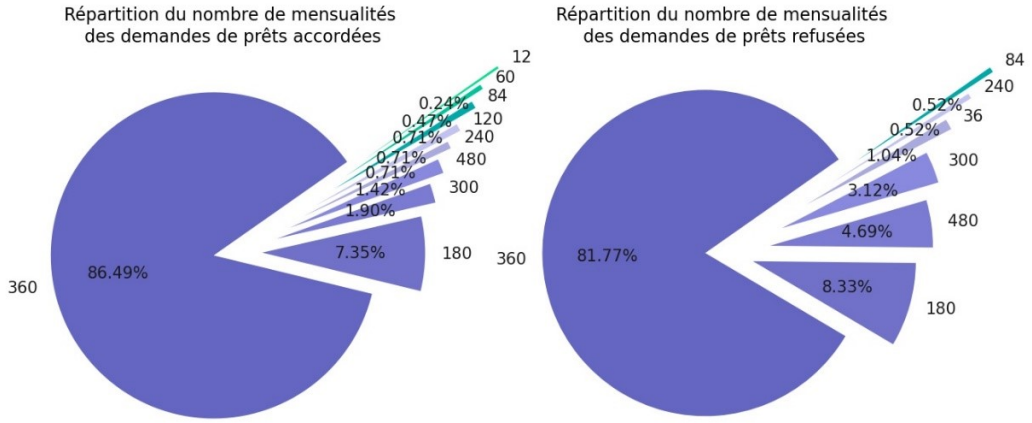
La répartition des réponses favorables aux candidats diplômés est plus importante que celle des candidats non-diplômés.



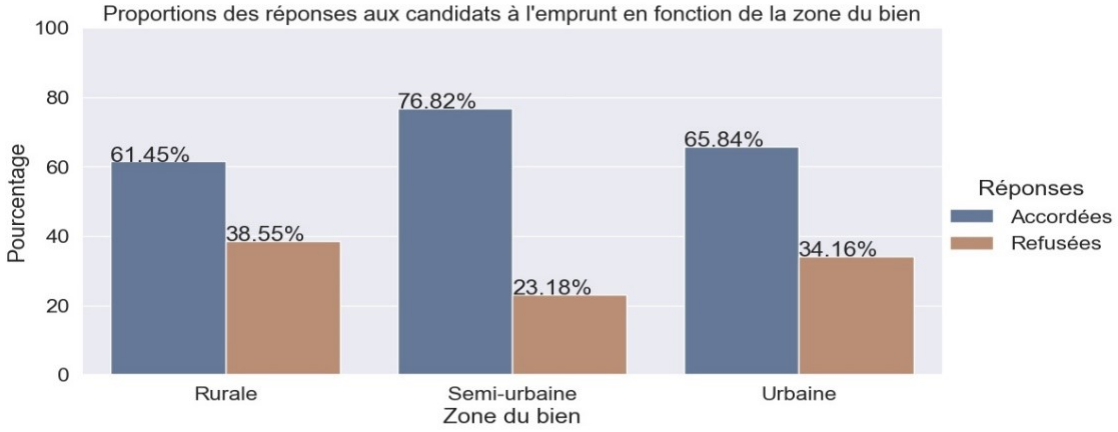
La répartition des réponses est très similaire que les candidats soient indépendants ou non.



La répartition du nombre de mensualités est similaire selon que la réponse soit favorable ou non. Cependant, le nombre de mensualités de 120, 60 ou 12 mois ont toutes été accordées contrairement aux demandes de 36 mensualités qui ont été refusées.

