

Scoring des Risques de Crédit

Mathys Rosse & Romain Habib

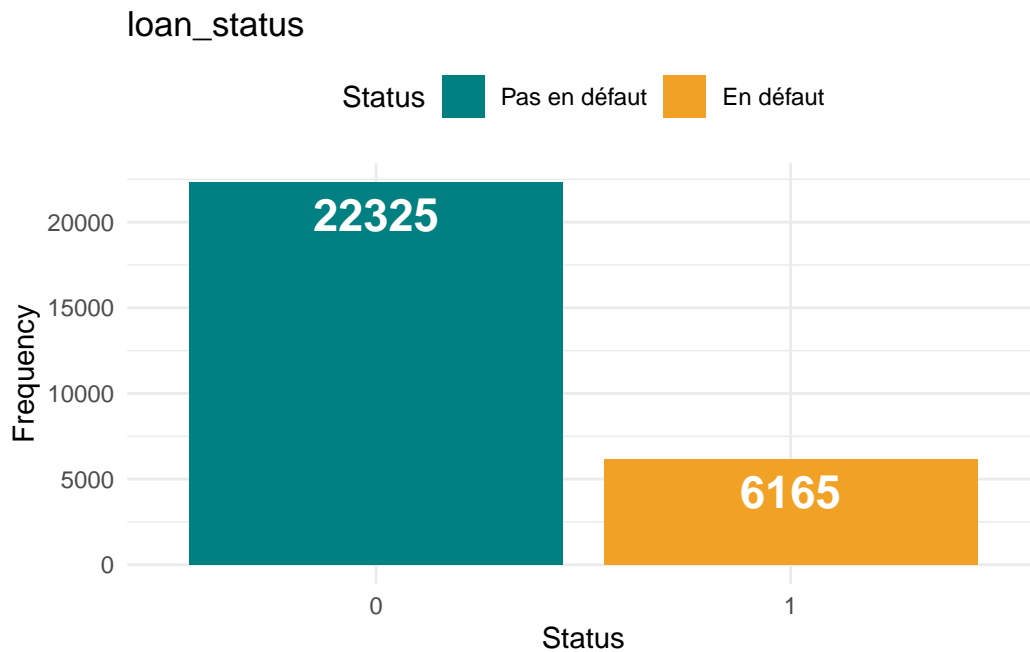
Exploration & Transformation des Données

Statistiques univariées :

Présentons les statistiques descriptives univariées en séparant les variables qualitatives et quantitatives :

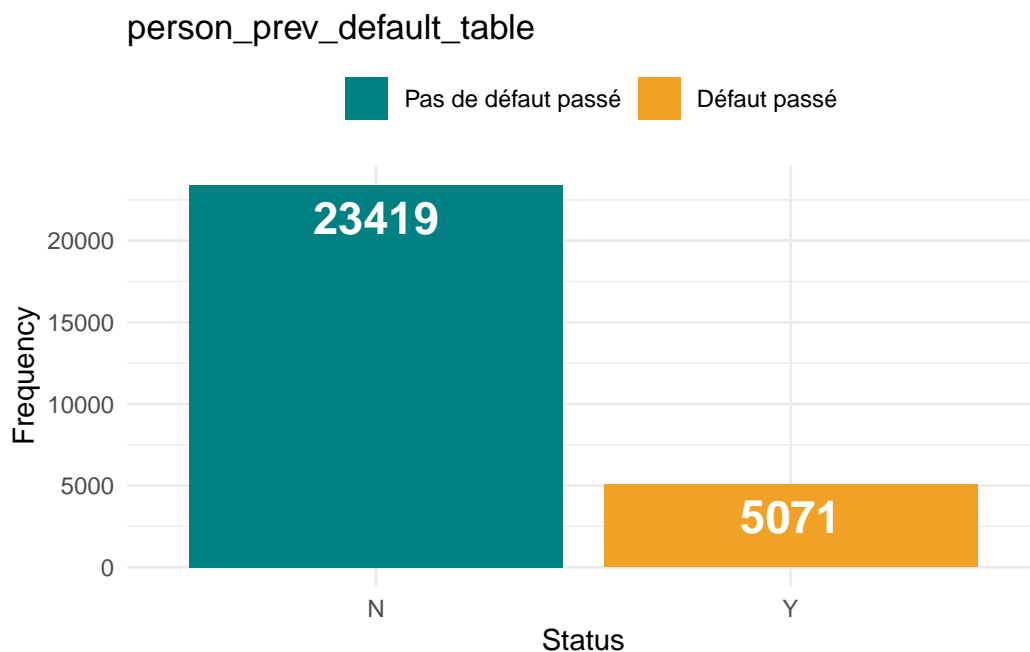
Les variables qualitatives sont : `person_prev_default`, `loan_status`, `loan_intent`, `person_home_ownership`

Présentons les :



