


Université D'ANTANANARIVO

Domaine Sciences et Technologies

Mention Mathématiques et Informatique

MISA

Mémoire en vue de l'obtention du diplôme de Master 2 en Mathématiques Informatique et Statistique Appliquées

---

Mise en place d'un système d'attribution de points d'une demande de prêts (CREDIT SCORING)

Présenté par :

Sandratra Herimanana ANDRIANIRINA

Devant le jury composé de :

Président :	M. *Solofoniaina* Joelson	*Université d'Antananarivo*
Examinateur :	M. *Robinson* Olivier	*Université d'Antananarivo*
Encadreur :	M. *Randrianarivony* Arthur	*Université d'Antananarivo*
Co-Encadreur :	M. *Rasoanaivo* Andry	*Université d'Antananarivo*

Année universitaire 2017/2018

Ambohitsaina, le 08 Février 2019

Remerciements

Je tiens à exprimer ma gratitude et mes remerciements à tous ceux qui m'ont, de près ou de loin, apporté leur aide et qui ont contribué à l'élaboration de ce mémoire. Mes plus vifs remerciements s'adressent à ma famille pour leur soutien durant toutes mes années d'études.

Je tiens à remercier, Messieurs Tahiry ANDRIAMAROZAKANIAINA, Olivier ROBINSON, Andry RASOANAIVO ainsi que le corps enseignant de la MISA de m'avoir permis de suivre cette formation, pour les connaissances et les conseils qu'ils ont prodigués.

Je réitère mes remerciement pour Monsieur Andry RASOANAIVO , en tant qu'encadreur de stage, pour ses directives et son encadrement pour mener à terme ce mémoire de fin d'études.

J'exprime mes remerciements à Monsieur Arthur RANDRIANARIVONY, mon encadreur pédagogique, pour ses conseils et sa disponibilité.

Je suis reconnaissant envers les membres du jury qui ont accepté de juger mon travail.

Enfin, je ne saurai manquer d'exprimer ma gratitude à mes collègues et amis de la promotion de Master MISA 2018 pour le partage de connaissances et d'entraides.

Merci.

Table des matières

Glossaire

AUC Aire sous la courbe, est calculé avec les taux de vrais positifs et taux de vrais négatifs.

BNI Banky Nasionalin'ny Indostria.

FN Faux Négatif.

FP Faux Positif.

FPR False Positive Rate (en français TFP).

K-Fold Validation croisée ou « cross-validation » sur K sous échantillons.

LDA Analyse Discriminante linéaire.

RNA Réseaux de neurones Artificiels.

ROC receiver operating characteristic.

SWAP Espace d'échange servant à étendre la mémoire utilisable par un système d'exploitation.

TPR True Positive Rate (en français TVP).

VN Vrai Négatif.

VP Vrai Positif.

Introduction

Depuis toujours, les instituts financiers apprécient la capacité de remboursement des prêts de leurs clients à partir de leurs données professionnelles et aussi personnelles. Ces institutions prennent en compte notamment les informations telles que la rémunération, la stabilité du travail, l'âge, etc... Ce n'est que depuis le début des années 1990 que les instituts financiers se sont intéressés à l'utilisation des outils statistiques et de machine Learning pour apprécier la capacité de remboursement des clients à partir des données professionnelles et personnelles. Bons nombres de travaux sur le crédit scoring ont été fait depuis. Effectivement, ce domaine suscite grandement l'intérêt des instituts financiers puisqu'un minimum de discernement entre bons payeurs et mauvais payeurs leur rapporte à la fois plus de profits, en fournissant du crédit aux bons payeurs, et moins de pertes en refusant du crédit aux mauvais payeurs. Le procédé du crédit scoring est pratiquement le même : L'apprentissage du modèle se fait avec des données historiques. D'abord, la collecte de données, car il faut au moins deux ans de données pour avoir assez d'informations et de cas distincts pour l'apprentissage. Sur ce point, on retrouve quelques standards sur les données collectées. Ensuite, la phase de collecte de données est suivie de la création du modèle qui permettra l'appréciation du client communément dite crédit Scoring. Notons que la plupart des travaux consacrés sur le crédit scoring sont sur cette dernière phase. En d'autre termes, on a voulu comparé les différents algorithmes de modèle de prédiction sur le Credit Scroring. Néanmoins, le modèle le plus performant varie selon les données d'apprentissage. Notre travail de six mois se déroulera comme suite : Tout d'abord, nous allons brièvement faire une analyse des existants dans le domaine du credit scoring en général, suivi d' une analyse spécifique du crédit scoring antérieur de la BNI. En second lieu, nous présenterons notre travail de construction du nouveau système de la BNI. En comparant plusieurs algorithmes d'apprentissage, nous avons choisi un modèle parmi les quatre modèles que nous avons construits pour le nouveau système de crédit scoring de la BNI. En dernier lieu, nous avons pris part à un concours organisé par Kaggle et Home Credit sur le même thème, c'est à dire le Credit Scoring. Durant ce concours, nous avons construit un nouveau modèle créé à partir des données de Home Credit. Ainsi, à partir de ce dernier modèle, nous pouvons suggérer quelques propositions d'amélioration au modèle de la BNI.

Chapitre 1

Contexte

1.1 Contexte du crédit scoring

La survie de chaque banque est toujours affectée par plusieurs types de risques à savoir, le risque de marché, de crédit, ... [10]. Le risque qui nous intéresse ici est un risque de crédit aussi appelé risque de contrepartie. Lors d'une demande de crédit, qu'il s'agisse d'une carte de crédit, d'un prêt automobile, d'un prêt personnel ou d'une hypothèque, le créancier voudra connaître le niveau de risque de crédit de chaque emprunteur. Devant la crise financière mondiale actuelle, notamment les échecs successifs de certaines banques internationales célèbres, les méthodes classiques de gestion du risque de crédit ont été remises en cause dans la plupart des pays. Ce risque doit être traité par des méthodes plus sophistiquées. Nous pouvons dire que les banques craignent un non-paiement des crédits empruntés et cherchent ainsi à réduire ce risque. En effet, le non-remboursement des prêts peut compromettre toute activité de la banque en la rendant inactive. Ce qui nous amène à étudier la gestion de risque de crédit et d'en analyser sa politique. Depuis que cette exposition au risque de crédit des banques ne cesse de s'accroître, les superviseurs et les banques elles-mêmes devraient être capables de monter un scénario se basant sur les expériences passées et de prévenir à ce risque. Ainsi, cela va aussi bien conscientiser chaque banque à identifier, surveiller et contrôler son risque de crédit aussi bien qu'à prévoir le capital suffisant contre ces risques. Grâce à cette analyse, chaque banque pourra se sortir indemne, de la manière la plus sûre, des risques à encourir. D'où l'utilité du crédit scoring.

1.2 État de l'art

1.2.1 Définition

Nous rappelons que le crédit est une mise à disposition d'argent sous forme de prêt, consentie par un créancier (prêteur) à un débiteur (emprunteur). Le « crédit scoring » est une action d'évaluer le risque de défaillance de crédit. C'est le processus d'évaluation du risque de crédit. Cet outil est mis en œuvre lors de l'analyse de risques d'une demande de prêts. Le « crédit scoring » est un ensemble de modèles de décision et de leurs techniques sous-jacentes qui aide les prêteurs dans l'octroi du crédit à la consommation. Il figure parmi les applications les plus réussies de la modélisation statistique et est le plus utilisé par les organismes financiers, les banques et les agences de crédits. Il leur est indispensable pour faciliter la prise de décision sur l'octroi du crédit, ne se contentant pas de la bonne foi du demandeur mais en se fiant aux divers critères de cet outil. Il permet également d'évaluer la probabilité de remboursement du prêt par l'emprunteur. Grâce à cet outil, les organismes tels que les banques sont en mesure de prendre une décision objective concernant l'octroi d'un crédit avec la garantie en plus que chaque demande de crédit est traitée selon les mêmes normes. Aussi, cela leur permet de distinguer les bons des mauvais payeurs et pour y arriver, elles ont besoin de plus de données à traiter et analyser pour mesurer le score de chaque client.

1.2.2 Méthodes

Pour pouvoir attribuer un score à un emprunteur, il est nécessaire de posséder les données sur ce dernier. Le créancier pourra ainsi faire des analyses et aboutir à la modélisation des données existantes qui vont être utiles pour cette étude.