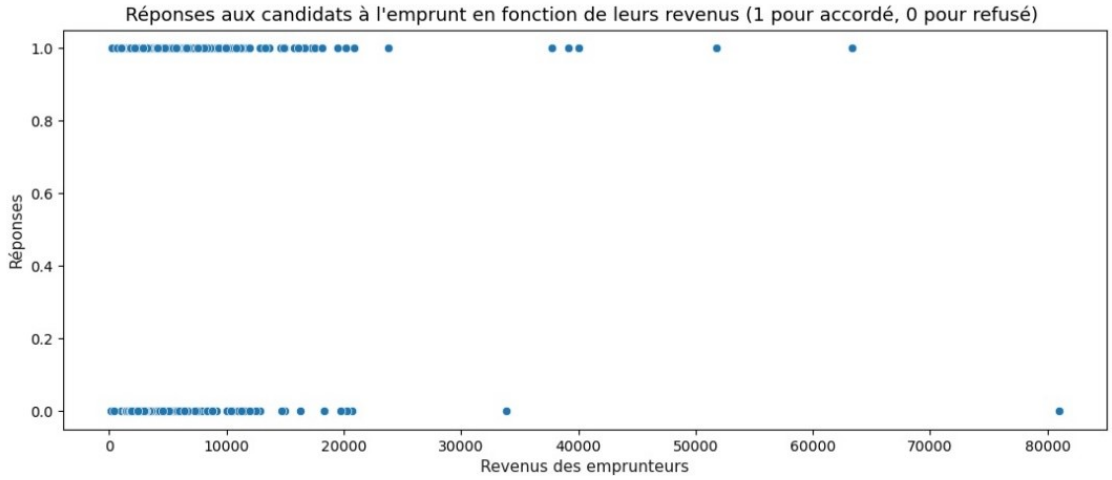


La proportion des réponses favorables est plus importante lorsque le bien du demandeur est situé en zone semi-urbaine.

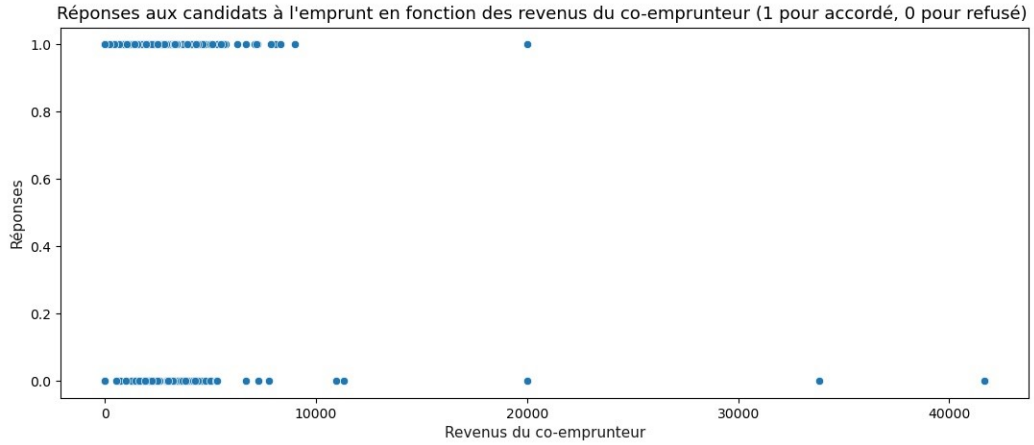


4.2.3 Les variables qualitatives et quantitatives

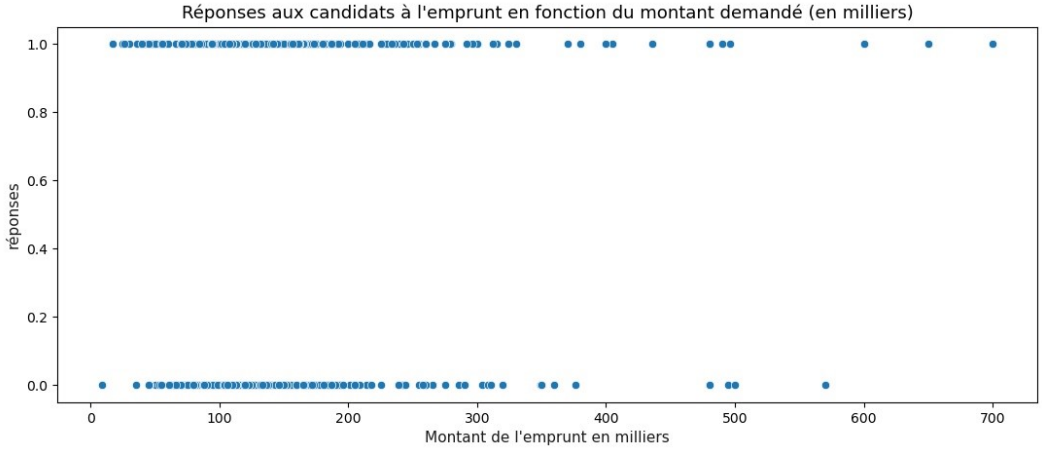
Il n'apparait pas de différences de réponses en fonction des revenus de l'emprunteur. Cependant, à partir 33848 les 5 demandes ont toutes obtenue une réponse favorable, mis à part la demande à 81000 », dont le 'Credit_History' est à 0 (défavorable).



