Projet de Datavisualisation

Sujet : Prédiction de la probabilité de défaut

Réalisé par :
HAMMAMI Wejdene
MENNESSIER Lyzie

Supervisé par :
Mr. KEZHAN Shi

Niveau : M2 Actuariat
Année Universitaire : 2024-2025

Table des matières

1 Introduction Générale

La probabilité de faire défaut est défini dans la règlementation Bale II comme "la probabilité qu'un débiteur ne puisse faire face à ses obligations de remboursement." Le risque de défaut représente donc un enjeux majeur pour de nombreux acteurs comme par exemple les banques lors de l'octroi d'un crédit.

Nous pouvons donc nous demander quels caractéristiques influencent la probabilité ou non de faire défaut.

Pour cela, nous avons à travers ce projet effectué une analyse de notre base de données, une modélisation et un travail de visualisation graphique afin d'identifier les facteurs influençant le risque de faire ou non défaut.

2 Méthodologie

2.1 Préparation des données

Notre base de données recense des milliers de prêts accordés via la plateforme Lending Club qui est une plateforme de prêts entre particuliers. Les personnes qualifiées de fiables se veront donc proposer des prêts plus facilement et à des taux d'intérêt plus avantageux contrairement aux personnes jugées moins fiables.

Notre base contient 887 380 lignes. Pour un souci de simplicité au niveau de l'affichage dans Rshiny nous ne conservons que 30% de la base de données initiales.

La première étape de ce projet consiste en la préparation des données. Nous avons tout d'abord commencé par remplacer les valeurs manquantes présentes. Pour cela nous avons choisi pour les colonnes numériques de remplacer par la moyenne et pour les colonnes catégorielles de remplacer par le mode.

Nous effectuons ensuite quelques transformation sur les données. Nous souhaitons créer des catégories de revenus, pour cela nous calculons les quantiles de la variable *annual_ inc*, ce qui nous permet de créer une nouvelle variable *income_ category* prenant trois mode : *Low*, *Medium* ou *High*. Nous souhaitons aussi catégoriser les conditions de prêt. La variable *loan_ condition* distingue les bons prêts (*Good Loan*) des mauvais prêts (*Bad Loan*).

Nous préparons ensuite les données pour qu'elles soient utilisable dans le cadre d'une visualisation graphique. Nous souhaitons faire une carte, pour cela nous créons une table de comptage des prêts par Etats que nous fusionnons avec les données géographiques de la cartographie. Un travail d'agrégation est nécessaire pour identifier la moyenne des taux d'intérêt par catégorie de revenus, le nombre de bons ou de mauvais prêts émis ainsi que la répartition des prêts par finalité.

Viens ensuite une étape de contôles des données. Nous supprimons les valeurs aberrantes qui se trouvent

en dehors des bornes définies par l'IQR (écart interquartile) pour les trois variables suivantes : *int_rate*, *dti* et *annuel_inc*.

Nous finissons par vérifier l'absence de valeurs manquantes grâce à un diagramme.

2.2 Interface Rshiny

Notre interface Rshiny contient trois sections permettant de présenter nos données, une autre d'analyser les données et enfin la dernière permettant la prédiction de la probabilité de défaut.

La section de présentation des données nous permet de visualiser et filtrer sur les première lignes de notre base.

La seconde section d'analyse de données nous donnes tous les graphiques réalisés. Ces résultats seront détaillés dans la prochaine section.

La dernière section qui concerne la prédiction nous permet d'estimer la probabilité de défaut de l'emprunteur en paramétrant ses caractéristiques. Nous pouvons faire varier le grade, l'ancienneté professionnelle, l'objectif du prêts, la durée du prêts, le revenu annuel de l'emprunteur et le taux d'intérêt.