Loan_status	Fréquence
0	0.78
1	0.22

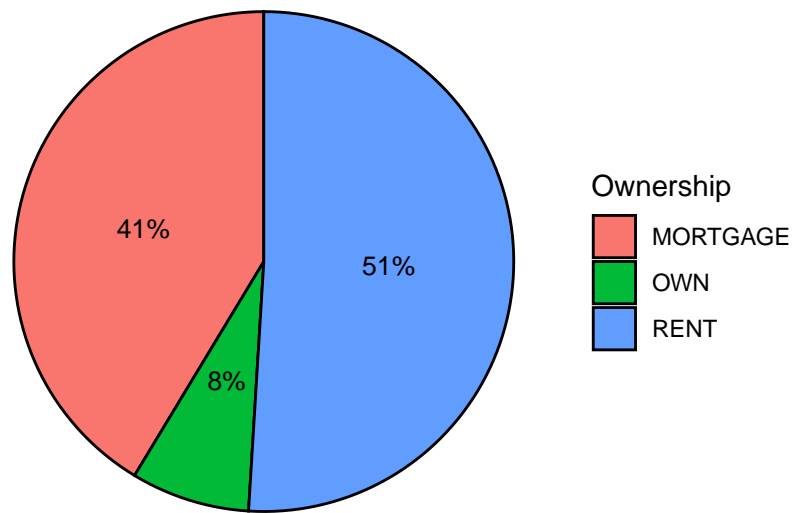
Pour commencer, étudions premièrement la proportion de clients en défaut de paiement. Grâce au graphique effectués nous voyons qu'il y a plus de clients qui ne sont pas en défaut de paiements que l'inverse. Nous en retirons une fréquence de 78.4% de clients qui ne sont pas en défaut de paiement contre 21.6% qui le sont.



person_prev_default	Fréquence
N	0.82
Y	0.18

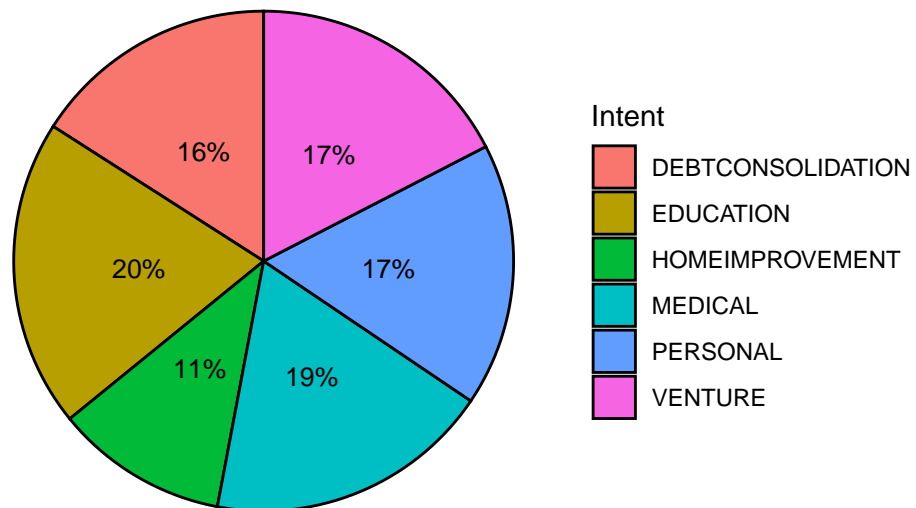
Parmi les clients, peu importe qu'ils soient en défaut de paiement ou non, 82.2% d'entre eux ont déjà été en défaut de remboursement dans le passé.

Home Ownership



Nous pouvons en retirer que la population observée est majoritairement locataire à 51%, ou propriétaire avec un emprunt à la banque à hauteur de 41.3% de notre échantillon. Très peu sont propriétaire sans emprunt soit 7.7% de la population totale.