Il n'apparait pas de différences de réponses en fonction des revenus du co-emprunteur.
Cependant, il y a 2 exceptions ; emprunteurs diplômés aux salaires assez bas, conjoints dont les salaires sont supérieurs à 30000, et qui pourtant répondent aux critères d'un « bon emprunteur », avec un 'Credit_History' positif (favorable).

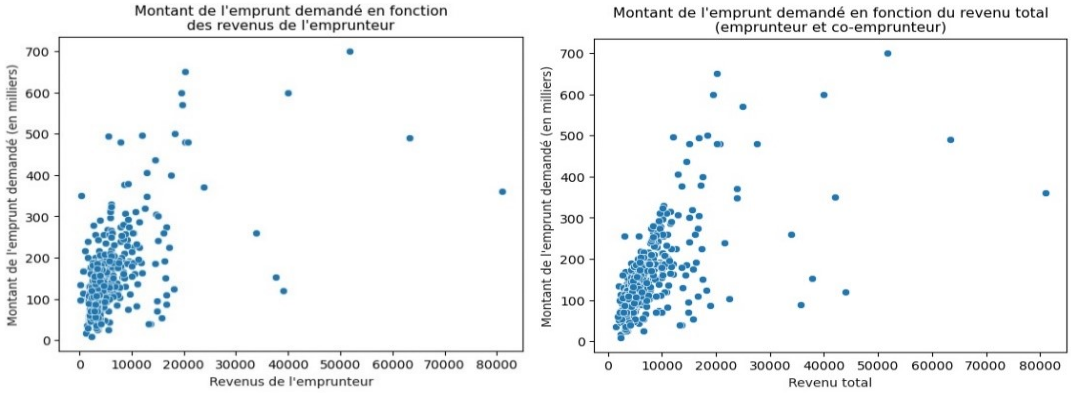


Il n'apparait pas de différences de réponses en fonction des montants demandés.
Cependant, à partir 600k les 4 demandes ont toutes obtenue une réponse favorable (3 sont visibles car 2 sont à 600k).



4.2.4 Les variables quantitatives

Plus les revenus de l'emprunteur est important, plus le montant du prêt l'est également, ce qui confirme la corrélation r² à 0.57 du tableau des corrélations. La corrélation r² augmente à 0.63 lorsque les revenus concernent l'emprunteur et le co-emprunteur (revenu total).



5. CHOIX D'UN MODÈLE DE PREDICTION

L'objet de cette étude étant de créer un modèle afin de prédire avec le moins d'erreurs possible si un crédit va être accordé ou non, nous allons au préalable préparer les données de test et d'entrainement, mettre en concurrence quatre algorithmes de classification, puis nous ferons notre choix parmi l'un des suivants :

- La régression logistique (LogisticRegression)
- La forêt d'arbres décisionnels (RandomForestClassifier)
- L'arbre de décision (DecisionTreeClassifier)
- Les K plus proches voisins (KNeighborsClassifier)

5.1 Les données d'entrainement et de test

Nous utiliserons StratifiedShuffleSplit pour diviser les données afin de conserver le même ratio de classes cible que dans les données dans leur totalité.

Nos données vont être séparées en 2, une partie 'train' représentant 80% de nos données pour l'entrainement de nos algorithmes, et une seconde partie 'test' représentant les 20% des données restantes pour déterminer le pourcentage de bonnes réponses pour chacun des algorithmes.