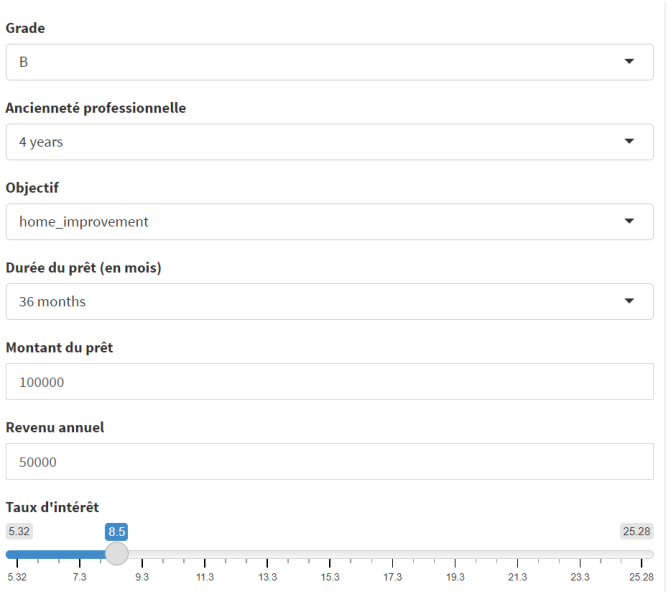

FIGURE 1 – Variables de prédiction pour la simulation de la probabilité de défaut

2.3 Visualisation des données

Dans cette partie, nous allons interpréter tous les graphiques générés dans l'application Rshiny.

Commençons par observer la répartition de défaut de notre base.

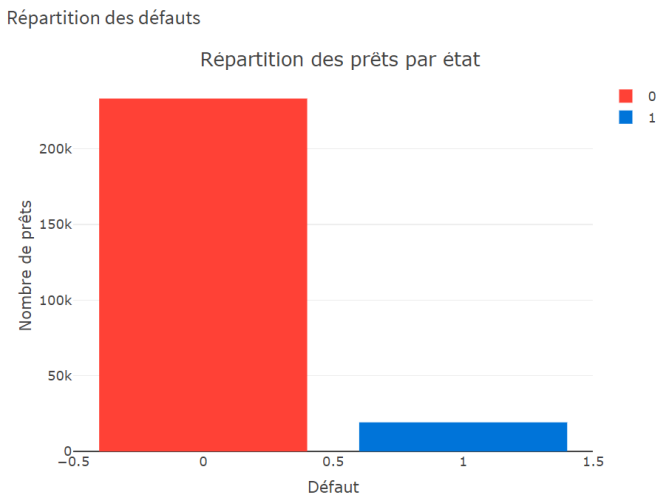

FIGURE 2 – Répartition des montants de prêt en défaut ou non

On constate que la majorité des prêts ne sont pas en défaut dans notre base. Environ 250 000 prêts ne sont pas en défaut et 20 000 sont en défaut. On remarque que la base est déséquilibréé, il sera important de traiter cela avant la modélisation.

Intéressons nous maintenant à l'influence géographique. La carte ci-dessus donne la répartition géographique du nombre de prêts.

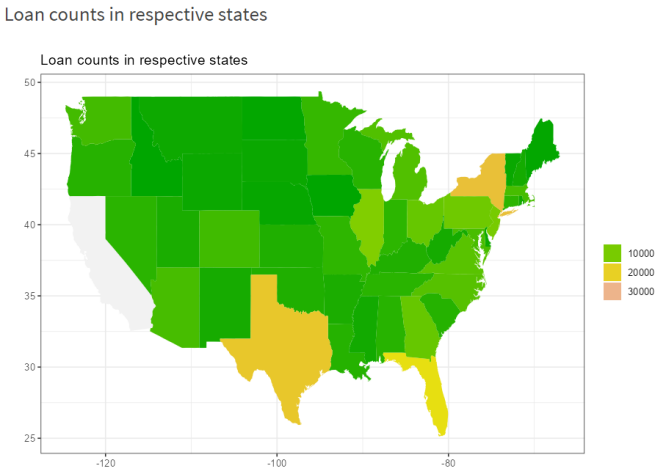

FIGURE 3 – Répartition géographique des prêts

On constate un nombre de prêts homogène aux alentours de 15 000 au niveau national même si quelques états se distinguent par un nombre plus élevé aux alentours des 20 000 notamment au sud et au Nord Est des Etats-Unis. Cela peut simplement s'expliquer par une population plus importante dans ses Etats.

Regardons maintenant la courbe de survie des défauts par grade en fonction de la durée.

