Toulouse School of Economics

# Recherche des facteurs d'influence pour les défauts de paiement de cartes de crédit *(Etude de cas)*

Javier Alcaraz Ortega
Luke LO SEEN
Roman Gambelin

*Toulouse School of Economics*
*Magistère d'économiste-statisticien*

01$^{st}$ July 2017

# Table des matières

**Résumé**

Nous étudions ici le cas des défauts de paiement à Taïwan. Notre objectif est d'identifier les variables d'influence en utilisant deux modèles de prédictions : le premier basé sur une régression linéaire et le second sur une forêt aléatoire. La qualité des modèles est évaluée par le rappel et la précision (dans le cas des forêts aléatoires) des résultats. Au final, nous trouvons que pour diverses raisons, il est préférable de se fier l'analyse des résultats de la forêt aléatoire. Celle-ci montre que les facteurs les plus importants sont les statuts de paiement des 3 derniers mois, la quantité de dette, les quantités de paiement et surtout relevés de compte ainsi que l'âge. Ainsi, la moajorité des variables influentes sont liées aux données bancaires.

# 1 Introduction

Le prêt est l'une des majeures sources de revenu pour les banques de dépôt. Construire des modèles de prédiction du défaut de paiement fiables afin d'automatiser la prise de décision pour l'attribution de prêts constitue donc un enjeu dont l'importance a crû avec l'augmentation du volume d'activité des banques ces dernières décennies. En particulier, nous étudierons ici les défauts de paiement de cartes de crédit à Taïwan. L'île a fait face en 2006 à une crise de crédit de consommation, causée par le fait que les banques taïwanaises avaient depuis plusieurs années distribué des cartes de crédit en masse afin d'alimenter leur croissance, et ce sans tenir compte de la solidité financière des récipientaires.

Notre objectif est de trouver les facteurs d'influence des défauts de paiement. Cette problématique a déjà été abordée dans la littérature. Cependant, alors que nous étudions le lien entre le profil des clients et leur propension à faire faillite, la plupart des travaux susmentionnés (e.g. [Agarwal, S., & Liu, C., 2003], [Gross, D. B., & Souleles, N. S., 2002]) prennent une approche différente et se concentrent sur le lien entre la probabilité de défaut et des données macroéconomiques. L'intérêt de ces modèles est d'aider à évaluer la vraisemblance d'une crise de crédit à un certain moment, pour un certain endroit, ce qui est très intéressant pour les analystes. À l'inverse, notre but est de produire une analyse principalement destinée aux professionnels, en identifiant les charactéristiques des clients permettant de dire si oui ou non leur probabilité de se retrouver en défaut de paiement est élevée.

Nous allons dans un premier temps utiliser un logit pour tenter de prédire la probabilité de défaut d'un client, à partir de ses charactéristiques et de son activité bancaire sur les six précédents mois. Au préalable, nous aurons étudié plus en détail le lien entre les variables de notre échantillon et la variable à expliquer, afin d'affiner la spécification du modèle (i.e. voir si différentes catégories de population ont une propension différente à tomber en défaut de paiement, et quelles sont les catégorisations de population qui font le plus de sens au vu des données). Enfin, nous tenterons

aussi une autre approche : les forêts aléatoires ne nécessitant aucune hypothèse sur la distribution des données (contrairement au logit).

nous présenterons les méthodes utilisées dans la section 2 puis l'échantillon de données dans la section 3, et parlerons de l'indice à utiliser pour mesurer l'efficacité du modèle. Les résultats de l'analyse seront présentés puis expliqués dans la section 3, et enfin nous terminerons avec la conclusion dans la section 4.

# 2 Présentation des méthodes

## 2.1 Régression Logistique

La régression logistique est une extension des modèles de régression linéaire au cas des variables binaires. Dans le cas du logit, nous supposons que la probabilité d'obtenir 1 suit une loi logistique, dont l'espérane est déterminée par la spécification de notre modèle. La régression logistique est un modèle simple d'utilisation qui renvoie un résultat facile à interpréter, cependant il ne détectera qu'une relation linéaire entre la variable prédite et les variables prédictives. Aussi, une forte corrélation entre les variables explicatives corrompra l'estimation des coefficients.