1.2.2.1 Types de données

En général, pour se procurer les données, la banque a à son usage la base de données de ses clients (sinistrés ou non) ainsi que les données des emprunteurs dont la demande de crédit a été acceptée. Quel que soit l'organisme qui utilise le crédit scoring, les facteurs pris en compte pour l'établissement des scores sont souvent :

- Indicateur financier
- Indicateur démographiques
- Indicateur d'emploi

Chaque indicateur est composé de données qui vont aider à l'identification du type de l'individu étudié. Comme exemple, nous avons le cas du Crédit Scoring Survey ayant quatre indicateurs dont le surplus est un indicateur comportemental dans lequel on peut suivre les mouvements d'un client particulier.

- Indicateur financier
  - Total des actifs de l'emprunteur
  - Revenu brut de l'emprunteur
  - Revenu brut du ménage
  - Coûts mensuels du ménage
- Indicateur démographique
  - Âge de l'emprunteur
  - Sexe de l'emprunteur
  - Situation matrimoniale de l'emprunteur
  - Nombre de personnes à charge
  - Statut domicile (propriétaire ou locataire)
  - District d'adresse
- Indicateurs d'emploi
  - Type d'emploi
  - Durée de l'emploi actuel
  - Nombre d'emplois au cours des x dernières années
- Indicateurs comportementaux
  - Vérification du compte
  - Solde moyen
  - Prêts en cours
  - Prêts en souffrance ou délinquant
  - Nombre de paiements par année

Les données contenues dans ces indicateurs sont indispensables et fiables du fait que ces critères permettent de cerner la nature, le comportement et le respect des engagements du candidat. De ce fait, insérer un maximum de critères comme l'âge, l'état civil, le type d'habitat, la situation du logement, la profession et le nombre d'enfants du candidat dans

la base de données permettra à l'organisme d'avoir un aperçu global sur la situation du candidat et son aptitude à faire face à ses responsabilités. Mais si les critères de solvabilité procurent quelques renseignements sur sa personnalité, elles ne permettent de déterminer ni son niveau de vie, ni sa capacité à rembourser.

Il faut également prendre en considération les critères de solvabilité du candidat. Pour ce faire, son revenu, sa profession, son ancienneté et sa catégorie professionnelle seront donc également insérés dans l'outil pour savoir si le candidat dispose des ressources nécessaires pour effectuer le remboursement du crédit.

Pour le recueil des données, seules celles qui sont issues d'une demande acceptée sont à utiliser pour pouvoir qualifier le client de « Bon » ou « Mauvais » payeur. Quelques contraintes sont cependant observées du fait que les dossiers refusés ne sont pas pris en considération. Par ailleurs, la proportion des données de clients non sinistrés et sinistrés doit être équitable pour ainsi construire un bon modèle de données.

1.2.2.2 Critères d'un bon modèle de données

Le modèle est bon quand le taux d'erreur est minimum c'est-à-dire le modèle a mal classé le minimum de dossiers (dossiers déclarés sinistrés par le modèle mais qui sont en réalité acceptables et inversement)
Parmi les données à disposition, seules les données propres sont à utiliser c'est-à-dire les données dont les informations sont complètes (pas de manque, pas de données anormales)
Pour une bonne modélisation, seules les variables discriminantes seront utilisées. Ces variables feront l'objet d'une analyse descriptive :

- Pour les variables continues (comme l'âge, montant du prêt, Revenu mensuel) on calculera le nombre d'observations, la moyenne, l'écart-type, la médiane, le minimum et le maximum afin d'observer la dispersion des données de ces variables et d'en déduire leur pertinence ;
- Pour les variables modales (Type de contrat, Situation matrimoniale, ...), les fréquences et les nombres de dossiers concernés seront calculés en pourcentage.
- Analyses multidimensionnelles

1.2.2.3 Techniques de Crédit Scoring

: Il y a plusieurs techniques de crédit scoring

- Approche géométrique, connue sous le nom de règle de Mahalanobis –Fisher
- Techniques basées sur la statistique

  - Classifieur Bayésien ‘
  - Analyse discriminante
  - régression linéaire
  - régression logistique
  - Arbre de décision

- Techniques basées sur l’intelligence artificielle
  - Réseaux de Neurones
  - Algorithme génétique

Approche géométrique ou Règle de Mahalanobis-Fisher :

Dans cette approche, les données des anciens clients seront utilisées pour définir à l’avance des groupes d’individus, ces groupes sont définis de telle façon que les individus au sein d’un même groupe aient à peu près les mêmes caractéristiques. Par exemple, on va définir deux groupes à l’avance, celui des bons payeurs et celui des mauvais payeurs, les anciens clients sont alors classés dans ces groupes selon leurs statuts. Et pour chaque groupe, le centre de gravité sera calculé et ce dernier sera un nouvel individu considéré comme le représentant de son groupe. Pour un nouvel individu, le calcul des distances entre le nouvel individu et le centre de gravité de chaque groupe va déterminer le classement de cet individu. En effet, ce nouvel individu va appartenir au groupe le plus proche.

L’approche géométrique, connue sous le nom de règle de Mahalanobis – Fisher, consiste tout simplement à classer cet individu dans le groupe le plus proche. On sait qu’un individu et un représentant d’un groupe sont deux points de R^p. Le groupe recherché est donc celui pour lequel la distance entre son centre de gravité et le point individu concerné est la plus petite.

La métrique considérée pour calculer cette distance est celle de Mahalanobis définie par l’inverse de la matrice intra groupes W.

$$W = \sum_{k=1}^{m} \pi_k V_k \text{ [7]}$$

π_k poids du groupe k, $\pi_k = \sum_{i \in E_k} p_i$,
et p_i est le poids associé à chaque individu i
V_k est la matrice de variance, covariance du groupe k. [7]
Pour un nouvel individu $X = (x_1, x_2, \ldots, x_j, \ldots, x_p)$, on a $d^2(X, g_k)$ le carré de la distance

entre X et g_k (centre de gravité du groupe k)

$$d^2(X, g_k) = (X - g_k)'W^{-1}(X - g_k) = X'W^{-1}X + g_k'W^{-1}g_k - 2X'W^{-1}g_k \text{ [7]}$$

Et le premier terme du dernier membre précédent ($X'W^{-1}X$) ne dépend pas des groupes alors on peut se limiter à :

$$S_k(X) = X'W^{-1}g_k - (1/2)g_k'W^{-1}g_k \text{ [7]}$$

Cette quantité sera calculée pour chaque groupe et le groupe qui aura la plus grande quantité sera le groupe d'affectation du nouvel individu.

Techniques basées sur les Statistiques :

Classifieur Bayésien

Beaucoup de méthodes de classification dans le cadre probabiliste se basent sur ce classifieur de Bayes [7]. Le principe de cette technique est le même que pour l'approche géométrique, c'est-à-dire on cherche à classer un individu dans l'un de quelques groupes définis à l'avance sauf qu'on se situera dans un cadre probabiliste. Comme dans l'approche géométrique, les données des anciens clients seront utilisées pour définir des groupes à l'avance. En effet, soit une population E de n individus repartis entre m groupes définis à priori. Étant donné un individu e qu' on cherche à classer dans l'un des groupes, cet individu peut être considéré comme le résultat d'une expérience aléatoire de tirage au hasard d'un élément de E.
La formule de Bayes permet d'exprimer la probabilité à posteriori d'appartenir au groupe k sachant que la variable prend la valeur x ;

$$P(e \in E_k \text{ sachant } X(e) = x) = \frac{p_k f_k(x)}{\sum_{k=1}^{m} p_k f_k(x)} \text{ [7]}$$

où :
p_k : Probabilité d'appartenance au groupe k. $p_k = P(e \in E_k)$ appelé probabilité à priori
$f_k(x)$: densité de probabilité du vecteur X dans le groupe k lorsque X est absolument continue ($f_k(x) = P(X = x/e \in E_k)$

Analyse discriminante

Une autre méthode de l'approche statistique est l'analyse discriminante de Fischer, cette méthode est la plus ancienne des méthodes statistiques de classement [7]. Elle est une variante du classifieur de Bayes car elle permet également de classer les individus dans différents groupes définis à l'avance et on se situe encore dans le cadre probabiliste. Le

modèle de notation de crédit basé sur une approche discriminante est essentiellement utilisé pour l'analyse statistique afin de classer des groupes de variables en deux catégories ou plus [1]. Dans cette technique, les descripteurs X sont des variables aléatoires continues et supposées suivre une loi normale à chaque groupe E_k. C'est donc une méthode où les probabilités conditionnelles à estimer sont supposées relevées de lois de probabilité données mais dépendant néanmoins de paramètres inconnus à estimer à partir des données mises à disposition (estimation paramétrique) [7]. De nombreux chercheurs sont convenu que le discriminant approche est toujours l'une des techniques les plus largement établies pour classer les clients comme bons crédits ou mauvais crédits [1].