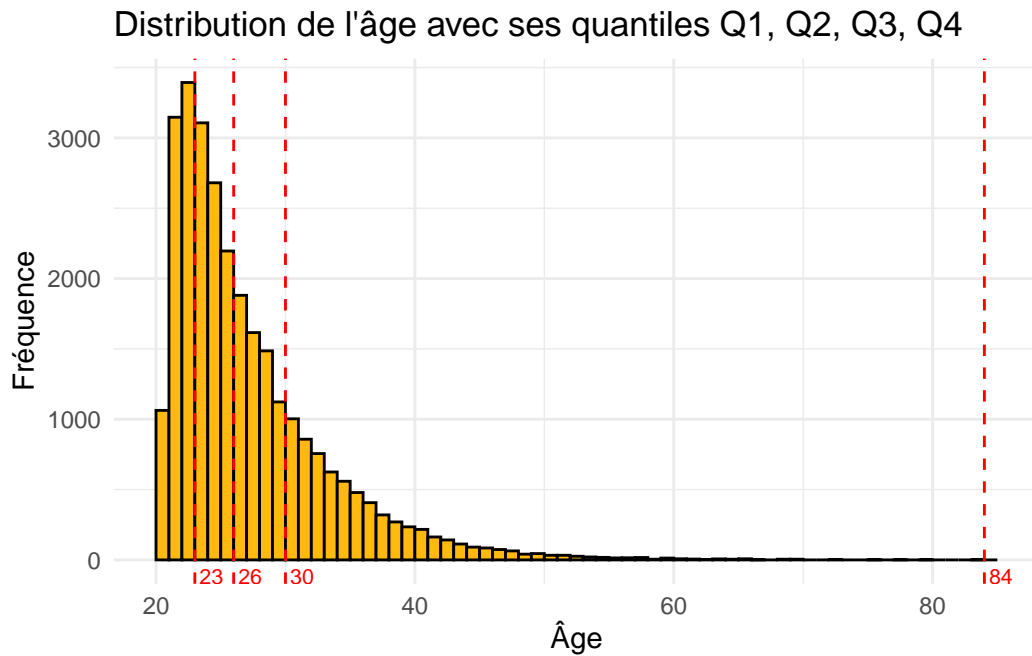
Loan Intent



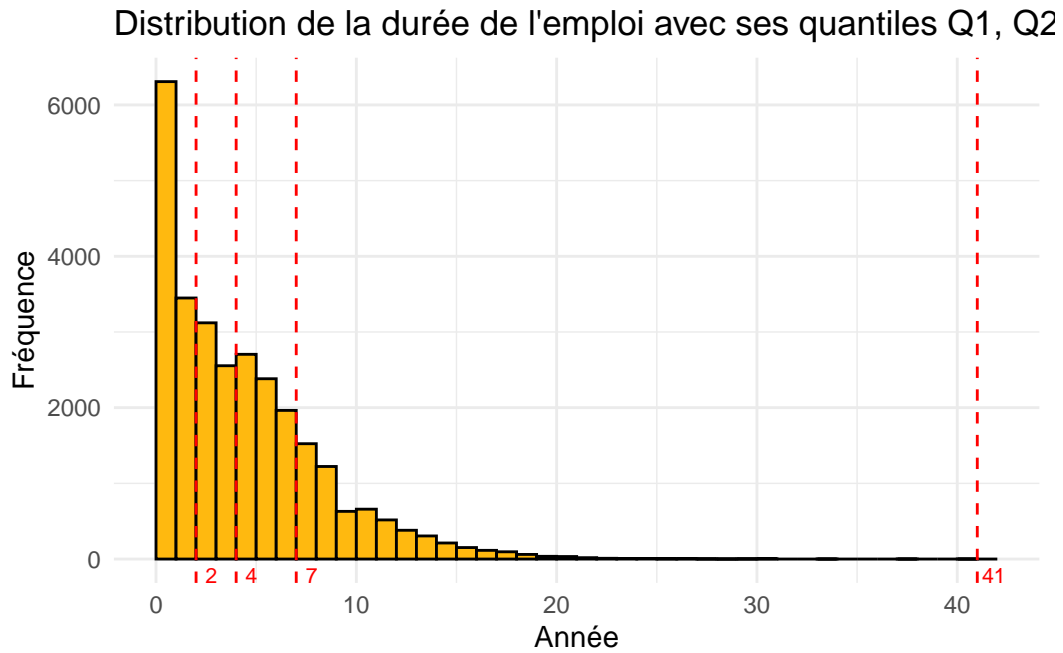
Le motif de crédit de tous ces clients est représentés par une proportion plutôt homogène soit 16% pour consolider un bien, 19.9% pour l'éducation, 11.2% pour une amélioration de leur maison, 18.5% pour le médical, 17% pour un usage personnel et 17.4% pour une entreprise (venture).

Pour ce qui est des autres variables, elles sont quantitatives.

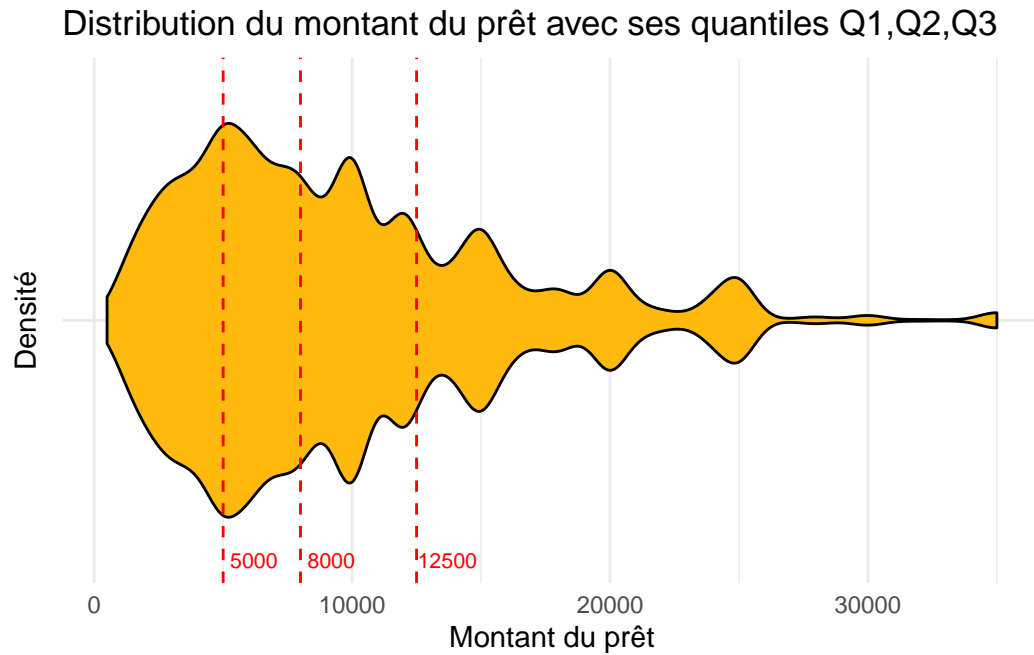
Présentons les :