Chacune de ces 2 parties comprendra d'une part ;

-les variables indépendantes (X_train, X_test), desquelles on retire la variable 'Self_Employed', car nous l'avons vu précédemment, elle a une corrélation r^2 pratiquement nulle avec la decision de la banque, il nous reste 10 variables indépendantes :

'Gender', 'Married', 'Dependents', 'Education', 'ApplicantIncome', 'CoapplicantIncome', 'LoanAmount', 'Loan_Amount_Term', 'Credit_History' et 'Property_Area'.

Et de l'autre :

-la variable dépendante (y_train, y_test), notre variable binaire cible :

'Loan_Status' .

Les caractéristiques de ces 4 parties sont :

X_train : (491 lignes, 10 colonnes), variables indépendantes d'entrainement.

y_train : (491,) , soit 491 valeurs, variable cible d'entrainement.

X_test : (123, 10 colonnes), variables indépendantes de test.

y_test : (123,) , soit 123 valeurs, variable cible de test.

5.2 Mise en concurrence des algorithmes

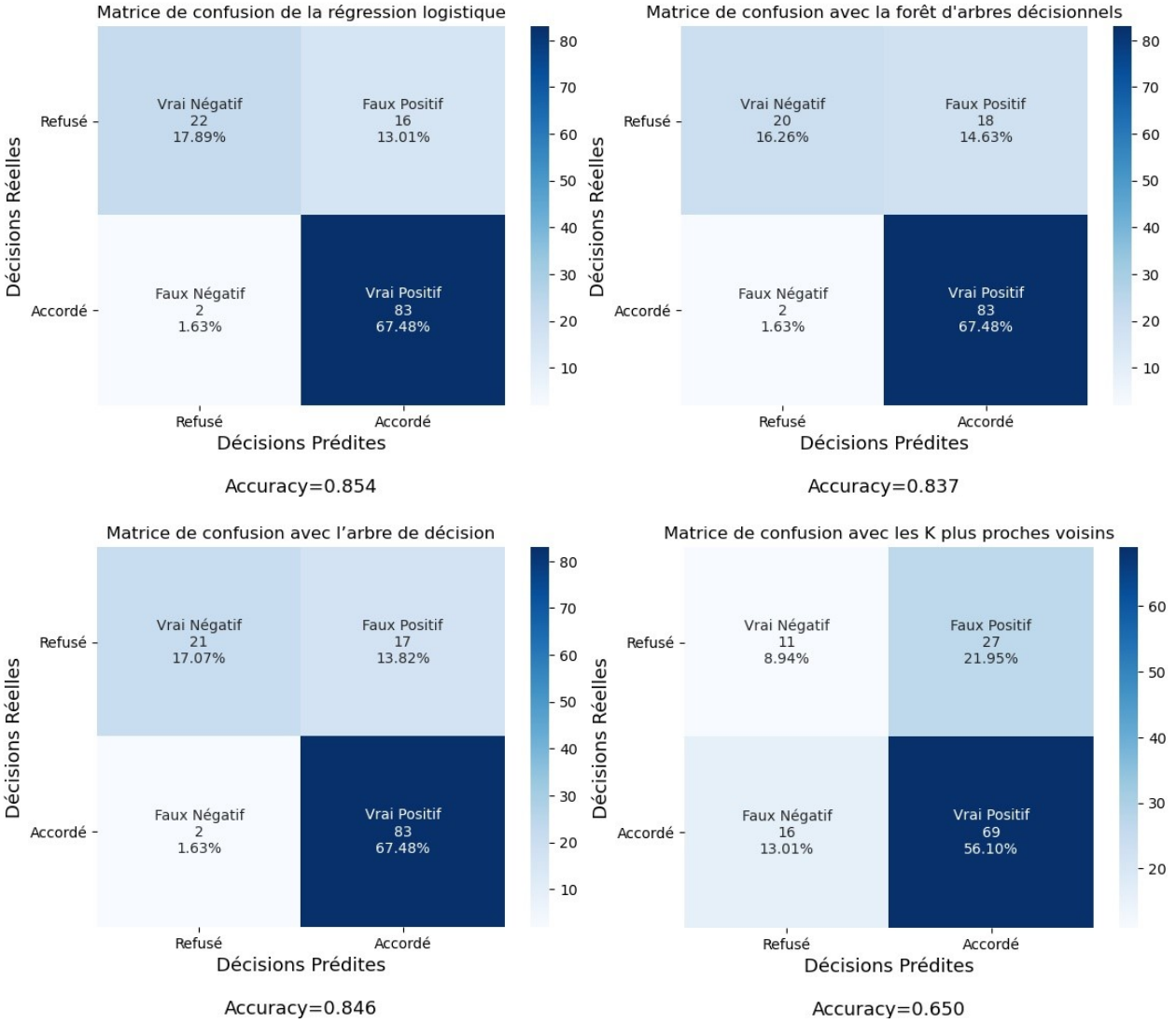
5.2.1 Précision des modèles

La précision "accuracy", proportion de réponses correctement prédites :

Pour la régression logistique, la précision du modèle est de 0.8536.
Pour la forêt d'arbres décisionnels, la précision du modèle est de 0.8373.
Pour l'arbre de décision, la précision du modèle est de 0.8455.
Pour les K plus proches voisins, la précision du modèle est de 0.6504.

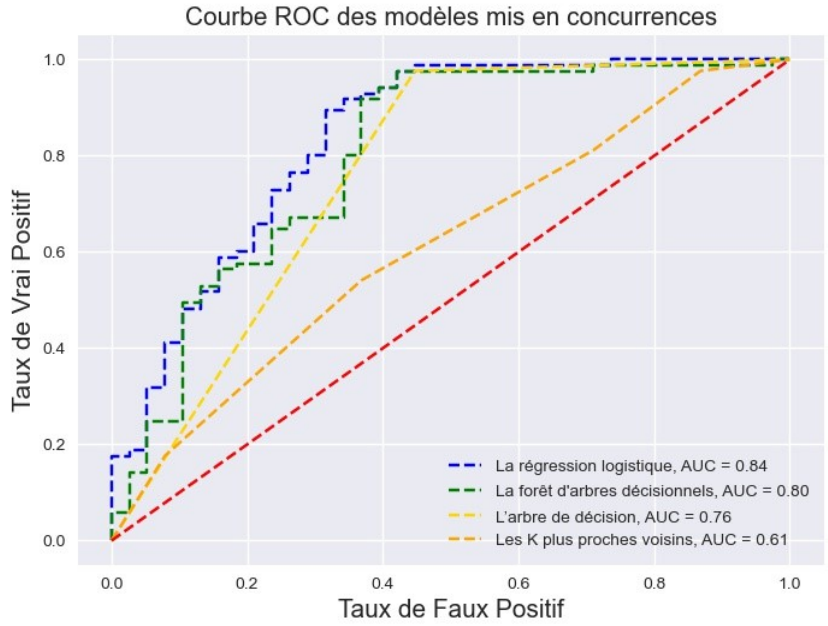
5.2.2 Les matrices de confusion

La comparaison des quatre matrices de confusion pour chacun des algorithmes indique que la classification par régression logistique a le meilleur résultat de prédiction, avec le moins d'erreurs, ce qui correspond au résultat de la précision "accuracy" avec plus de 85% de succès.