## 2.2 Forêts aléatoires

L'algorithme des forêts d'arbres décisionnels, aussi appelé des forêts aléatoires, présenté par ([Breiman, 2001]), est un algorithme du domaine de l'apprentissage automatique. Il est basé sur la construction d'un bon nombre d'arbres décisionnels, qui sélectionnent un sous-ensemble de variables à chaque nœud et trouvent la séparation optimale des observations en fonction de ces variables, en termes d'homogénéité des deux sous-ensembles choisis. Ceci peut être réalisé avec différents critères, notamment en utilisant le critère d'entropie de Shannon ([Shannon, 2001]) ou de diversité de Gini ([Breiman et al.]). Finalement, il agrège les prédictions de ces arbres et produit une prédiction, normalement par vote à la majorité. Cependant, il est possible de définir un seuil particulier pour le pourcentage d'arbres prédisant l'appartenance à une classe à partir duquel le modèle global prédira l'appartenance à cette classe. Cet algorithme a l'avantage de ne supposer aucune distribution des données, et de pouvoir fonctionner sur un ensemble non homogène de variables qualitatives et quantitatives. En contrepartie, la présence d'un grand nombre d'arbres rend l'interprétation du modèle difficile.

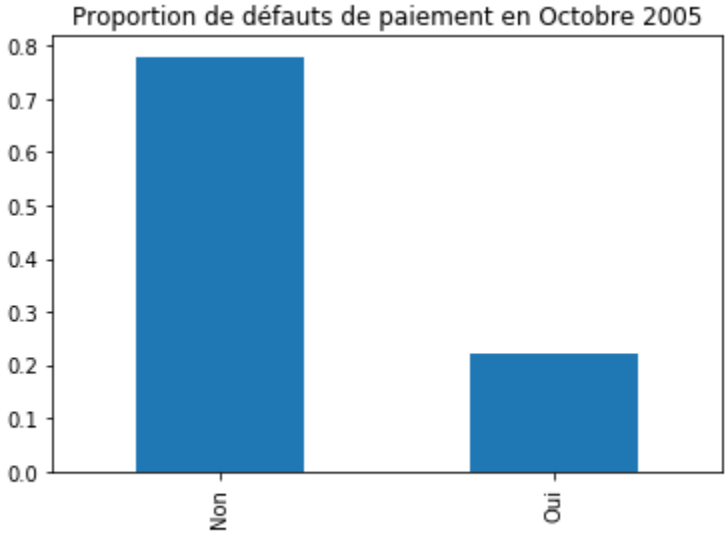
# 3 Présentation des Données

Les données sont issues d'une étude comparative de techniques de data mining ([Yeh& Lien, 2009]) et ont été mises à disposition du plus grand nombre via UCI ML, le compilateur d'échantillons de [Bache, K., & Lichman, M., 2013]. Nous disposons de 30000 observations de données bancaires issues d'une grande banque Taïwanaise, datant de 2005. Nous disposons des informations suivantes :

- ID : ID de chaque client
- LIMIT_BAL : Quantité de crédit en Dollars Taïwanais (NT), incluant celle des autres membres de la famille
- SEX : (1=home, 2=femme)
- EDUCATION : (1=Graduate School, 2=université, 3=lycée)
- MARRIAGE : (1=marié, 2=célibataire, 3=autre)
- AGE : en années (terriennes)
- PAY_0 : Statut de remboursement en September 2005 ($-2$ = aucune opération en cours, $-1$=paiement en règle, 0 =crédit revolving, 1 =paiement ajourné d'un mois, 2 =paiement ajourné de deux mois, ... 8 =paiment ajourné de 8 mois, 9 =paiement ajourné de neuf mois et plus)
- PAY_2 : Statut de remboursement en Août 2005 (même échelle)
- PAY_3 : Statut de remboursement en Juillet 2005 (même échelle)
- PAY_4 : Statut de remboursement en Juin 2005 (même échelle)
- PAY_5 : Statut de remboursement en Mai 2005 (même échelle)
- PAY_6 : Statut de remboursement en Avril 2005 (même échelle)
- BILL_AMT1 : Relevé de carte de crédit Septembre, 2005 (NT dollar)
- BILL_AMT2 : Relevé de carte de crédit Août, 2005 (NT dollar)
- BILL_AMT3 : Relevé de carte de crédit Juillet, 2005 (NT dollar)
- BILL_AMT4 : Relevé de carte de crédit Juin, 2005 (NT dollar)
- BILL_AMT5 : Relevé de carte de crédit Mai, 2005 (NT dollar)
- BILL_AMT6 : Relevé de carte de crédit Avril, 2005 (NT dollar)
- PAY_AMT1 : Quantité de paiement au mois précédent en Septembre, 2005 (NT dollar)
- PAY_AMT2 : Quantité de paiement au mois précédent en Août, 2005 (NT dollar)
- PAY_AMT3 : Quantité de paiement au mois précédent en Juillet, 2005 (NT dollar)
- PAY_AMT4 : Quantité de paiement au mois précédent en Juin, 2005 (NT dollar)
- PAY_AMT5 : Quantité de paiement au mois précédent en Mai, 2005 (NT dollar)
- PAY_AMT6 : Quantité de paiement au mois précédent en Suède, 2005 (NT dollar)
- default.payment.next.month : défaut de paiement ou non au mois d'Octobre