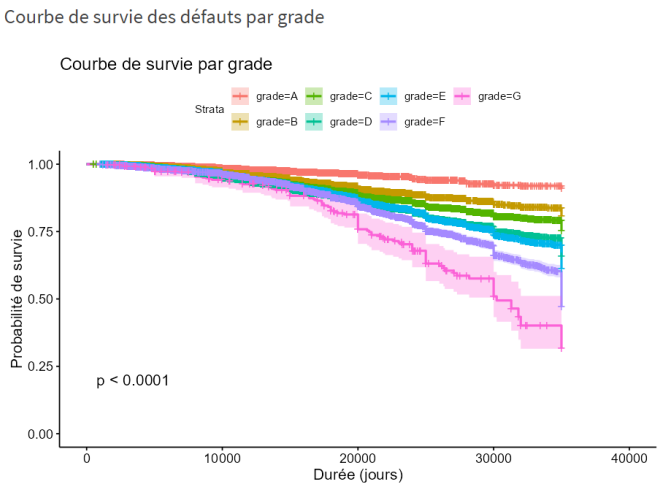

FIGURE 4 – Probabilité de survie des défauts par grade

On en déduit que les grades A sont ceux qui ont la plus haute probabilité de survie de défaut, ils sont ensuite classé par ordre alphabétique et donc les grades G sont ceux qui ont la plus basse probabilité de survie de défaut. On remarque aussi que cette probabilité diminue avec la durée d'autant plus que l'on avance vers le grade G. Le grade est donc un excellent indicateur de risque de défaut.

Regardons ensuite le taux d'intérêt en fonction du grade.

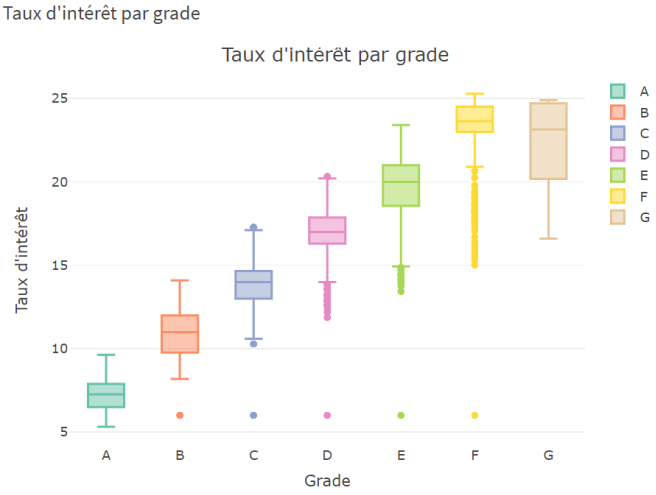

FIGURE 5 – Taux d'intérêt par grade

De manière générale, on constate que le taux d'intérêt median augmente avec le grade. On observe une dispersion très importante pour le grade G et la présence de valeurs extrêmes pour tous les grades sauf A et G.

Regardons l'observation précédente un peu plus en détail.

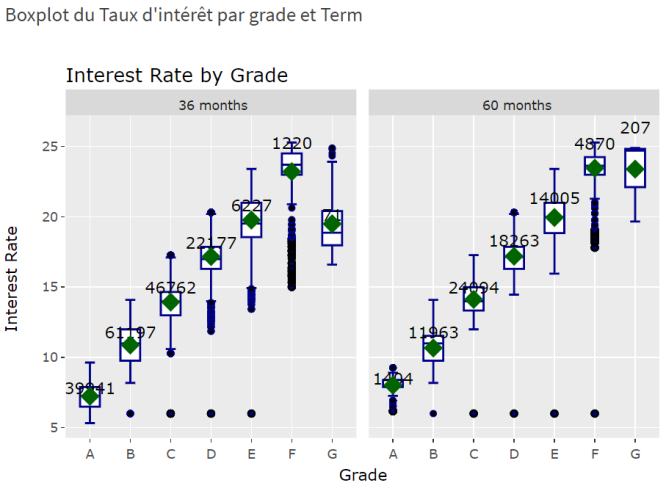

FIGURE 6 – Boxplot du taux d'intérêt en fonction du grade par durée d'emprunt

Les deux bloxplot représentent le montant du taux d'intérêt en fonction du grade pour un emprunt d'une durée de 36 mois et de 60 mois. On constate globalement que le taux d'intérêt augmente avec le grade et la durée de l'emprunt sauf pour la classe G. Cela est logique car les grades les plus élévés sont les plus risqués. Ce risque est alors compensé par un taux d'intérêt plus élevé. Le taux d'intérêt est aussi plus élevé pour une durée de prêt plus longue toute chose égale par ailleurs. Encore une fois, cela est logique car l'augmentation de la durée du prêt augmente le risque de défaut. La taille des boîtes indique la dispersion, la dispersion est plus importante lorsque le grade augmente. Cela montre une variabilité des taux plus importante pour ces grades. On peux aussi observer que la distribution du nombre de prêt par grade est inégale, les grades B, C et D sont ceux ayants le plus de prêts. Ici encore, on remarque la présence que quelques valeurs extrêmes. Nous avons réduit le taux de valeurs aberrantes par variable grâce au critère d'IQR mais nous ne pouvons pas supprimer toutes les valeurs extrêmes sans perdre des informations importantes.