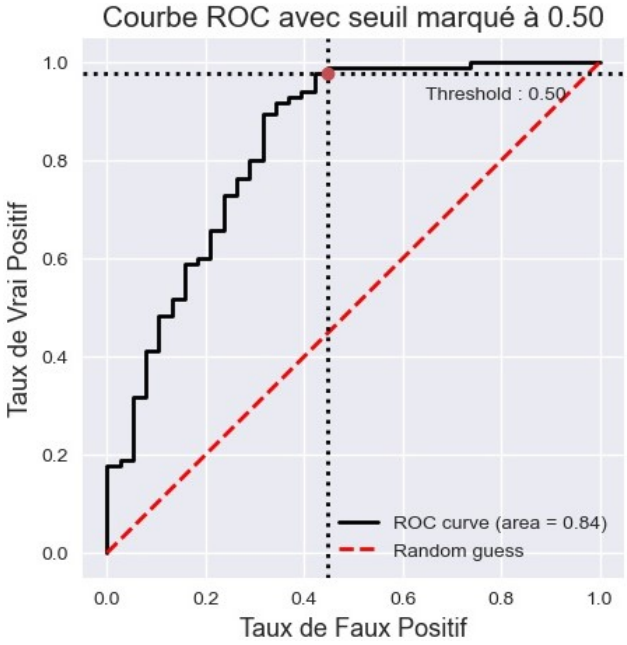
5.2.3 Les courbes ROC

Le comparatif des quatre courbes ROC confirme le choix de la régression logistique avec une aire sous sa courbe ROC de 84%, la meilleure performance.



5.3 Optimisation du modèle

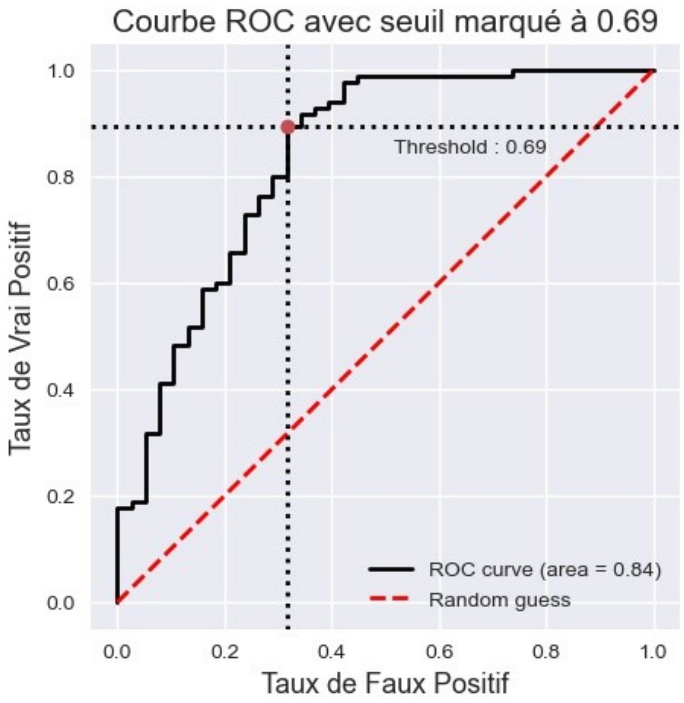
Nous cherchons à ce que le taux de faux positifs soit le plus bas possible afin de ne pas accorder un crédit à des candidats auxquels nous n'aurions pas dû, sans perdre trop de clients qui auraient dû l'obtenir.



Le point de seuil par défaut de 0.50 de la probabilité d'obtention du crédit, occasionne l'accord à la demande de crédit à 16 candidats qui n'auraient pas dû l'obtenir et le refus à 2 candidats qui auraient dû obtenir un accord (matrice de confusion page précédente).

16 candidats qui n'auraient pas dû obtenir de crédit sur 99 accords, cela parait beaucoup, car pour une banque c'est autant de risques de ne pas être remboursée.

Il est difficile de définir le point seuil optimal, cependant si nous devons choisir la valeur du seuil (Th reshold) la plus adaptée afin d'accorder une quantité moindre de crédits à des candidats auxquels il n'aurait pas dû être accordé sans perdre trop de candidats admissibles, nous devons maximiser (TV P-TFP) qui dans le graphique est représentée par la distance verticale entre la ligne pointillée rouge et la courbe noire.