Ainsi, nous avons des informations sur le profil démographique du client (âge, marié ou célibataire, niveau d'étude...) et sur les opérations liées à leurs carte de crédit (quantité de crédit, statut de paiement...) entre Avril et Septembre. Des informations sur la répartition d'âge, le ratio H/F

ou encore la répartition du niveau d'éducation dans notre échantillon se trouvent en annexe. Nous cherchons à prédire, à partir de ces données, si un client fera défaut en Octobre ou pas.




Nous constatons que la proportion d'observation "d'intérêt" est minoritaire (22.12%).Dans notre sujet, il s'agit de prédire le défaut de paiement, qui est une variable binaire, en fonction de plusieurs variables tant qualitatives que quantitatives. De plus, nous ignorons les distributions de ces variables et les relations entre elles (dans ce sens, l'application des forêts aléatoires semble pertinente). En particulier, il s'agit d'un cas simple de classification en deux classes, que l'on peut considérer comme un modèle qui émet des avertissements sur les clients à risque d'être en défaut de paiement. Une considération importante est que le défaut de paiement est un évènement atypique, qui présente des conséquences potentiellement graves. Au contraire, le « non-défaut » de paiement est un évènement courant, et la prédiction d'un défaut de paiement qui ne se matérialise pas ne présente pas de coûts importants a priori.

Simplement regarder le taux d'erreur afin de mesurer la précision du logit risque de mal capturer le nombre de faux négatifs. Nous regarderons à la place la proportion de défauts de l'échantillon prédite par le modèle (i.e. le rappel). Concernant les forêts aléatoires, nous allons combiner deux mesures courantes de performance d'une classification : la précision et le rappel (la précision est la proportion d'alarmes de défaut qui se matérialisent en défaut).

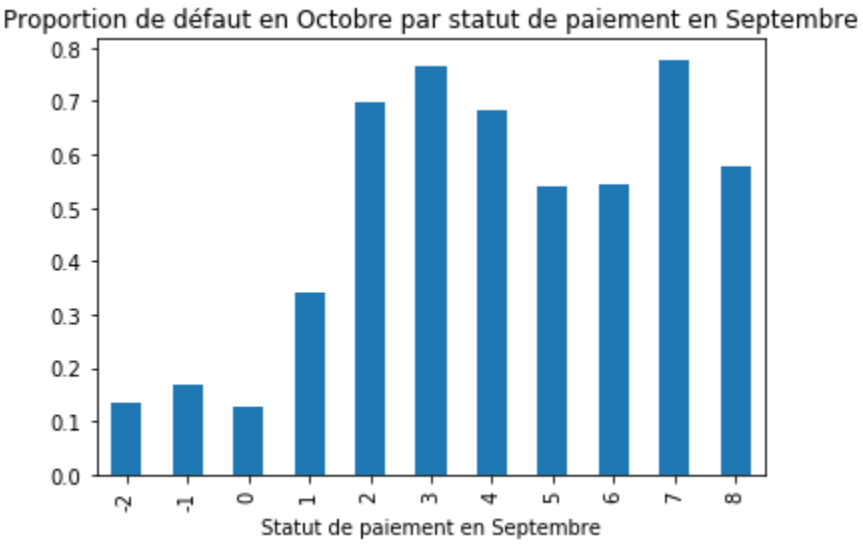
# 4 Présentation et Analyse des Résultats

## 4.1 Exploration de l'échantillon

Dans cette partie, nous avons testé (à un niveau de confiance de 95%) l'existence de différentes propensions au défaut de paiement pour différentes catégories de populations déterminées par une variable donnée.