Existant

Boubacar Diallo du laboratoire d'économie d'Orléans [4] a construit le modèle de crédit scoring pour une institution de microfinance malienne à partir d'un échantillon de 269 emprunteurs de l'institution avec l'analyse discriminante et a obtenu un résultat de plus de 70% de bonne prédiction, soit environ 30% de taux d'erreur.

Régression linéaire :

La régression linéaire est parmi les méthodes utilisées pour classifier. C'est une technique de datamining visant à prédire une valeur numérique (telle que le score). Il y a plusieurs types de régression mais ceux qui nous intéressent sont ceux utilisés pour le crédit scoring. Tout d'abord on a la régression linéaire, cette technique permet d'établir une relation linéaire entre une ou plusieurs variables indépendantes appelées aussi variables expliquées et les variables dépendantes ou variables explicatives. On a une relation de la forme :

$$Y = a_0 + a_1 X_{i,1} + ... + a_k X_{i,k} + \epsilon_i$$

Où :

Y_i est la variable expliquée pour un individu i - cette variable sera donc dans notre cas le score de l'individu.

$X_{i,k}$ sont les variables explicatives - les informations de l'individu.

a_i sont les coefficients de régression

Les méthodes de régression linéaire sont devenues une composante essentielle de toute analyse de données pour décrire la relation entre une variable de réponse et une ou plusieurs variables indépendantes. La régression linéaire a été utilisée dans les applications de notation de crédit, le problème de deux classes peut être représenté en utilisant une variable fictive [1] mais, son inconvénient est que des fois il n'existe aucune relation linéaire entre les variables expliquées et les variables explicatives, alors ce modèle ne sera pas approprié.

Régression logistique

À part la régression linéaire, la régression logistique est l'une des techniques les plus utilisées en crédit scoring. L'une des façons les plus courantes, réussies et transparentes requises pour faire une classification binaire entre «bon» et «mauvais» sont via une fonction logistique [8]. Cette technique prédit les valeurs prises par une variable binaire à partir d'une série de variables continues et/ou catégorielles. Dans le cadre du crédit scoring, la variable binaire qu'on notera Y représente le défaut d'un client (Y= 0 pour absence de défaut et Y = 1 si présence de défaut), et les séries de variables sont les informations relatives au client dont on veut prédire sa défaillance. Le modèle logistique est la suivante :

$$P(Y = 1) = e^u/(1 + e^u)$$

Où :

$u = a_0 + a_1X_1 + ...a_kX_k$ une expression de régression linéaire

a_i coefficients de régression des variables indépendantes X_i

C'est une fonction qui prend en entrée les caractéristiques du client et génère la probabilité de défaut. [8]

L'avantage de ce modèle par rapport au modèle linéaire est que les variables indépendantes peuvent être de types différents (quantitative, qualitative, ...), et que la valeur prédite se situe entre l'intervalle 0,1. Le fait que la valeur prédite soit bornée est très importante car cela permet aux banques ou aux organismes financiers de définir une échelle de risque pour l'octroi d'un crédit.

Existant :

L'institution de microfinance malienne a également utilisé la régression logistique pour construire un modèle de crédit scoring. Le résultat obtenu pour ce modèle était à peu près le même que celui du modèle créé avec l'analyse discriminante, soit 70% de bonne prédiction c'est-à-dire environ 30% de taux d'erreur.

Steiner et Carnieri [3] dans leur étude ont également utilisé la régression logistique pour prédire la probabilité de défaut d'un emprunteur. Leur modèle a été construit sur un échantillon de 9942 clients dont 6658 classées solvables et 3284 non solvables ou en défaut. Contrairement au modèle de l'institution financière malienne précédente, celui de Steiner et Carnieri a disposé de plus de données pour la construction de leur modèle et a obtenu environ 80% de bonne prédiction soit 20% de taux d'erreur. Selon eux, ce taux d'erreur élevé est dû au manque de données comportementaux de chaque client, leur solution était

donc de diviser l'intervalle de réponse [0,1] en 3 parties : [0 ; 0.35) -> client défaut ; [0.35 ; 0.65] -> client douteux et classifié manuellement par le gestionnaire ; (0.65 ; 1] -> client sain.

Réponses	Saine	Défaut
(0.65 ; 1]	161/247	45/247
[0.35 ; 0.65]	65/247	52/247
[0 ; 0.35)1	21/247	50/247

Cette division de l'intervalle de réponse leur a permis de diminuer leur taux d'erreur à 13% et donc avoir un taux de bonne prédiction jusqu'à 87% sur 494 clients. Visant à effectuer la reconnaissance de modèle de clients sains et défaillants, le modèle de régression logistique est considéré comme l'un des meilleurs parmi les méthodes utilisées par Steiner [3] . Cette proposition de subdiviser l'intervalle a considérablement diminué leur taux d'erreur mais l'inconvénient de ce système de classement est qu'il y a encore une part de client (ceux qui ont obtenu une réponse entre [0.35 ; 0.65]) qui va être classée manuellement alors que l'objectif d'utilisation de cet outil est de classer automatiquement un emprunteur.

Arbre de décision

L'arbre de décision est une technique statistique classée parmi les apprentissages supervisés. Elles fonctionnent en partitionnant récursivement les données d'entraînements afin d'obtenir des sous-ensembles aussi purs que possibles pour une classe cible. Cette technique est non paramétrée dans l'analyse de variables dépendantes et/ou catégorielles en fonction des variables explicatives continues [1]. Chaque nœud d'un arbre est associé à un ensemble particulier d'enregistrements E qui est divisé par un test spécifique sur une caractéristique. Par exemple, une division sur l'attribut Revenu mensuel qui va être induit par le test Revenu mensuel X ; l'ensemble des enregistrements E sera donc partitionné en deux sous-ensembles qui mènent à la branche de gauche ou droite,

$$E_g = \{t \in T / t(\text{revenu mensuel}) < X\}, E_d = \{t \in T / t(\text{revenu mensuel}) >= X\}$$

Notons qu'il existe déjà plusieurs algorithmes comme le CART, CHAID, C4.5, QUEST qui permettent d'établir un arbre de décision.

Existant

Sousa M de M et Figueiredo R. S. [12] dans leur étude sur l'analyse de crédit avec le datamining ont utilisé l'arbre de décision pour construire un modèle de scoring. Ils ont établi leur modèle avec une vingtaine de critères et à l'aide de l'algorithme d'arbre de décision appelé C4.5 avec lequel ils ont obtenu 41 feuilles (c.-à-d. 41 règles de décision). Et le résultat

de classement de 321 nouveaux individus est le suivant :

Actuel/Prédit	Défaut	Saine
Défaut	121	11
Saine	8	181

Sur les 321 individus, 302 individus sont bien classés et 19 individus mal classés, soit 94,08% de bonne prédiction et 5,92% d'erreur

Techniques basées sur l'Intelligence Artificielle :

Selon Huang et al, en 2004, le problème majeur dans l'application des méthodes statistiques, est que la validité des résultats trouvés par ces techniques sont tributaires de leurs hypothèses restrictives qui sont rarement satisfaites dans la vie réelle, en l'occurrence l'hypothèse de la normalité de la distribution de chacune des variables retenues et l'hypothèse de l'indépendance entre celles-ci, ce qui peut rendre ces méthodes théoriquement invalides.

C'est ainsi que nous nous tournons vers d'autres méthodes basées sur l'intelligence Artificielle

Les réseaux de Neurones ou RNA

Ce sont des outils flexibles et non paramétriques inspirés des systèmes biologiques neuronaux. Cette technique mathématique est un programme de résolution de problèmes qui apprend à partir des procédés d'exercice d'erreurs et d'épreuves ayant recours à la logique du cerveau humain. C'est à partir de la base d'apprentissage que proviennent les variables d'entrée qui seront les neurones d'entrée afin d'obtenir un résultat révélant la présence d'un risque de crédit ou non. Ces bases d'apprentissage sont les données d'anciens clients de la banque ou de l'organisme financier possédant déjà un statut de bon ou de mauvais payeur. Et c'est à partir de chaque donnée et des résultats obtenus que le réseau se modifie lui-même et apprend peu à peu pour arriver à la résolution du problème.