On s'intéresse maintenant au lien entre le statut du prêt et le taux d'intérêt.

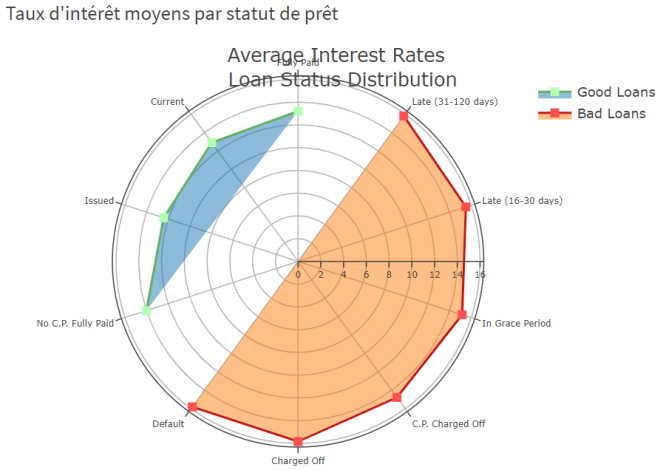

FIGURE 7 – Taux d'intérêt en fonction du statut du prêt

Ce diagramme radar montre que les mauvais prêts sont ceux qui présentent les taux d'intérêts les plus élevés et que les bons prêts ont un taux d'intérêt plus avantageux.

Intéressons nous aussi au montant des prêts par grade.

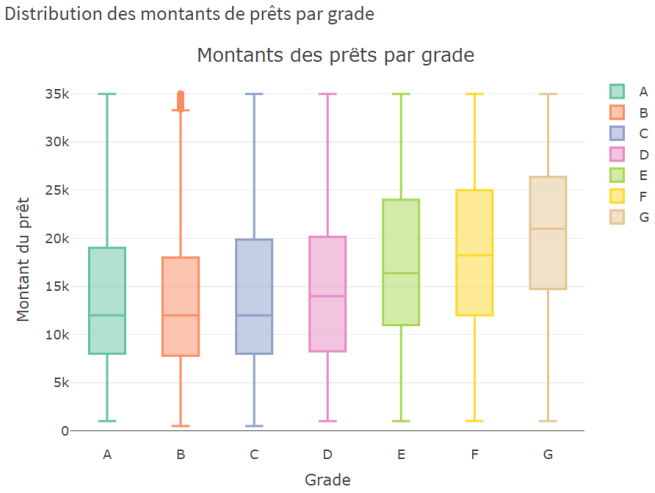

FIGURE 8 – Montant des prêts par grade

Ce boxplot met en avant que les grades A à D sont ceux qui empruntent le moins avec des montants médians entre 12 000 $ et 14 000 $. Les grades E à G empruntent quant à eux plus avec des montants médians allant de 16 000 $ à 21 000 $. La dispersion des montants de prêts est globalement la même pour la plupart des grades. Les emprunteurs risqués empruntent globalement des sommes plus importantes.

Nous avons ensuite jugé intéressant de regarder le montant des prêts par grade et par statut de propriété.