Concernant les charactéristiques individuelles, il est ressorti que pour le sexe, la situation matrimoniale et le niveau d'étude, il y avait des différences significatives entre les différentes classes. En revanche, en divisant l'âge en 4 catégories (20-30 ans, 30-40 ans, 40-50 ans, 50-70 ans), nous avons remarqué qui'il n'y avait pas de différence significative entre la propension de défaut des 20-30 ans et celle des 40-50 ans. Nous en avons ainsi conclu qu'il est possible de simplifier notre modèle en fusionnant ces deux catégories d'âge dans notre spécification.

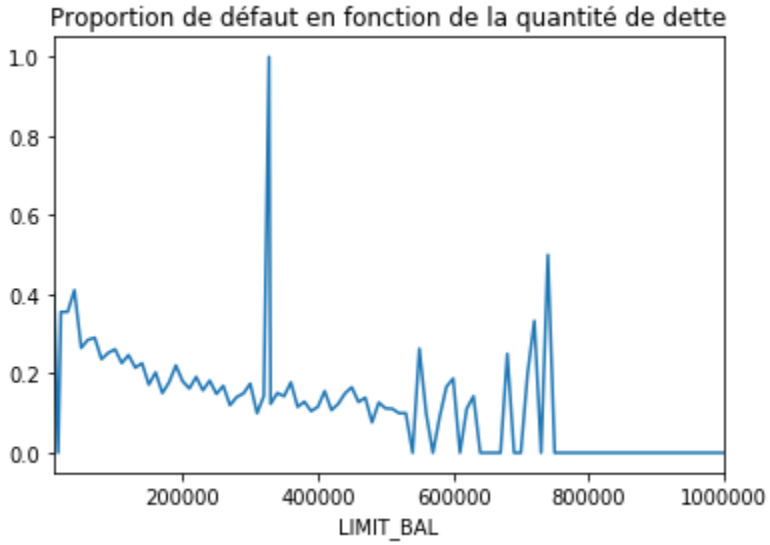
Pour les données bancaires, nous connaissons le statut de paiement (pay_i), les relevés de carte (Bill_AMTi), la quantité de remboursement (Pay_AMTi) sur les 6 derniers mois, ainsi que le montant de la dette (Limit_Bal). Ci-dessous, nous avons représenté la proportion de défauts en fonction du statut de paiement le mois précédent :




Les statuts -2 ;-1 et 0 correspondent à des situations où le paiement n'est pas décalé ce mois-ci,ce qui peut être un signe de solvabilité. Cette intuition semble être confirmée par le graphique ci-dessus. On observe aussi que la proportion de défaut semble être beaucoup plus élevée à partir d'un ajournement de 2 mois et plus.La valeur 1 semble être intermédiaire par rapport aux deux autres groupes. Ainsi, 3 catégories de Pay_0 semblent se dessiner pour notre spécification. Nous avons répété l'opération pour les autres statuts de paiement, et le motif semble se répéter. Par

ailleurs, nous avons aussi trouvé une forte corrélation entre Pay_0 et les autre valeurs de Pay (allant de 0.67 pour Pay_2 à 0.48 pour Pay_6). Et que la variable "default.payment.next.month" (i.e. défaut de paiement au mois d'Octobre) était surtout corrélée avec Pay_0. Ainsi, nous en avont conclu que le mieux était de ne garder que Pay_0 et de diviser cette variable en fonction des catégories spécifiées plus haut.

Pour la quantité de dette, après avoir confirmé ue cette variable était bien significative, nous avons représenté la proportion de défaut en fonction du montant de la dette :




Nous observons un pic autour des 35,000 NT de dette. La proportion de défaut n'est pas une fonction strictement croisssante du montant de crédit, contrairement à ce qu'on pourrait attendre. Une explication de ce phénomène serait que différentes catégories de population ont différents niveaux de dettes : le pic de 35000 NT correspondrait à la quantité maximale de crédit que peut supporter la classe populaire/moyenne. Et seuls les lus riches peuvent atteindre la queue de la distribution sans tomber en défaut de paiement au préalable. Ainsi nous distinguons trois classes pour cette variable : la catégorie des emprunteurs "vertueux", celle des emprunteurs "populaires" à risque, et celle des emprunteurs "riches" à risque (i.e. 0-200k ;200-400k ;+400k).