Le modèle du neurone suit la procédure suivante : Pour N entrées, à chaque entrée est affecté un poids synaptique . Le neurone va commencer par en faire la somme pondérée, c'est sa fonction d'entrée, ce qui va donner son état interne. Le résultat de cette somme est transformé par une fonction de transfert (appelée aussi fonction d'activation) qui produit la sortie du neurone. Cette fonction de transfert est très importante et détermine le fonctionnement du neurone et du réseau [2].

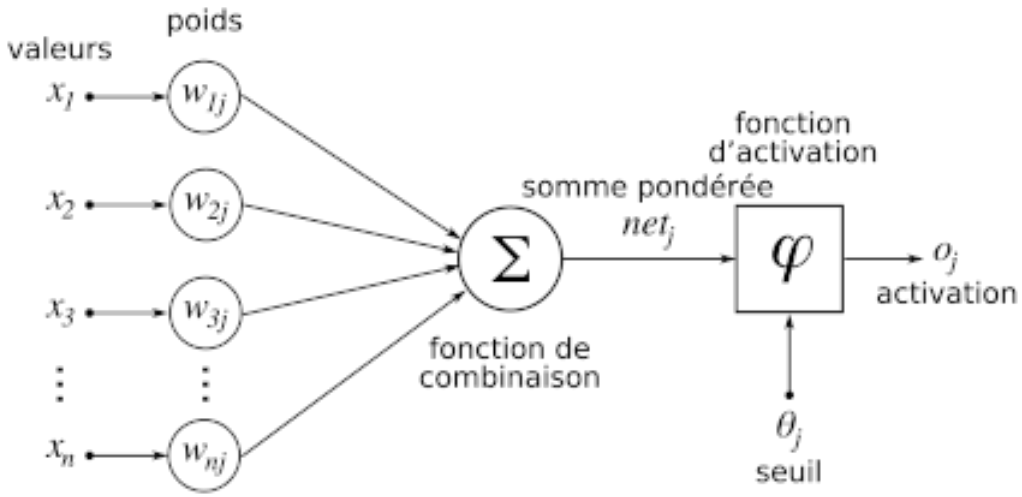


FIGURE 1.1 – Schéma du réseau de neurones

Existant

Sousa M de M et Figueiredo R. S. [12] dans leur étude sur l'analyse de crédit avec le datamining ont également utilisé l'arbre de décision pour construire un modèle de scoring. Ils ont établi leur modèle avec une vingtaine de critères et à l'aide de l'algorithme basée sur les réseaux de neurones artificiels appelé algorithme MLP. Le résultat de classement de 321 nouveaux individus est le suivant :

Actuel/Prédit	Défaut	Saine
Défaut	118	14
Saine	13	176

Sur les 321 individus, 294 individus sont bien classés et 27 individus mal classés, soit 91,59% de bonne prédiction et 8,41% d'erreur.

Algorithme génétique

La programmation génétique est l'une des plus récentes techniques utilisées dans le crédit scoring [1].Cet algorithme essaie de répliquer le processus de sélection naturelle durant lequel les gènes passent d'une génération à la suivante. Inspiré par l'évolution biologique, il offre un mécanisme de résolution de problèmes qui est efficace. La solution au problème évolue après plusieurs cycles de traitement, qui à chaque cycle, s'améliore de plus en plus.

Système immunitaire artificiel

Le système immunitaire artificiel est une intelligence artificielle inspirée du système immunitaire du corps. Il apprend et mémorise les composants et construit un modèle de

reconnaissance.

Comparaison de la performance des modèles :

On peut évaluer ces modèles et comparer leurs performances pour pouvoir décider celui qui est approprié à nos attentes, à nos objectifs. Pour mesurer la qualité d'une règle de décision, on calcule son efficacCurveité par la mesure appelée : « taux d'erreurs de classement ».

Pour ce faire, nous avons recours à quelques méthodes :

- Calculer le taux d'erreur à partir de la matrice de confusion
- Comparer la performance à l'aide de la courbe ROC
- Comparer la performance à l'aide de la courbe de Lift

Matrice de confusion

Une matrice de confusion est aussi appelée tableau de contingence et sert à l'évaluation de la qualité d'une classification.

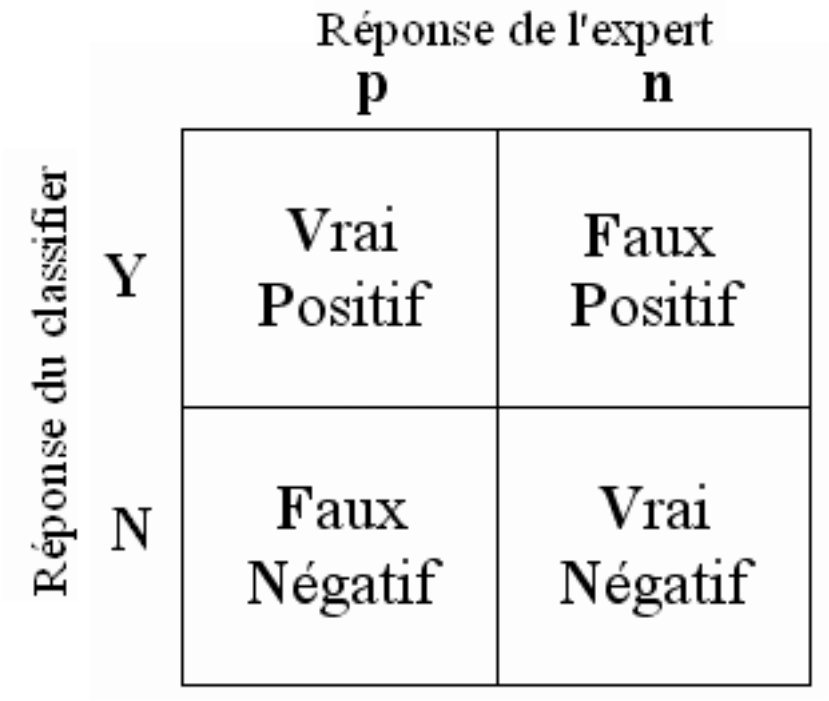


FIGURE 1.2 – Matrice de confusion

Cela permet de connaître le nombre de dossiers bien classés et mal classés par le modèle, on connaîtra alors le nombre de vrais positifs VP, vrais négatifs VN, faux positifs FP et faux négatifs FN.

À partir de ces données, on pourra calculer le taux d'erreur, qui va permettre d'estimer la probabilité de mal classer un individu.

Ce taux se calcule par la formule :

$$\text{Taux d'erreur} = \frac{FN + FP}{VP + FN + VN + FP}$$

Notons qu'on a 2 types d'erreur :

- Type 1 : Le modèle ne prédit pas le sinistre existant

$$\text{Taux d'erreur type 1} = \frac{FP}{VP + FN + VN + FP}$$

- Type 2 : Le modèle prédit un sinistre inexistant

$$\text{Taux d' erreur type 2} = \frac{FN}{VP + FN + VN + FP}$$

La comparaison des modèles se fera via ces taux d'erreurs, bien sûr le modèle qui a le taux le plus faible sera sans doute meilleur.

La seule contrainte qui se pose est que : lorsque les classes sont très déséquilibrées, la matrice de confusion et surtout le taux d'erreur donnent souvent une fausse idée de la qualité de l'apprentissage

Méthode de la courbe ROC

La courbe ROC ou en anglais Receiving Operating Characteristics est une manière d'évaluer un modèle de prédiction. Cette courbe est un outil d'évaluation et de comparaison des modèles

Atouts

La méthode est indépendante des matrices de coûts de mauvaise affectation, elle permet toujours de savoir si un modèle M_1 sera toujours meilleur qu'un modèle M_2 quelle que soit la matrice de coût.

Elle est toujours opérationnelle même dans le cas de distribution très déséquilibrée (sans les effets pervers de la matrice de confusion liés à la nécessité de réaliser une affectation)

Les résultats restent valables même si l'échantillon test n'est pas représentatif (Tirage prospectif ou tirage rétrospectif : les indications fournies restent les mêmes)

C'est un outil graphique qui permet de visualiser les performances. Un indicateur synthétique y est associé (aisément interprétable)

Procédé

Nous sommes confrontés à un problème à 2 classes. Le modèle de prédiction fournit P(Y = +/X) ou toute grandeur proportionnelle à P(Y = +/X) qui permettra de classer les

observations.