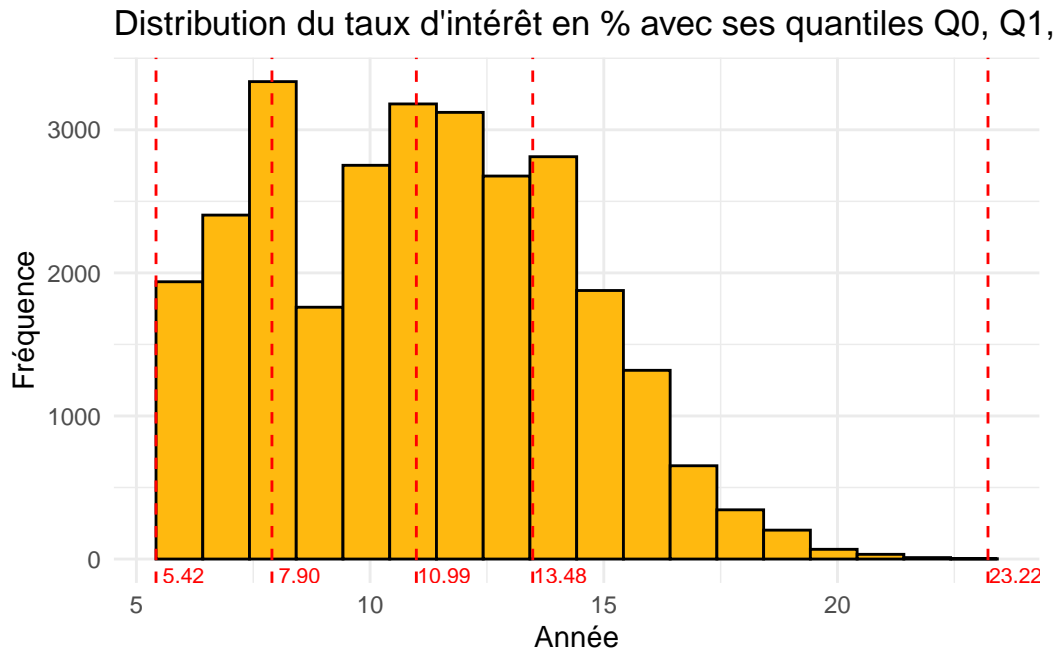
Concernant l'âge de notre population, nous avons plutôt une jeune population avec une densité plus importante d'individus ayant la vingtaine selon la distribution de l'âge qui forme une courbe qui, à partir de 23 ans, décroît en tendant vers 0. Nous trouvons donc qu'il y a 75% de notre échantillon qui se trouve en-dessous de 30 ans avec une moyenne d'âge de la population de 27 ans.