Concernant les relevés de cartes de crédit et la quantité de paiement, nous avons observé que leur corrélation avec la variable à prédire ne dépasse pas -7.6%. Nous les avons donc enlevées de notre modèle.

## 4.2 Régression

Nous nous retrouvons donc avec la spécification suivante pour le logit :

$$P(Default_i \mid X_i) = \frac{\exp(\beta X_i)}{1 + \exp(\beta X_i)}$$

Avec :

$$X_i = (sexe_i, Edu_i, Mariage_i, Age1_i, Age2_i, Pay1\_0_i, Pay2\_0_i, creditA_i, creditB_i)$$
$$\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9)$$

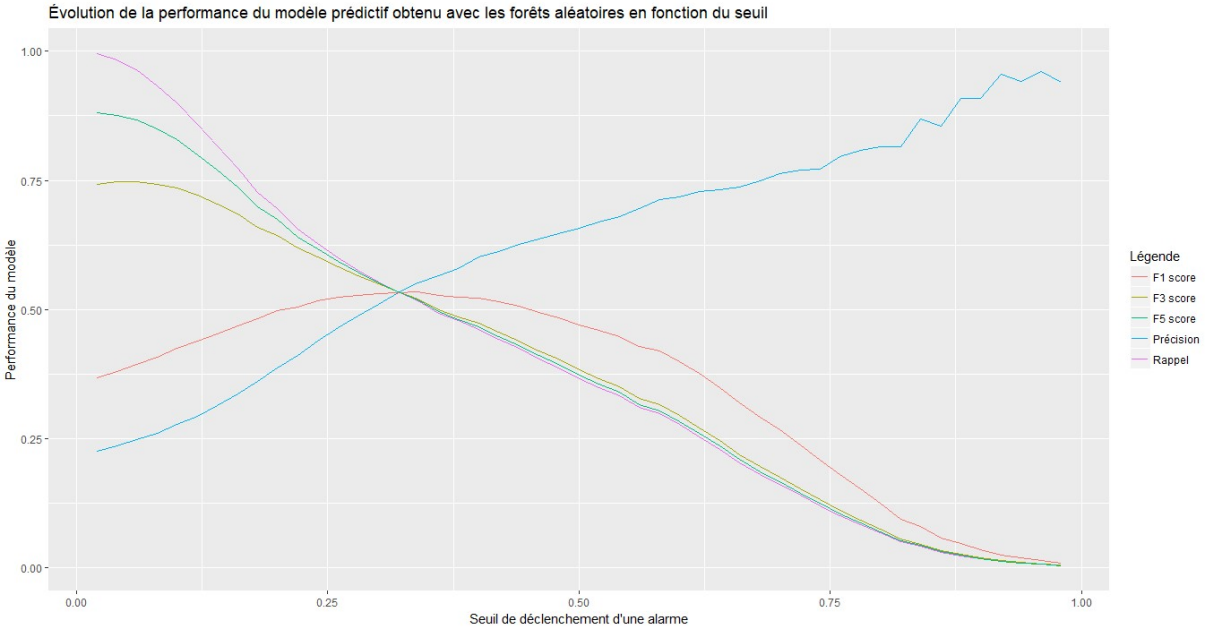
Ici, **Age1** représente la catégorie de gens de 20-30 ans/40-50 ans et **Age2** celle des gens de 30-40 ans. **Pay1_0** représente les gens qui ont ajourné de 1 mois leur paiement en Septembre, et **Pay2_0** ceux qui ont ajourné de au moins deux mois leur paiement. Pour finir, **creditA** est un booléen indiquant si le client a moins de 20,000 NT de dette, et **creditB** si cette dette est comprise entre 20,000 et 40,000 NT.

Nous avons fait face à une légère déconvenue en vérifiant l'efficacité du modèle : nous nous sommes retrouvé avec une valeur de 0.33 pour le recall. En d'autres termes, notre modèle ne prédit qu'un tiers des défauts de paiement de l'échantillon. Ce résultat est à mettre en relation avec [Yeh& Lien, 2009], qui montre qu'n terme d'efficacité, la régression logistique n'est pas grandement inférieure aux autres algorithmes utilisés en data mining.