Principe de la courbe ROC :

- $P(Y = +/X) >= P(Y = -/X)$ équivaut à une règle d'affectation P (Y=+/X) >= 0.5 (seuil = 0.5)
- Cette règle d'affectation fournit une matrice de confusion MC1 et donc 2 indicateurs :
  - TPR1 = Rappel = sensibilité = VP / Positifs
  - FPR1 = 1 – Spécificité = FP / Négatifs

    Avec :

    $$\text{Sensibilité} = \frac{VP}{VP + FN} \text{ et Spécificité} = \frac{VN}{VN + FP}$$

  - Pour un autre seuil, on obtiendra MC2 et donc TPR2 et FPR2, etc...

L'idée est de faire varier le SEUIL de 0 à 1 et pour chaque cas, calculer TPR et FPR que l'on reporte dans un graphique : en abscisse le FPR et en ordonné le TPR

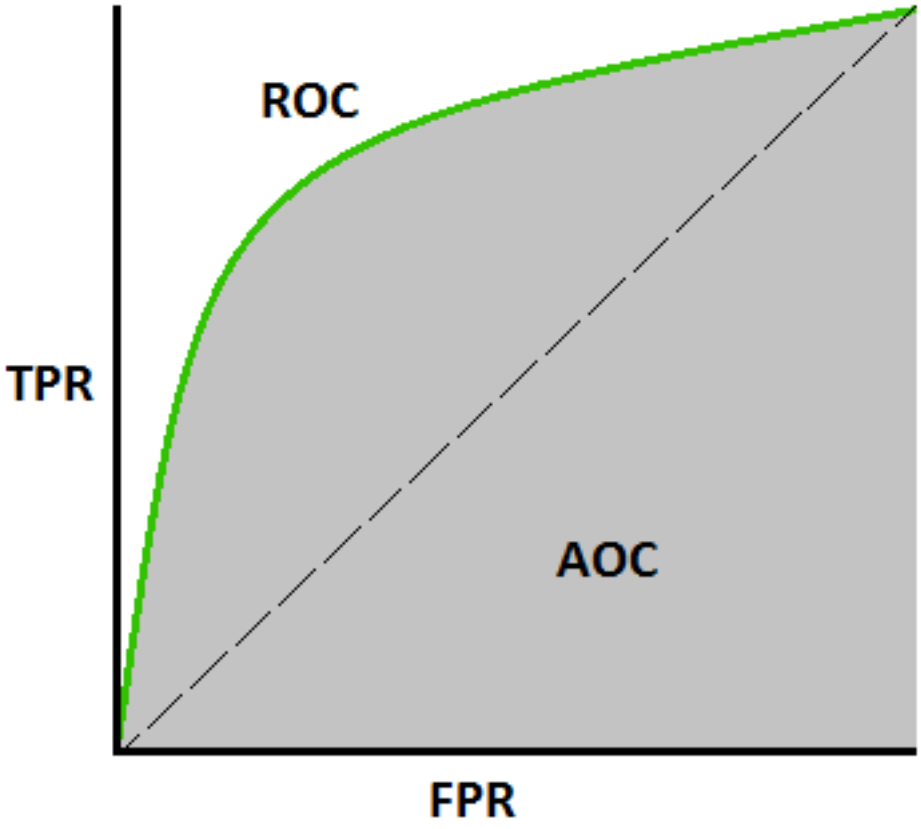


FIGURE 1.3 – Courbe ROC [11]

Interprétation

On désigne par AUC, l'aire sous la courbe formée par le calcul des TPR ET FPR. AUC indique la probabilité pour que la fonction SCORE place un positif devant un négatif (dans le meilleur des cas AUC = 1)

Ex : Si SCORE classe au hasard les individus (c.-à-d. le modèle de prédiction ne sert à rien), AUC = 0.5. Pour comparer les modèles, il suffit de visualiser le graphique obtenu et extraire celui qui est dominant afin d'obtenir le meilleur modèle.

Courbe de Lift

Cette courbe est une mesure de la performance d'un modèle prédictif ou descriptif, mesuré par rapport au modèle du choix aléatoire.

- Construire un tableau à l'aide de la base de données clientèle constituée de :
  - Variable ayant pour valeur l'appétence du client : Clients appétents (+) et clients non-appétents (-)
  - Score pour trier la base selon l'appétence du client
  - 2 critères
    - Taux de retour (proportion de (+) parmi les ciblés : rendement
    - Rappel (part des (+) retrouvés) : part de marché
- À l'aide de ce tableau, on construit un graphique dont l'abscisse est le pourcentage cumulé de la population cible (rendement) et en ordonnée, la proportion de (+) retrouvés (Rappel)

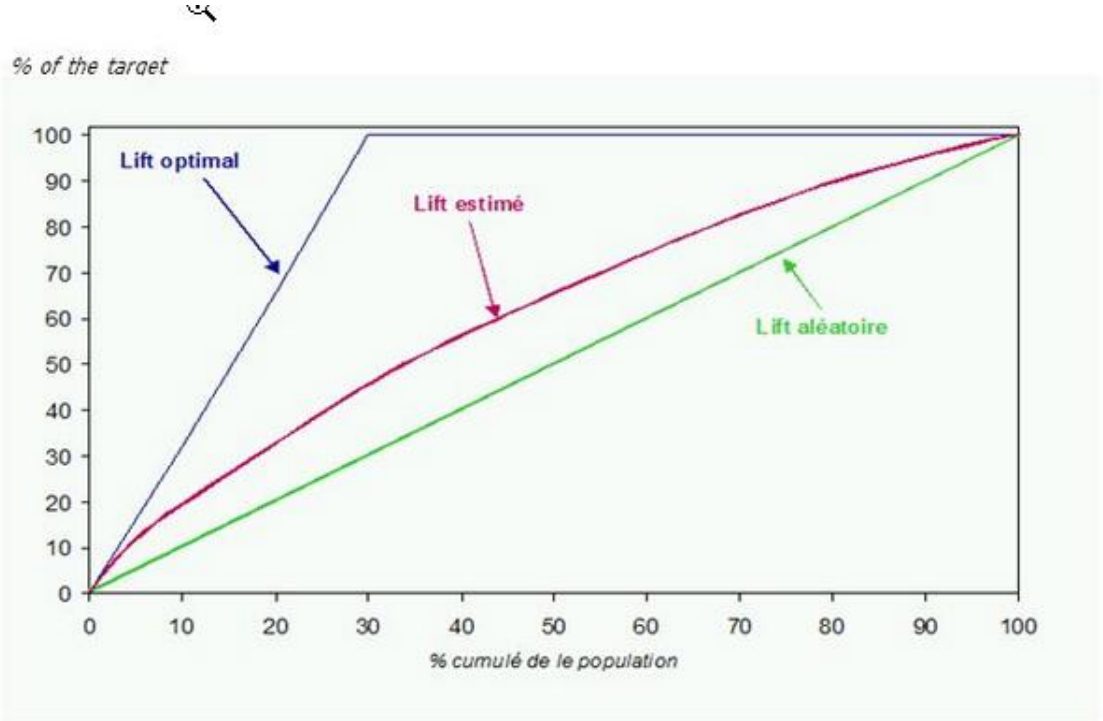


FIGURE 1.4 – Courbe de Lift [9]

Interprétation

On appelle coefficient χ le rapport (aire entre la courbe de lift et la diagonale) / (aire entre

la courbe de lift et la courbe idéale). On a aussi $\chi = 2AUC - 1$
Pour comparer les modèles, il suffit de comparer leur coefficient χ et le modèle possédant un χ élevé sera sans doute le modèle le plus performant parmi les modèles comparés.

Avantages du crédit scoring :

Le crédit scoring est un outil avantageux pour l'organisme de crédit, car il lui permet de limiter les risques en veillant à ce que l'institution financière puisse classer les clients demandeurs (en BON ou MAUVAIS payeur). Il facilite et accélère la prise de décision d'octroi de crédits.

En interne, son utilisation apporte un avantage de charge de travail dans la mesure où l'exploitant et le processus de décision sont considérablement accélérés. Ainsi, cela permet une appréciation rapide et relativement fiable. En outre, il diminue le taux d'impayés.

Il facilite la délégation des décisions : le personnel peut ne pas être capable de mener à terme le processus traditionnel de décision, mais ce personnel peut facilement utiliser la méthode de scoring.

Il peut également présenter un avantage commercial du fait que le client reçoit une réponse en quelques minutes et permet aux prêteurs d'adapter les efforts marketing aux besoins spécifiques de marketing sur les différents segments de marché.

Inconvénients du crédit scoring :

Bien que l'utilisation de cette méthode présente plusieurs avantages, nous notons cependant quelques inconvénients.