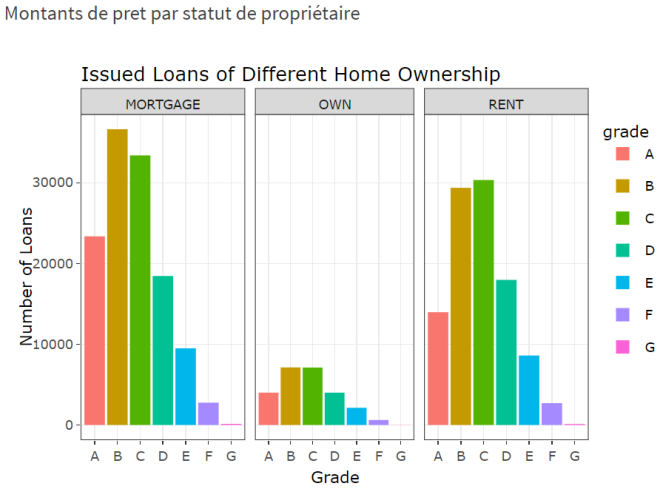

FIGURE 9 – Montant des prêts par grade et statut de propriété

Les histogrammes montrent que chez les propriétaires avec hypothèque, les grades B et C sont les plus fréquents, suivis par le grade A. Les grades D, E, F et G sont beaucoup moins représentés. Ainsi les

personnes ayant une hypothèque ont un profil de risque moyen à faible. Pour les propriétaires sans hypothèque, la distribution est plus concentrée sur les grades B et C, avec moins de prêts pour les autres grades. Le nombre total de prêts pour ce groupe est également beaucoup plus faible que pour les propriétaires avec hypothèque ou les locataires. Pour les locataires, on observe une distribution plus étalée, avec des grades B et C majoritaires, mais aussi une représentation non négligeable des grades D et E. Les grades F et G sont toujours moins fréquents, mais plus présents que chez les propriétaires avec hypothèque. Cela suggère que les locataires ont un profil de risque plus varié, avec une proportion plus importante d'emprunteurs risqués.

Regardons le lien entre le revenu annuel et le montant du prêt.

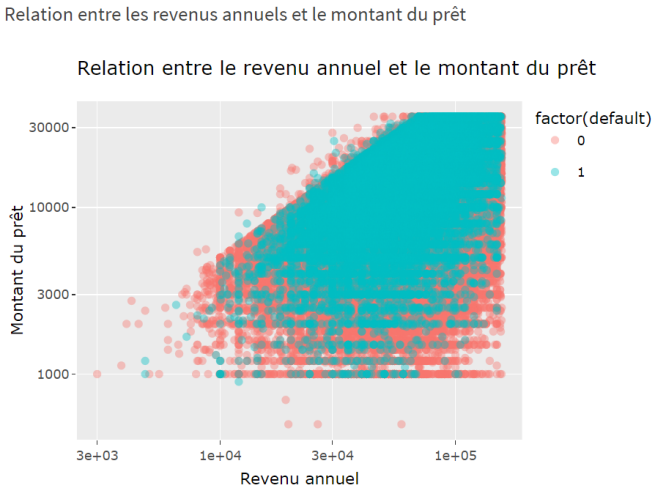

FIGURE 10 – Répartition des montants de prêt en défaut ou non

Ce nuage de point montre une corrélation positive entre le revenu annuel et le montant du prêt. Les défaut semblent présent sur toute la distribution mais avec une concentration plus forte sur les revenus les plus faibles.

On observe ici la relation entre le revenu annuel moyen et le montant moyen de prêt par grade.

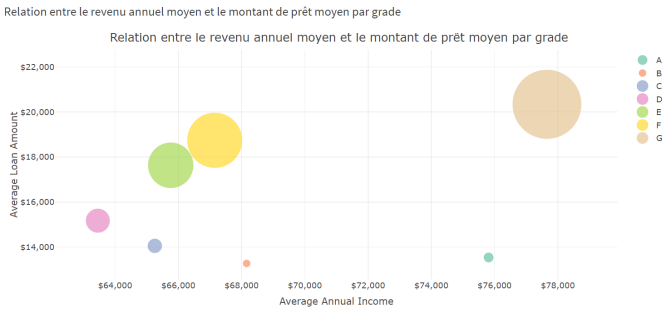

FIGURE 11 – Montant moyen de prêt par grade en fonction du revenu moyen annuel

Ce diagramme à bulles montre une tendance générale à l'augmentation du montant du prêt moyen avec l'augmentation du revenu annuel moyen. Cette tendance se voit sur les grades les plus représentatifs. On constate un déséquilibre au niveau de la population par grade.

Nous regardons aussi les variables corrélées entre elles pour limiter le biais dans nos prédictions.