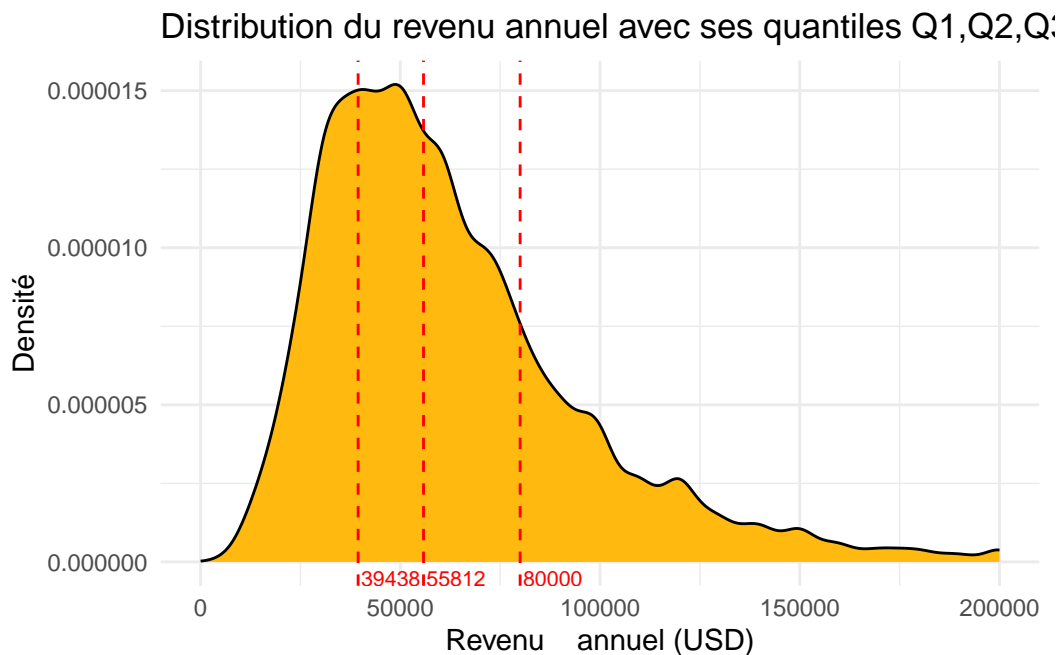
Parlons des durées des emplois des individus. Nous avons une durée dans l'emploi qui décroît au fil du temps ce qui est logique car il est moins fréquent d'avoir le même emploi durant 20 ans que 2 ans. La moyenne de la durée dans l'emploi est d'un peu plus de 4 ans et demi (4.785 ans). On a 50% de gens possédant une durée d'emploi inférieure à 4 ans et 50% des gens supérieur à 4 ans. Nous avons un individu possédant une durée d'emploi de 41 ans.



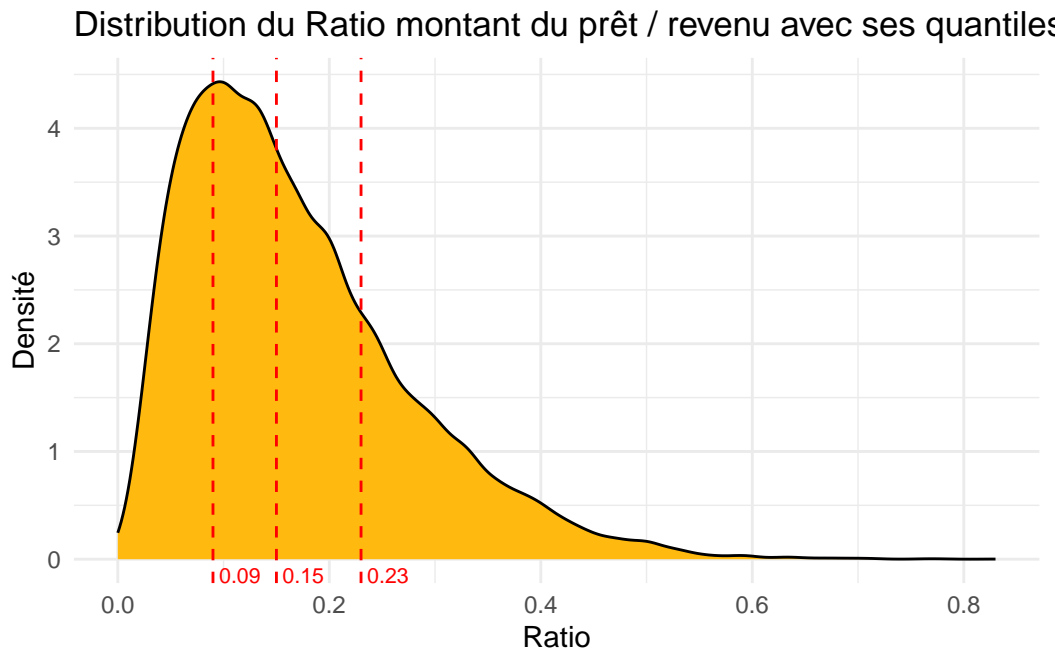
Globalement, le montant moyen du prêt est de 9,651\$ et varie de 500 jusqu'à 35,000\$. Seulement 10% des individus ont contracté un prêt supérieur à 19,400\$. Selon la distribution du montant des prêt, le prêt contracté le plus fréquent se trouve être d'une valeur autour de 5,000\$.



Le taux d'intérêt moyen est de 11.04%. On a 25% de l'échantillon qui possède un taux d'intérêt inférieur à 7.90%, de plus, le taux d'intérêt le plus bas est de 5.42%. La distribution des taux d'intérêts suggère que 50% des gens ont un taux d'intérêt compris entre 7.90% et 13.48%.



La population totale gagne en moyenne 66,404\$/an, avec un individu aberrant touchant jusqu'à 2,039,684\$/an. Cependant, 90% de la population gagne moins de 112,000\$/an.