Le crédit scoring tend à déshumaniser le processus de la demande de prêt. Avec des formules mathématiques complexes, prenant en compte des données, les dossiers sont refusés ou acceptés, sans une possibilité de défendre le projet ou convaincre le banquier.

Le crédit scoring handicape le candidat qui peut se voir refuser sa demande de prêt sur la base de ses données standards, qui peuvent ne pas refléter la situation réelle du candidat.

Il y a une absence de garantie car le scoring est basé sur une probabilité et non une certitude.

En se basant essentiellement sur cet outil, l'organisme de crédit peut se voir rater des opportunités de vente, ou même perdre des clients, ce qui pourrait affecter les interactions entre prêteurs et emprunteurs ainsi que l'accessibilité et la tarification du crédit.

Le crédit scoring peut réduire l'accès au crédit pour ceux qui n'ont pas d'historique de crédit.

1.3 Présentation de la BNI Madagascar et de ses données

La « Banky Nasionalin'ny Indostria » où en abrégé BNI est une banque existant à Madagascar depuis 1919 sous la nomination d'Institut foncier de Madagascar. Cette banque fournit différents services dont le compte épargne, les crédits à la consommation et immobiliers, et enfin une assurance et prévoyance. Les données fournies par la BNI sont ses propres données historiques. Ces données datant des années 2011, 2012, 2013, 2014 et 2017. Les données comprennent les informations sur les clients auxquels la BNI Madagascar a fourni du crédit et leur capacité à rendre ce crédit. Le tableau suivant rend compte ainsi sur les données de chaque dossier, c'est-à-dire les noms des champs et leurs descriptions.

NOM DES CHAMPS	DESCRIPTIONS	Type de variable
RADICAL	RADICAL DU CLIENT	Texte
CLE	CLE DU COMPTE DU CLIENT	Texte
MONTANT ACCORDE	MONTANT CAPITAL PRETE	Nombre
DUREE	DUREE DU PRÊT en mois	Nombre
MENSUALITE	SOMME MONTANT ASSURANCE - MONTANT INTERET ET AMORTISSEMENT CAPITAL	Nombre
TAUX D'ENDETTEMENT	TAUX D'ENDETTEMENT REEL	Nombre
ANCIENNETE BNI (en mois)	ANCIENNETE DE LA RELATION avec BNI en mois	Nombre
NOMBRE SALAIRE VIRE	NOMBRE DE VIRT SALAIRE DU CLIENT	Nombre
CATEGORIE EMPLOYEUR	CATEGORIE OU QUALITE DE L'EMPLOYEUR	Texte
COULEUR SCORE	SCORE DEMANDEUR/ Cf. également ISBA CPCR/QICOUL	Texte
NATURE CAP	NATURE DU CREDIT AUX PARTICULIERS	Texte
AGE	AGE DE L'EMPRUNTEUR	Nombre
STATUT MARITAL	STATUT MARITAL	Texte
CONVENTIONNE	CONVENTION PARTENARIAT (OUI/NON)	Texte
ANCIENNETE EMPLOI	ANCIENNETE DE L'EMPRUNTEUR en mois	Nombre
TYPE DE CONTRAT EMPRUNTEUR	NATURE DE CONTRAT DE L'EMPRUNTEUR	Texte
REVENU MENSUEL	REVENUS DU MENAGE PRIS EN CPTE	Nombre
LOGEMENT	TYPE DE LOGEMENT DE L'EMPRUNTEUR	Texte
CODE_DOUTEUX	DOSSIER AYANT REMBOURSE LE PRET (0 oui et 1 NON)	Nombre

TABLE 1.1 – Variables des données de la BNI

1.4 Analyse des existants de la BNI

Depuis 2009, la BNI utilise un outil d'analyse délivrant un score permettant de déterminer le circuit d'octroi d'un crédit aux particuliers (CAP). Le score délivré par cet outil est sous forme de couleur. Chaque couleur est associée à un niveau de risque :

- Vert : Dossier sain, sans risque
- Orange avec/sans Garantie : Dossier fiable mais nécessitant des études
- Rouge : Dossier à risque
- Noir : synonyme de dossier irrecevable

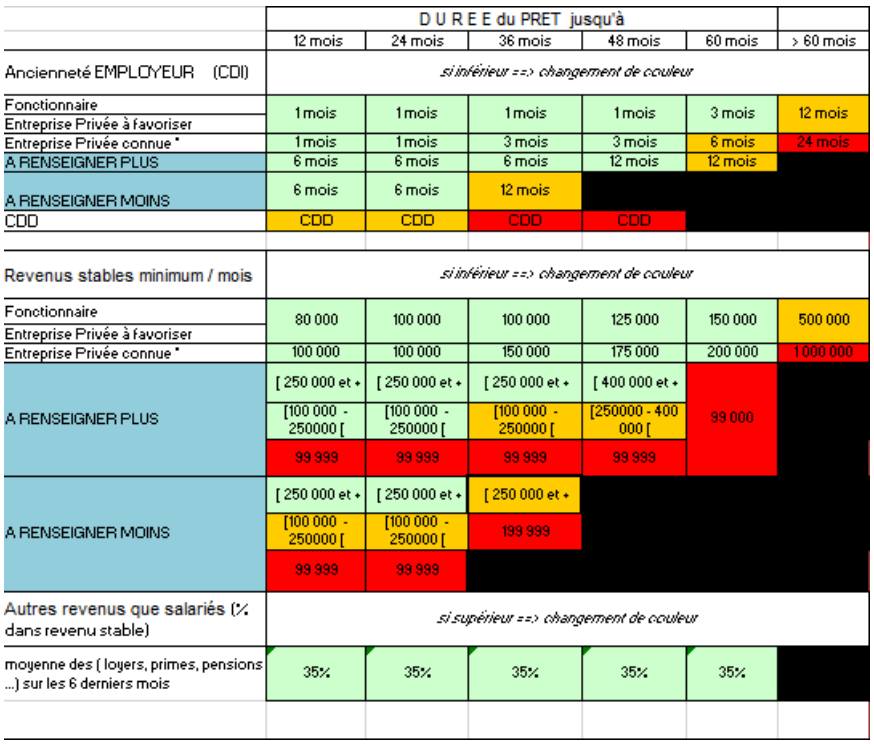
Cet outil se base sur onze (11) critères qui sont :

- L'ancienneté de l'emprunteur à son emploi (en mois)
- Le revenu stable mensuel de l'emprunteur
- La catégorie de l'employeur (Fonction publique, Entreprise privée à favoriser, Entreprise privée connue, Entreprise privée à renseigner)
- L'âge de l'emprunteur
- Le montant du crédit débloqué (en Ariary)
- Le taux d'endettement de l'emprunteur
- L'ancienneté à la BNI (en mois)
- Le fonctionnement du compte (nombre d'impayés sur les 12 derniers mois)
- Le remboursement anticipé

À partir de ces critères, le score est établi à l'aide de la grille de score dans la figure 1.5

L'ordre des couleurs étant lié aux risques est : Vert, Orange, Rouge, Noire.
À chaque condition non vérifiée dans la grille, la dégradation de couleur suit cet ordre ;

- Si une condition de couleur "Vert" n'est pas vérifiée, on change sa couleur par celle associée à cette condition et sera donc "Orange"
- Si une condition de couleur "Orange" n'est pas vérifiée, on change sa couleur par celle associée à cette condition et sera donc "Rouge"
- Si une condition de couleur "Rouge" n'est pas vérifiée, on change sa couleur par celle associée à cette condition et sera donc "Noir"



	DUREE du PRET jusqu'à					
	12 mois	24 mois	36 mois	48 mois	60 mois	> 60 mois
Ancienneté EMPLOYEUR (CDI)	*si inférieur ==> changement de couleur*					
Fonctionnaire Entreprise Privée à favoriser	1 mois	1 mois	1 mois	1 mois	3 mois	12 mois
Entreprise Privée connue *	1 mois	1 mois	3 mois	3 mois	6 mois	24 mois
A RENSEIGNER PLUS	6 mois	6 mois	6 mois	12 mois	12 mois	
A RENSEIGNER MOINS	6 mois	6 mois	12 mois			
CDD	CDD	CDD	CDD	CDD		
Revenus stables minimum / mois	*si inférieur ==> changement de couleur*					
Fonctionnaire Entreprise Privée à favoriser	80 000	100 000	100 000	125 000	150 000	500 000
Entreprise Privée connue *	100 000	100 000	150 000	175 000	200 000	1 000 000
A RENSEIGNER PLUS	[250 000 et +	[250 000 et +	[250 000 et +	[400 000 et +	99 000	
	[100 000 - 250000 [	[100 000 - 250000 [	[100 000 - 250000 [	[250000 - 400 000 [		
	99 999	99 999	99 999	99 999		
A RENSEIGNER MOINS	[250 000 et +	[250 000 et +	[250 000 et +			
	[100 000 - 250000 [	[100 000 - 250000 [	199 999			
	99 999	99 999				
Autres revenus que salariés (% dans revenu stable)	*si supérieur ==> changement de couleur*					
moyenne des (loyers, primes, pensions ...) sur les 6 derniers mois	35%	35%	35%	35%	35%	