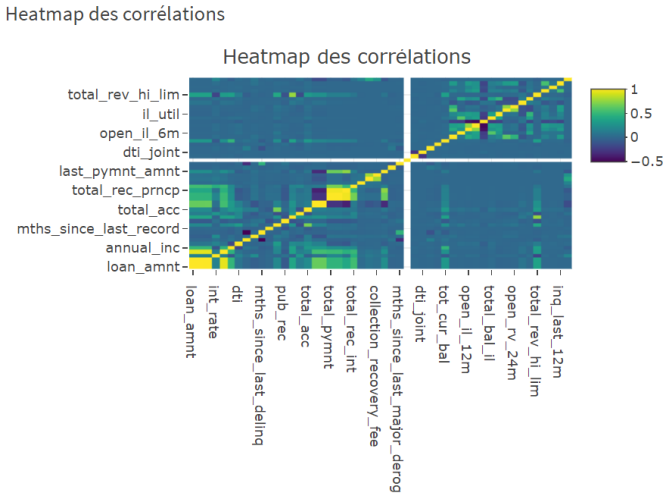

FIGURE 12 – Carte des corrélation

On observe que *total_pymnt* est fortement corrélée positivement avec *total_rec_prncp*, *total_rec_int* et *last_pymnt_amnt*. Cela est dû au fait que le paiement total est la somme du capital et des intérêts. De même *loan_amnt* est positivement corrélé avec *total_pymnt*, *total_rec_prncp* et *total_rev_hi_lim*. Plus la capacité d'emprunt est élevé et plus les prêt pourra être important avec des paiements conséquents. Au contraire, on constate que *int_rate* est faiblement corrélée négativement avec *loan_amnt* ce qui peut signifier que les prêts importants bénéficient d'un taux d'intérêt avantageux. *dti* montre une corrélation négative avec *annual_inc*, une revenu plus élevé pour une même dette aura un ration dette/revenu plus faible. Les autres variables présentent une corrélations faibles ou nulles.

Nous avons aussi jugé utile de regarder en détail le liens entre les variables *Loan Amount*, *Principal Received* et *Loan Condition*.

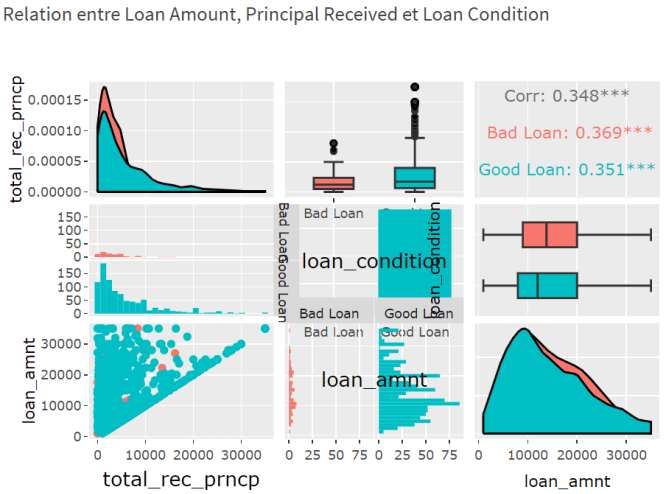

FIGURE 13 – Relation entre les variables *Loan_Amount*, *Principal_Received* et *Loan_Condition*

La couleur bleu indique les prêts sans défaut de remboursment et le rouge les prêts en défaut. Le

premier graphique en haut à gauche représente la distribution des paiements et des defauts de paiement. L'histogramme au milieu à gauche détaille l'analyse. On constate que les défauts se distinguent des paiements principalement en début de remboursement. Les défauts sont surtout concentré en début de remboursement même si on en trouve quelques uns en queue de distribution. En haut et au centre ainsi qu'au milieu à droite, on observe deux boxplots donnant la distribution des bons et des mauvais prêts. On observe plus de bons prêts que de mauvais prêts avec dans les deux cas la présence de valeurs extrêmes même si elles sont plus nombreuses dans le cas des bons prêts. En haut à droite, on a le coefficient de corrélation (0.348) entre le montant remboursé et le montant du prêt. Les trois astérisques (***) indiquent que cette corrélation est statistiquement significative ($p < 0.001$). Les corrélations pour les bons (0.351) et les mauvais prêts (0.369) sont également significatives. En bas à droite et en bas au milieu on observe un graphique de densité et un un histogramme, montrant la distribution du montant du prêt. La distribution semble légèrement asymétrique à droite, suggérant que la plupart des prêts ont un montant moyen, avec quelques prêts de montants beaucoup plus élevés. En bas à gauche, un nuage de points montre la relation entre le montant du prêt et le montant remboursé. On observe une corrélation positive : plus le montant du prêt est élevé, plus le montant remboursé tend a être élevé également.