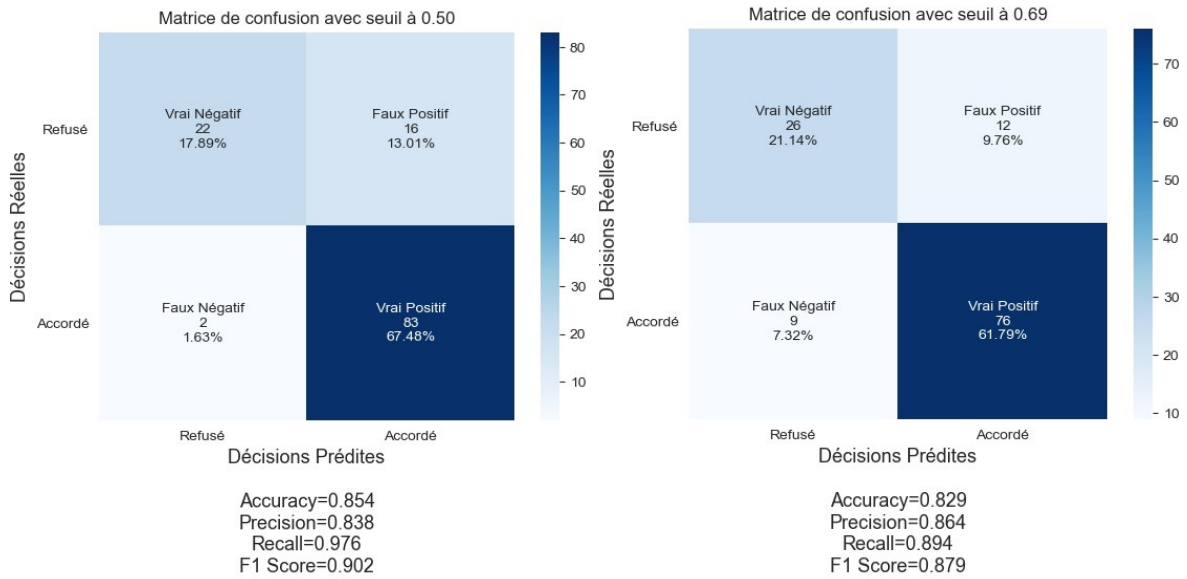


De cette manière, nous améliorons le score de précision (à ne pas confondre avec la précision "accuracy", proportion de réponses correctement prédites) :
Car notre préoccupation étant d'avoir moins de faux positifs, notre but est d'augmenter le score de précision (TP/(TP+FP)), car moins il y aura de faux positifs, plus le score de précision sera haut.

La valeur seuil optimisée répondant à ce critère est de 0.69, et elle répond à nos attentes :

Précision "accuracy" avant et après optimisation de la valeur seuil :	0.8536, 0.8292
Score de précision avant et après optimisation de la valeur seuil :	0.8383, 0.8636
Recall Score avant et après optimisation de la valeur seuil :	0.9764, 0.8941
F1 Score avant et après optimisation de la valeur seuil :	0.9021, 0.8786

5.4 Matrice de confusion avant et après optimisation du seuil



L'adoption d'un seuil optimisé à 0.69 occasionnera une perte de 7 clients (Vrai Positif 83-76) auxquels le crédit aurait dû être accordé, pour un gain de 4 clients (Faux Positif 16-12) auxquels nous ne donnerons pas le crédit par erreur comparativement au seuil par défaut qui est à 0.50.

6. CONCLUSION

Un modèle de régression logistique nous a permis de prédire correctement avec plus de 85% de réponses correctes la décision d'accorder un crédit ou non.

Malgré cette précision satisfaisante, la banque accordera le prêt par erreur à 13% des demandes qui devraient être refusées et n'accorderait pas le prêt a moins de 2% des candidats qui auraient dû l'obtenir.

Un ajustement est possible, la banque accordera le prêt par erreur à 9,76 % des demandes qui devraient être refusées et n'accorderait pas le prêt a 7,32 % des candidats qui auraient dû l'obtenir, ce qui baissera la précision «accuracy » du modèle à 83%.

La banque déterminera laquelle des deux possibilités, ou quel ajustement lui conviendra le mieux en fonction des critères qu'elle aura définis.

Des données supplémentaires telles que les garanties fournies par l'emprunteur ou son âge pourraient améliorer notre modèle de prédiction.