Pour le ratio prêt/revenus, un ratio faible proche de zéro suggère un montant de prêt modeste par rapport au revenu annuel de l'emprunteur, indiquant une gestion prudente des revenus pour le remboursement, tandis qu'un ratio élevé proche de 1 révèle un niveau de dette plus important par rapport aux revenus, augmentant le risque de défaut de paiement. On voit que dans notre population, 75% des individus possèdent un ratio montant du prêt/revenu inférieur à 0.23. Le ratio moyen étant de 0.1694.

Statistiques Bivariées

Discrétisation des variables

Afin de rendre nos données adaptées à la création de notre modèle de scoring, nous allons tout d'abord segmenter nos variable continues en 4 modalités suivant les valeurs de leurs quartiles. Ensuite, nous procéderons à la création de dummy de chacune de ces modalités afin de pouvoir déceler les similarités entre modalités lors de l'estimation de notre modèle de régression logistique. Cette transformation en dummy concernera également les variables originellement catégorielles (loan_intent & person_home_ownership)

Estimation du Modèle de Regression Logistique

L'estimation de notre modèle se fera en plusieurs étapes. Premièrement, nous effectuerons une estimation "naïve" avec toutes les modalités de toutes les variables (hormis les modalités de référence de chaque variables).

Call:

```
glm(formula = loan_status ~ ., family = binomial(link = "logit"),
     data = credit_binary)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-6.473971	0.143914	-44.985	< 2e-16	***
person_age_gr_1	0.206826	0.050621	4.086	4.39e-05	***
person_age_gr_2	0.069974	0.048351	1.447	0.14784	
person_age_gr_3	0.007443	0.051925	0.143	0.88603	
person_home_ownership_MORTGAGE	1.351074	0.100828	13.400	< 2e-16	***
person_home_ownership_RENT	2.226293	0.097374	22.863	< 2e-16	***
loan_intent_DEBTCONSOLIDATION	0.956490	0.061407	15.576	< 2e-16	***
loan_intent_EDUCATION	0.180622	0.062222	2.903	0.00370	**
loan_intent_HOMEIMPROVEMENT	1.016128	0.068939	14.740	< 2e-16	***
loan_intent_MEDICAL	0.831164	0.059865	13.884	< 2e-16	***
loan_intent_PERSONAL	0.408175	0.063440	6.434	1.24e-10	***
person_prev_default	0.121791	0.045767	2.661	0.00779	**
loan_amnt_gr_1	1.084240	0.059887	18.105	< 2e-16	***
loan_amnt_gr_2	0.299183	0.055325	5.408	6.38e-08	***
loan_amnt_gr_4	0.056630	0.049820	1.137	0.25567	
loan_percent_income_gr_2	0.445577	0.058382	7.632	2.31e-14	***
loan_percent_income_gr_3	1.025986	0.062897	16.312	< 2e-16	***
loan_percent_income_gr_4	2.782044	0.067241	41.375	< 2e-16	***
loan_int_rate_gr_2	0.341227	0.058087	5.874	4.24e-09	***
loan_int_rate_gr_3	0.619471	0.056783	10.909	< 2e-16	***
loan_int_rate_gr_4	2.091285	0.057892	36.124	< 2e-16	***
person_emp_length_gr_1	0.372757	0.047533	7.842	4.43e-15	***
person_emp_length_gr_3	0.053606	0.053403	1.004	0.31547	
person_emp_length_gr_4	0.162805	0.058169	2.799	0.00513	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 29761 on 28489 degrees of freedom
 Residual deviance: 21593 on 28466 degrees of freedom
 AIC: 21641

Number of Fisher Scoring iterations: 5

Dans notre modèle, nous avons sélectionné nos modalités de référence pour chaque variable de sorte à ce que tous nos coefficients β soit positifs (la modalité de référence ayant un coefficient de 0, nous devons choisir la variable ayant le coefficient le plus faible).

Nous pouvons cependant observer que certaines de nos modalités ne présentent pas de coefficients différents de 0. 0 étant le coefficient de la modalité de référence, les modalités concernés ne sont en réalité pas significativement différente de la modalité de référence, nous procéderons donc à la fusion de ces modalités.

En pratique, nous fusionnerons les modalités de groupes d'âge 2, 3 et 4. Ne restera donc dans notre modèle que la variable `person_age_1` (age < 23) présentant un coefficient positif. Nous fusionnerons également la modalité `loan_amnt_4` avec la modalité de référence `loan_amnt_3` ainsi que la modalité `person_emp_length_gr_3` avec `person_emp_length_gr_2`.

Call:

```
glm(formula = loan_status ~ ., family = binomial(link = "logit"),
    data = credit_binary_2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-6.39880	0.13692	-46.735	< 2e-16 ***
person_age_gr_1	0.18132	0.04079	4.445	8.80e-06 ***
person_home_ownership_MORTGAGE	1.36047	0.10071	13.509	< 2e-16 ***
person_home_ownership_RENT	2.22749	0.09738	22.874	< 2e-16 ***
loan_intent_DEBTCONSOLIDATION	0.95618	0.06140	15.573	< 2e-16 ***
loan_intent_EDUCATION	0.17963	0.06220	2.888	0.00388 **
loan_intent_HOMEIMPROVEMENT	1.01645	0.06890	14.754	< 2e-16 ***
loan_intent_MEDICAL	0.82980	0.05985	13.865	< 2e-16 ***
loan_intent_PERSONAL	0.40613	0.06342	6.404	1.51e-10 ***
person_prev_default	0.12137	0.04575	2.653	0.00797 **
loan_amnt_gr_1	1.05934	0.05608	18.889	< 2e-16 ***
loan_amnt_gr_2	0.27285	0.05000	5.457	4.83e-08 ***
loan_percent_income_gr_2	0.44530	0.05835	7.631	2.33e-14 ***
loan_percent_income_gr_3	1.02911	0.06276	16.398	< 2e-16 ***
loan_percent_income_gr_4	2.79339	0.06629	42.142	< 2e-16 ***
loan_int_rate_gr_2	0.34336	0.05803	5.917	3.28e-09 ***

loan_int_rate_gr_3	0.62246	0.05665	10.988	< 2e-16	***
loan_int_rate_gr_4	2.09491	0.05768	36.318	< 2e-16	***
person_emp_length_gr_1	0.34377	0.03842	8.948	< 2e-16	***
person_emp_length_gr_4	0.13836	0.05107	2.709	0.00674	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 29761 on 28489 degrees of freedom
 Residual deviance: 21598 on 28470 degrees of freedom
 AIC: 21638

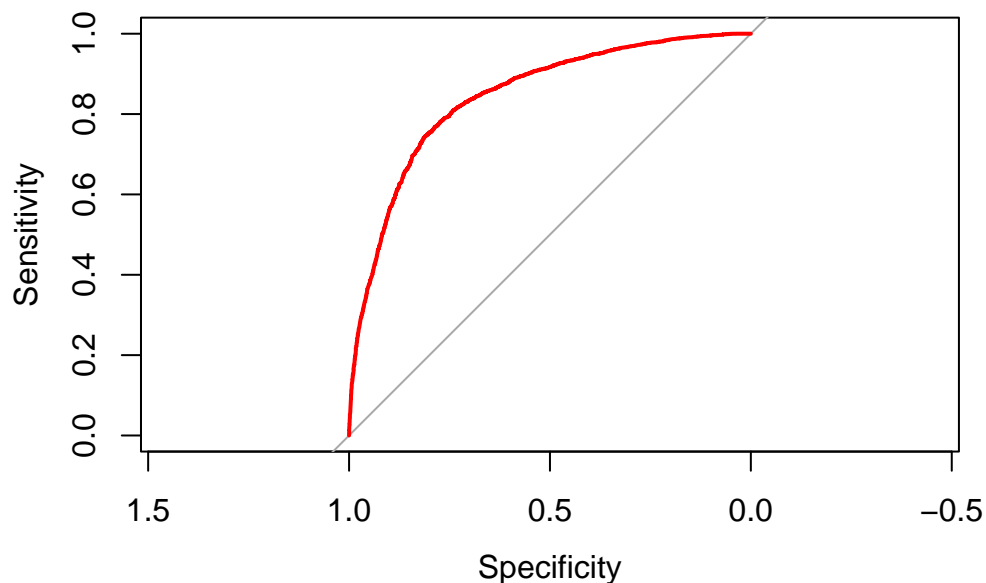
Number of Fisher Scoring iterations: 5

Maintenant que nous avons procédé à notre première vague de fusion de variables, toutes nos modalités restantes sont significativement différentes de 0 (et donc de leur modalité de référence) à un seuil de 1%.

Pour finir, nous procéderons à une série de tests de Wald afin de déterminer si nos modalités restantes présentent des impacts similaires sur la probabilité de défaut des individus. L'hypothèse H_0 de chacun de ces tests est $\beta_{m1} = \beta_{m2}$ avec $m1$ et $m2$ les deux modalités testées.

Nous pouvons conclure à la suite de ces tests que nos modalités ont toutes des impacts mesurés différents sur la probabilité de défaut des individus.

Nous procéderons maintenant à l'estimation du seuil de probabilité prédite nous permettant d'obtenir le modèle possédant le compromis sensibilité/spécificité le plus équilibré. Nous utiliserons la courbe ROC de notre modèle pour guider notre décision.



Area under the curve: 0.8441

[1] "Optimal Cutoff Point: 0.256"

Notre seuil sera donc 0.256. Si un individus présente une probabilité de défaut supérieur à ce seuil, il sera considéré comme à risque de défaut. Nous procéderons maintenant à la création de notre grille de score se basant sur notre modèle.