2.4 Modélisation de la probabilité de défaut

Nous souhaitons maintenant modélisé la probabilité de défaut.

Pour cela, nous avons constaté dans la partie visualisation des données que notre base est déséquilibrée, nous avons donc appliqué l'algorithme SMOTE (Synthetic Minority Over-sampling Technique) pour le rééquilibrer.

L'algorithme SMOTE est une méthode d'over-sampling qui permet de rééquilibrer les données dans un contexte de classification. L'algorithme génère des exemples synthétiques pour les classes identifiées comme minoritaires afin de limiter le biais du modèle.

A la suite de cela nous avons choisi, pour la partie prédiction, une régression logistique ainsi qu'un Random forest.

Nous obtenons les matrices de confusion suivantes :

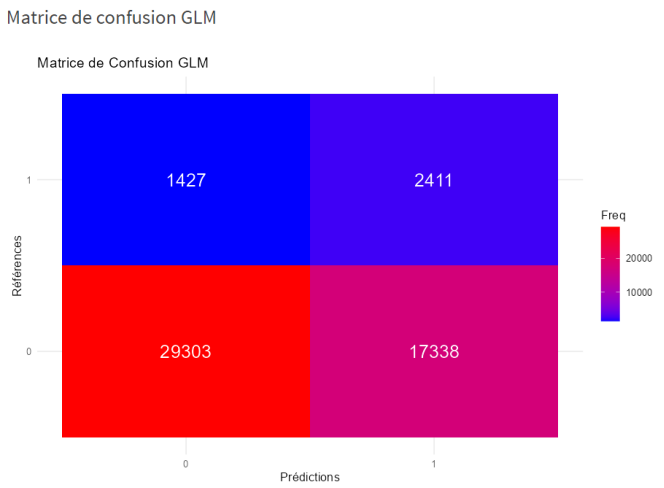

FIGURE 14 – Matrice de confusion pour le modèle GLM

Pour le modèle GLM, on observe 29 303 vrais négatifs, 2 411 vrais positifs, 1 738 faux positifs et 1 427 faux négatifs.

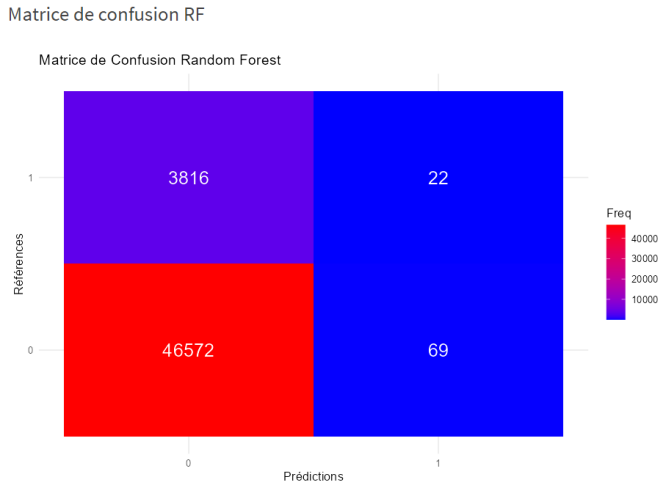

FIGURE 15 – Matrice de confusion pour le modèle RF

Pour le modèle Random Forest, on observe 46 572 vrais négatifs, 22 vrais positifs, 69 faux positifs et 3 816 faux négatifs.

Ainsi que les précisions suivantes :

	Précision
Régression logistique	62.83%
Random Forest	92.3%

TABLEAU 1 – Efficacité des deux modèles testés

Il apparait que le modèle Random Forest est beaucoup plus performant que la régression logistique.

Nous avons donc choisi de faire l'explicabilité du modèle Random Forest.