Aussi, lorsque l'on utilise toutes les varibales de l'échantillon dans notre régression, nous nous retrouvons alors avec un recall encore plus faible, proche de 0. Notre spécification améliore donc substantivement les résultats du modèle "de base" où l'on prend toutes les variables en compte.

## 4.3 Forêts aléatoires

L'algorithme est testé sur l'ensemble de notre base de données, en utilisant la librairie « randomForest » de R. Nous varions le seuil de déclenchement d'une alarme entre 0 et 1, avec un pas de 0.02, et nous représentons les résultats obtenus dans le graphique ci-dessous. Nous pouvons y voir l'évolution de la précision et le rappel, ainsi que plusieurs F-scores. Le F-score correspond à la moyenne harmonique entre précision et rappel, et le chiffre qui accompagne le F représente le poids relatif entre précision et rappel.



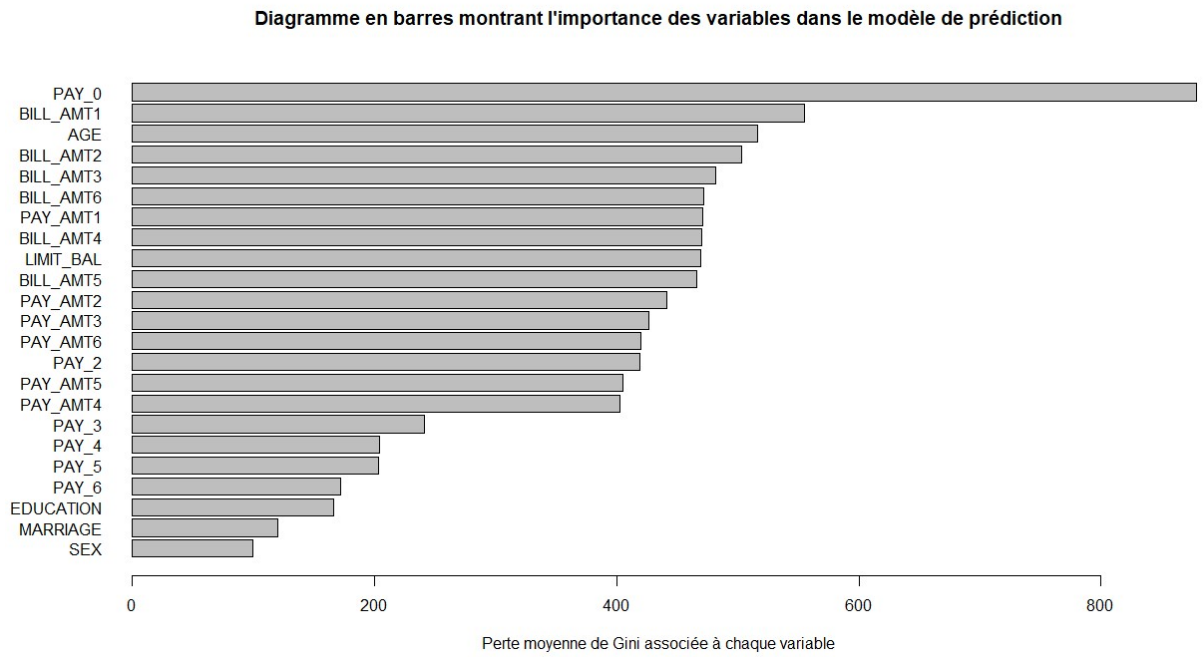

Comme dit précédemment, nous pensons que dans ce cas il est judicieux de maximiser le rappel. Si nous considérons qu'un défaut non prédit est 5 fois plus grave qu'une prédiction d'un défaut qui ne se matérialise pas, nous devons alors baser notre choix du seuil en fonction du score F5. Dans ce cas, cela conduit à un seuil de 0.02. Ce seuil est jugé excessif, donc nous allons utiliser plutôt le score F3, dont la maximisation conduit à un seuil de 0.06, qui est donc retenu pour la suite de l'étude. Afin de valider les résultats, nous entrainons le modèle avec ce seuil sur 80% des observations, et il est ensuite testé sur le 20% restant. La matrice de confusion est affichée ci-dessous :

```
> optimal_model$test$confusion
    0    1 class.error
0 380 1953  0.83711959
1  13  648  0.01966717
```