FIGURE 1.5 – Grille de score

À part les dossiers irrecevables (score associé Noir), tous les dossiers ont été acceptés.
Le Backtesting de cet outil de score a permis d'établir les données suivantes :

- Cas de 2014 :
  - Nombre de dossiers acceptés : 6687
  - Nombre de dossiers sinistrés : 214
  - Nombre de dossiers sains : 6473
  - VERT : 4651 (dont 125 dossiers sinistrés)
  - ORANGE avec Garantie : 150 (dont 5 dossiers sinistrés)
  - ORANGE sans Garantie : 1612 (dont 54 dossiers sinistrés)
  - ROUGE : 274 (dont 30 dossiers sinistrés)

Ces données montrent que parmi les dossiers de score VERT, des dossiers qui devraient être sans risque, il y a encore bon nombre de dossiers sinistrés soit 125 sur 4651. D'ailleurs, parmi les dossiers de score ROUGE qui ont été acceptés, le nombre de dossiers sinistrés est faible (30 sur 274 dossiers) cependant ces dossiers étaient considérés comme des dossiers à haut risque de défaut.

Problèmes de la grille de score

- Selon les résultats du Backtesting de l'année 2014, ce score est peu fiable

- Il y a des critères pénalisant les non clients : l'ancienneté à la BNI, fonctionnement du compte.
- Des critères de discrimination ne sont pas adéquats. Par exemple, le montant du crédit ne devrait pas être considéré dans le calcul de score.

1.5 Objectifs du projet

Cette étude va se focaliser sur les quatre points suivants :

- La refonte du système de notation de la BNI pour prédire si un client est un bon ou un mauvais payeur.
- La création d'un système de notation à base de score (par exemple : de 0 à 100, mais des couleurs peuvent être attribuées selon des tranches de points choisis par la BNI)
- La création d'un système de notation plus fiable qui implique que l'octroi de crédit basé sur ce nouveau système de notation doit être :
  - Vert : Accord immédiat
  - Orange : fiable mais nécessitant des études
  - Rouge : non fiable
- L'analyse des critères pertinents pour le calcul du nouveau score (tout en considérant le cas des clients et des non-clients de la BNI) en identifiant :
  - les critères inutiles ou pénalisants pour le calcul de score
  - les critères pertinents (anciens et/ou nouveaux critères)

Chapitre 2

Le crédit scoring de la BNI

2.1 Méthodologie Générale

Cette partie du document fait l'objet des méthodologies suivies pendants trois mois, afin de modéliser les données de la BNI. Or, les dites données doivent tout d'abord être soumises à certaines analyses. Dans un premier temps, nous allons explorer et préparer les données dans le but de bien les évaluer avant qu'elles ne soient modélisées. Cette partie du travail concerne les quatre (4) premières étapes figurant dans le document. Ensuite, la modélisation pourra se faire et nous consacrerons trois (3) étapes à cette seconde partie. Ainsi, notre méthode suivra sept (7) grandes étapes qui seront détaillées par la suite.

2.1.1 Prétraitement commun des données

Le prétraitement des données se divise en quatre(4) grandes étapes ;

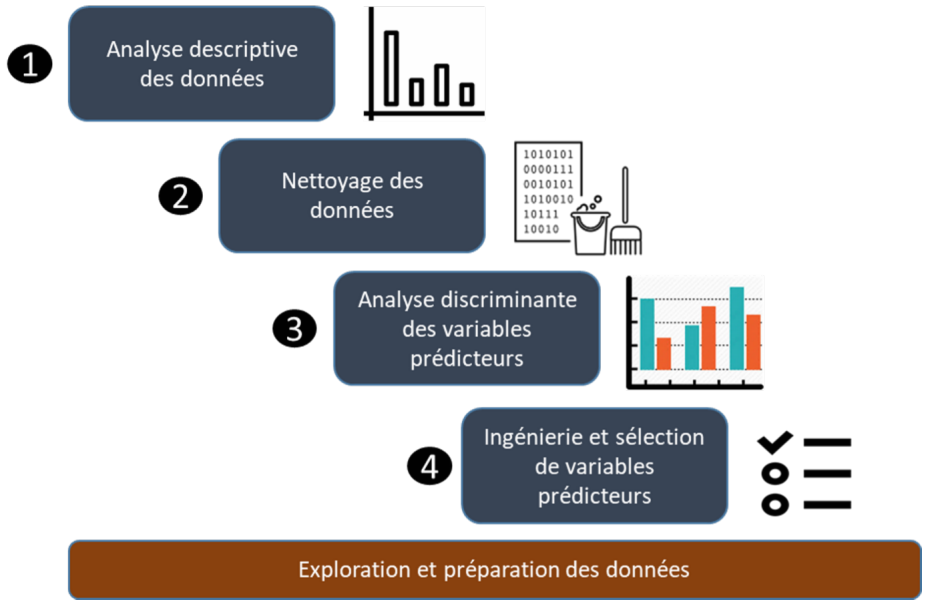


FIGURE 2.1 – Prétraitement des données

2.1.1.1 Exploration et préparation des données

Cette étape vise en premier lieu, à étudier d'une manière univariée les tendances dans le jeu de données, ceci afin d'identifier les valeurs aberrantes qui pourraient potentiellement nuire à la modélisation. En second lieu, nous essayerons de montrer le pouvoir discriminant de chaque variable prédictrice sur la variable cible afin d'orienter nos choix de variables prédictrices pour la construction du modèle optimal (modèle maximisant les critères de performance choisis).

2.1.1.2 Analyse des données

Pour une bonne maîtrise du périmètre d'étude, il nous est nécessaire de sortir la tendance de la population sur laquelle nous effectuons la modélisation. Les indicateurs utilisés sont : la distribution, la moyenne et l'écart type, minimum et maximum.

2.1.1.3 Nettoyage des données

À partir de l'analyse descriptive, nous pourrions rencontrer des cas atypiques : variables prédictrices non renseignées, valeurs trop éloignées de la moyenne (règles des 3 sigmas), modalité d'une variable n'apparaissant que pour un unique individu etc. Il est important de traiter ces cas au risque de nuire à l'apprentissage. Les modèles géométriques étant basés sur les notions de distance, ces derniers sont particulièrement sensibles aux valeurs extrêmes. Si ces cas ne représentent qu'une infime partie des données, nous pouvons les exclure sans risque de perdre de l'information.

2.1.1.4 Analyse discriminante des variables prédictrices

Le but de la modélisation est de généraliser, par l'intermédiaire de méthodes statistiques, géométriques et d'apprentissage automatique, les critères d'un client sain et ceux d'un client sinistre. Il est nécessaire de comprendre en amont si le fait qu'un individu présente une certaine valeur d'une variable prédictrice continue (ou une certaine modalité d'une variable catégorielle) peut accentuer la chance qu'il soit sinistre ou pas. Pour cela, plusieurs techniques seront utilisées : visualisation (histogramme et boîte à moustache), AUCROC, test de χ^2. De ces analyses et suivant ces critères, nous pourrons avancer et proposer les variables que l'on pourra qualifier d'« importantes », qui auront très certainement un impact non négligeable sur le modèle. D'une manière réciproque, les variables qui ne présentent aucune discrimination suivant ces analyses pourraient alourdir le modèle (interprétation, temps d'entraînement).

2.1.1.5 Ingénierie et sélection de variables prédictrices

De l'étude précédente, nous pouvons établir la liste des variables prédictrices les plus discriminantes pour passer à la modélisation. Nous proposerons aussi d'autres manières de représenter les variables par des transformations de type : discrétisation, normalisation, transformation par logarithme, changement d'unité, pour accommoder les modèles (certains modèles s'améliorent pour certaines transformations) et faciliter l'interprétation de certaines variables (âge en classe d'âge, ancienneté en années /semestre).