Nous avons étudié la dépendance partielle pour ce modèle pour les variables *annual_ inc*, *int_ rate* et *loan_ amnt*

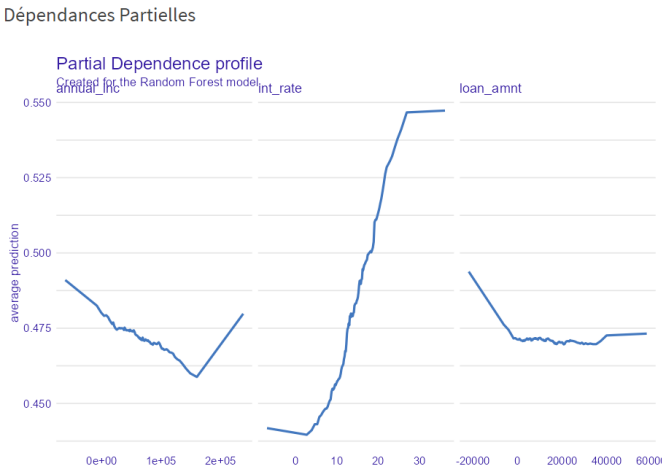

FIGURE 16 – Dépendance partielle pour le modèle Random Forest

Ce graphique montre comment la prédiction moyenne du modèle change lorsque l'on fait varier une variable, tout en moyennant l'effet des autres variables. Pour la variable *annual_ inc* observe une légère diminution de la prédiction moyenne lorsque le revenu annuel augmente jusqu'à environ 100 000 $. Cela suggère qu'un revenu plus élevé, est associé à une probabilité plus faible de défaut prédit par le modèle Au-delà de 100 000 $, la prédiction moyenne remonte légèrement, indiquant un effet moins prononcé du revenu sur la prédiction pour les revenus très élevés. Pour la variable *int_ rate* l'effet le plus marqué est une forte augmentation de la prédiction lorsque le taux d'intérêt augmente, surtout entre environ 5% et 20%. Cela indique qu'un taux d'intérêt plus élevé est associé à une probabilité plus élevée de prédiction de défaut. Au-delà de 20, l'augmentation de la prédiction moyenne se stabilise, suggérant que l'effet du taux d'intérêt devient moins important au-delà de ce seuil. Enfin pour la dernière variable *loan_ amnt*, on observe une légère diminution de la prédiction moyenne lorsque le montant du prêt augmente jusqu'à environ 20 000 $. Cela signifie qu'un montant de prêt plus élevé, est associé à une probabilité plus faible de prédiction de défaut. Au-delà de 20 000 $, la prédiction moyenne se stabilise, puis remonte légèrement au-delà de 40 000 $, indiquant un effet complexe du montant du prêt sur la prédiction.

On s'intéresse maintenant aux variables influentes pour la prédiction de défaut dans le modèle Random Forest.

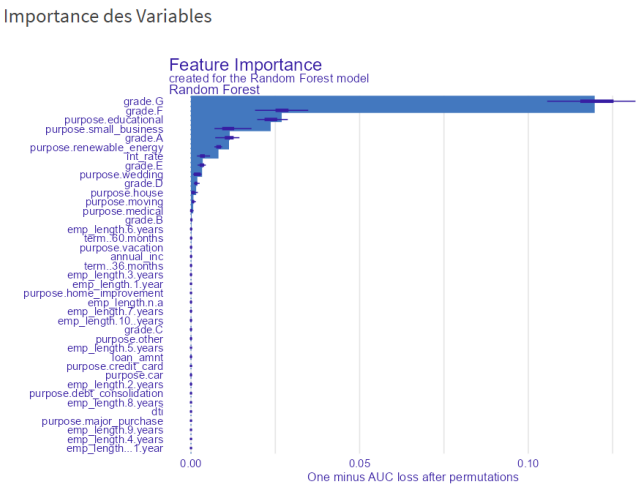

FIGURE 17 – Variables influentes

Les plus influentes dans ce modèle Random Forest sont *grade G*, *Grade F*, *purpose educational*, *purpose small_ business*, *grade A*, *purpose renewable_ energy*,*int_ rate*,*grade E*, *purpose wedding*, *grade D*, *purpose house* et *purpose moving*.

3 Conclusion

On a donc à travers ce projet étudier la prédiction de la probabilité de défaut. Après avoir analysé les données, implémenté notre application de visualisation Rshiny réaliser une modélisation pour la prédiction. L'interpretation graphique nous a permis de présenter les principales particularité de notre base importantes pour la modélisation. Nous avons identifié les variables influentes et leur impact tels que le grade, la durée du prêt, le montant du prêt, le revenu annuel et le taux d'intérêt.