Ces résultats sur l'échantillon de test conduisent à une précision d'environ 0.25 et un rappel de 0.98, qui est presque identique aux résultats obtenus sur l'échantillon d'entrainement. Ceci n'est pas surprenant puisque les forêts aléatoires ont une faible tendance au sur-ajustement. Finalement, nous testons le modèle également sur la base de données modifiée selon les critères expliquée dans la partie précédente, sur laquelle nous avons étudié la régression logistique. Cependant, comme nous pouvons voir dans la matrice de confusion ci-dessous, les résultats sont loin des attentes. Ce modèle semble donc assez sensible au nombre de variables, ce qui fait que le seuil devrait être choisi séparément dans chaque cas.

```
> optimal_model$test$confusion
     0   1 class.error
0 4358 236  0.05137135
1  782 458  0.63064516
```

Pour conclure, nous présentons ci-dessous l'importance des variables dans le modèle, mesurée par la perte moyenne de Gini lors d'une permutation de ses valeurs. Cette méthode est souvent utilisée, même s'il est connu qu'elle présente un biais positif vers les variables quantitatives. Comme lors de l'exploration de l'échantillon, nous trouvons que des variables comme l'âge ou Pay_0 ont un fort impact sur les défauts. Cependant, Nous pouvons aussi remarquer que des variables non spécifiées dans la régression logistique (i.e. les quantités de paiement) apparaissent en tête alors qu'elles avaient une très faible corrélation avec la variable prédite. Il est possible que ces variables n'aient pas une relation d'ordre linéaire avec les défauts, et donc qu'elles soient mieux intégrées dans le modèle par les forêts que par une régression. Étant donné que la régression logistique est incapable de capter une partie de l'information de l'échantillon, il est préférable de se baser sur les résultats de la forêt aléatoire pour la conclusion.




# 5 Conclusion

Dans cette étude, nous avons formulé une spécification pour le logit qui améliore la régression "de base" avec toutes les variables. Cependant, le modèle reste assez peu efficace puisqu'il ne prédit que 1/3 environ des défauts de notre échantillon. Ceci dit, [Yeh& Lien, 2009] montre que les autres modèles ne sont pas beaucoup plus efficaces. Quant aux explications de ce score assez faible, on peut être tenté de penser que la relation entre le défaut de paiement et les autres variables de l'échantillon n'est pas linéaire. Dans ce cas, des modèles tels que les forêts aléatoires ou les réseaux neuronaux devraient avoir de meilleures performances. Dans son papier, [Yeh& Lien, 2009] trouve que c'est le cas.

Nous avons donc ensuite testé les forêts aléatoires. En baissant le seuil d'alerte (i.e. la proportion minimale d'arbres prédisant un défaut pour que la forêt prédise un défaut), nous avons pû obtenir un bon niveau de rappel au prix d'un faible niveau de précision. Nous calculé la perte moyenne de Gini en permutant les valeurs. Nous trouvons que des variables écartées pour la régression logistique étaient en réalité significative. Ainsi, nous pouvons formuler l'hypothèses qu'une relation non-linéaire lie ces variables,qu'une régression logistique ne peut pas détecter.

En nous basant sur la perte de moyenne du Gini, nous observons que les variables isses des données bancaires (statut de paiement des 3 derniers mois, quantité de dette, quantité de de paiement et surtout relevés de compte) ainsi que l'âge sont les données plus influentes.

# Références

[Agarwal, S., & Liu, C., 2003] Agarwal, S., & Liu, C. (2003). Determinants of credit card delinquency and bankruptcy : Macroeconomic factors. Journal of Economics and Finance, 27(1), 75-84.

[Breiman, 2001] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[Breiman et al.] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

[Gross, D. B., & Souleles, N. S., 2002] Gross, D. B., & Souleles, N. S. (2002). Do liquidity constraints and interest rates matter for consumer behavior ? Evidence from credit card data. The Quarterly journal of economics, 117(1), 149-185.

[Bache, K., & Lichman, M., 2013] Bache, K., & Lichman, M. (2013). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2013. URL:http://archive.ics.uci.edu/ml

[Kocenda, E., & Vojtek, M., 2009] Kocenda, E., & Vojtek, M. (2009). Default predictors and credit scoring models for retail banking.

[Shannon, 2001] Shannon, C. E. (2001). A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 5(1), 3-55.

[Yeh& Lien, 2009] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.