Scoring et Interprétation

Afin de pouvoir établir notre grille de score, nous allons procéder à la mise à l'échelle de nos coefficients. L'impact d'une modalité sur le score final d'un individu sera défini par la formule : $100 \frac{\beta_m}{\text{somme des coefs max}}$ avec la somme des coefficients maximum la somme des coefficients correspondant aux modalités définissant l'individu théorique le plus à risque.

Dans notre cas, l'individu théorique le plus à risque est un client de moins de 23 ans, locataire, empruntant pour des rénovations immobilières un montant faible d'argent (< 5000\$) avec un taux d'intérêt élevé (> 13.48%), le montant de l'emprunt représente plus 23% de son revenu, sa durée d'emploi est faible (< 2 ans). il a également fait défaut d'un emprunt par le passé. La somme des coefficients associés à ces caractéristiques est 9.83804. Voici donc les nouveaux coefficients associés à chaque modalité.

person_age_gr_1	person_home_ownership_MORTGAGE
1.843003	13.828671
person_home_ownership_RENT	loan_intent_DEBTCONSOLIDATION
22.641626	9.719216
loan_intent_EDUCATION	loan_intent_HOMEIMPROVEMENT
1.825844	10.331854
loan_intent_MEDICAL	loan_intent_PERSONAL
8.434607	4.128208
person_prev_default	loan_amnt_gr_1
1.233693	10.767786
loan_amnt_gr_2	loan_percent_income_gr_2
2.773384	4.526327
loan_percent_income_gr_3	loan_percent_income_gr_4
10.460507	28.393771
loan_int_rate_gr_2	loan_int_rate_gr_3
3.490138	6.327053
loan_int_rate_gr_4	person_emp_length_gr_1
21.293932	3.494322
person_emp_length_gr_4	
1.406341	

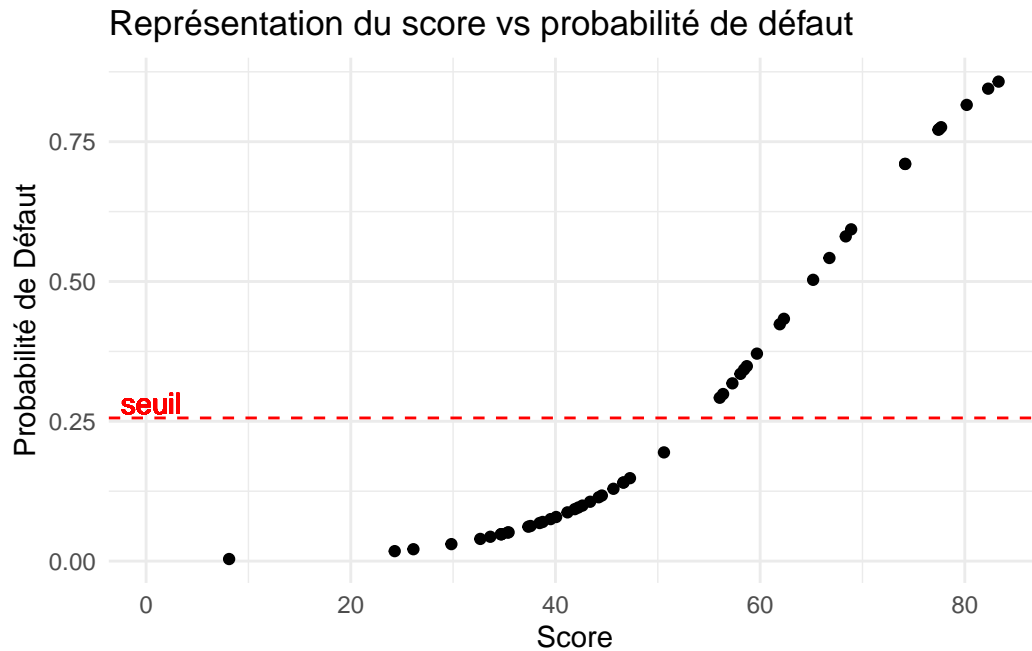
Nous observons bien que suivant ces nouveaux coefficients, un individu correspondant à la description ci-dessus obtiendra bien un score de 100%.

Suivant l'interprétation des coefficients, nous pouvons observer que les facteurs les plus à même de prédire le défaut d'un client sont le ratio emprunt/revenu, si l'emprunteur est locataire et le montant du taux d'intérêt.

Passons maintenant à l'étape de prédiction sur données de test, nous allons devoir dans un premier temps effectuer sur les données de test les mêmes transformations que nous avons effectué sur les données d'entraînement.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8.112	37.789	46.149	50.458	62.224	83.300

[1] 0.42



[1] 0.308424

L'intégralité de nos individus de tests ont obtenu des scores entre 8.112 et 83.3. Nous pouvons également observer grâce à notre graphique la relation entre le score et la probabilité de défaut de l'individu, il semblerait qu'un score supérieur à 55 est synonyme avec un risque significatif de défaut.

Il semblerait que 42% de nos individus de test ont été déclaré comme étant des individus à haut risque de défaut. Ce nombre est