2.1.2 Modélisation

Après le prétraitement, on procède par la modélisation :

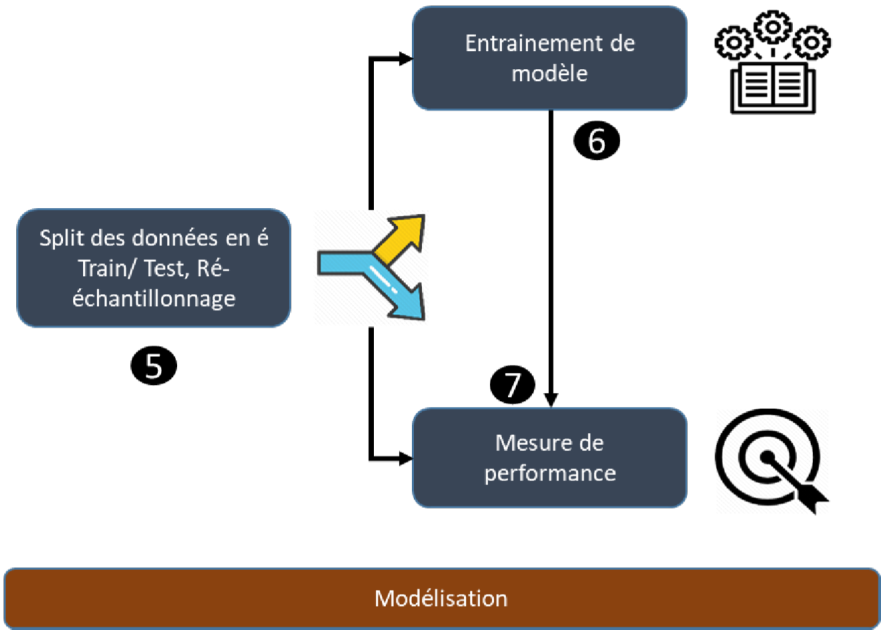


FIGURE 2.2 – Modélisation

2.1.2.1 Partage des données en partie entraînement et test

Pour pouvoir apprécier la performance des modèles construits, il est important de segmenter aléatoirement les données en deux catégories :

- Données d'entraînement : données qui serviront à estimer les coefficients et paramètres optimaux du modèle.
- Donnée de test : données qui serviront à estimer les performances du modèle

NB : ce split est important parce que les mesures de performance effectuées sur les données d'apprentissages sont trop optimistes et constituent un biais d'optimisme (on ne pourra avoir qu'une bonne performance si on teste un modèle sur les mêmes données qui ont servi à le construire).

2.1.2.2 Construction du modèle

Au cours de cette étape, nous allons estimer les paramètres du modèle à partir des données « train ». L'algorithme du modèle « apprend » à sa façon les critères d'un client « sinistre » et celui d'un client « sain ». NB : Il peut arriver que l'algorithme « apprenne » davantage sur les critères d'une classe au détriment d'une autre. Ce phénomène est causé par le non-balancement des classes dans la population « train ». Comme illustration, si notre population d'apprentissages est constituée de 98% de « sain » et 2% de « sinistre », l'algorithme aura plus tendance à « apprendre » les critères de la classe majoritaire « sain » et ignorer l'autre classe. Pour pallier ce problème, il est pratique de procéder à un ré-échantillonnage : diminuer la classe surpondérée ou augmenter de manière synthétique la classe sous-pondérée ou appliquer les deux.

2.1.3 Les modèles utilisés

2.1.3.1 Le Gradient Boosting

Le Gradient Boosting est l'une des idées d'apprentissage les plus puissantes introduites il y a vingt ans. Il a été conçu à l'origine pour les problèmes de classification. Le gradient Boosting classifieur combine plusieurs arbres de décision avec « faible » puissance de prédiction (qui prédit des petits scores) pour construire un modèle plus généraliste. C'est donc un modèle ensembliste.

2.1.3.2 Le Réseau de Neurones

Le Neural Network ou réseau de neurones est un algorithme d'apprentissage automatique qui a pour particularité de simuler le fonctionnement des neurones humains. C'est un des plus puissants classifieurs d'aujourd'hui. Il est modélisé mathématiquement par un réseau dans un graphe, plus ou moins complexe, hiérarchique sous forme de couches dont les nœuds élémentaires sont appelés neurones. Il est composé généralement de 3 types de couche : la couche d'entrée (input layer), la couche cachée (hidden layer) et la couche de sortie (output layer). La couche d'entrée est composée de neurones qui correspondent aux caractéristiques des données d'entrée représentées par une grille multidimensionnelle (par exemple la matrice de pixels d'une image ou la forme vectorielle d'une donnée). La couche de sortie représente les résultats de la tâche assignée au réseau. Par exemple pour une classification de 1000 classes, les 1000 neurones de la couche de sortie représentent la probabilité ou le score pour chaque classe. Les couches cachées sont les couches intermédiaires entre l'entrée et la sortie. L'ensemble du réseau ainsi formé est généralement vu comme une boîte noire.

2.1.3.3 La Régression Logistique

La régression logistique est un modèle permettant de mesurer l'association entre la survenue d'un évènement (variable expliquée) et les facteurs susceptibles de l'influencer (variables explicatives), dans notre cas, la variable expliquée Y étant le risque de crédit. Nous allons utiliser le modèle « logit » et chercher à estimer la probabilité P [Y=1], c'est-à-dire la probabilité de risque de crédit à partir des variables explicatives. $P[Y = 1] = eu/(1 + eu)$ Avec $u = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n$, X_i variables explicatives (ou facteurs susceptibles d'influencer Y) et a_i leur coefficient

2.1.3.4 Le LDA (Analyse disciminante linéaire)

L'analyse Discriminante est une technique statistique visant à décrire, expliquer et prédire l'appartenance d'un individu à des groupes prédéfinis d'un ensemble d'observations à partir d'une série de variables prédictrices. Cette technique peut être également interprétée de façon géométrique. Dans notre cas, l'objectif est de prédire le groupe d'appartenance (soit 0 –pour sains ou 1 pour douteux) d'un dossier à partir des valeurs prises par des variables prédictrices. Pour cela, on cherche à estimer la fonction discriminante linéaire $d(X) = c_0 + c_1X_1 + c_2X_2 + \ldots. + c_JX_J$
Où : les X_i sont les variables prédictrices c_i les coefficients des variables prédictrices.
Dans une Analyse Discriminante Linéaire où la variable à prédire Y prend deux modalités $(Y = 0, 1)$, la règle d'affectation est comme suit :

- Si $d(X) > 0$ alors $Y = 1$ et le dossier sera classé Sinistre
- Sinon $Y = 0$ et le dossier sera classé Sain

2.1.4 Mesures de performance des modèles

La sortie des prédictions n'est pas contrairement à ce que l'on pense, une réponse claire (sain ou sinistré) ; mais une probabilité de défaut. Dans le cas de la BNI, c'est la probabilité que le dossier considéré soit douteux, c'est-à-dire mauvais payeur. Pour trancher cette décision, nous introduisons le seuil de décision : si la probabilité est supérieure au seuil, alors le dossier est positif (douteux) ; par contre, si la probabilité est inférieure au seuil, alors le dossier est négatif (sain). Cependant, il est difficile de comparer deux modèles différents du fait que le seuil de décision varie d'un modèle à un autre. Ainsi, pour comparer les modèles nous nous baserons sur les courbes ROC et score AUC des modèles avant de choisir ce seuil de décision.

2.1.4.1 La courbe ROC et le score AUC

La courbe ROC suit le mieux pour comparer les modèles, sachant que la distribution est fortement déséquilibrée et que le seuil de décision varie d'un modèle à un autre. Nous pourrons ainsi choisir le modèle convenable à la BNI à partir du score AUC de chaque modèle.

2.1.4.2 La matrice de contingence

Cette méthode d'évaluation complète la précédente dans la mesure où elle permet de prendre en compte les erreurs du modèle et ainsi, on peut à la fois avoir une prévision du taux de remboursement et permet à la banque de prendre des mesures adéquates. On choisira donc à partir des valeurs de cette matrice le choix de décision auquel on désignera un dossier sinistre ou non.

2.1.5 Bilan

2.1.5.1 Le modèle choisi

Le score AUC des modèles nous permet d'avoir la discrimination de chaque modèle pour chaque jeu de données annuelles. Ainsi, nous pouvons choisir le modèle convenant. Le tableau suivant rend compte du score AUC de chaque modèle pour chaque année. De ce fait, le gradient Boosting a le plus de score AUC dans tout le tableau. L'algorithme Neural Network a le plus de score AUC pour les données les plus récentes. En ce qui concerne l'algorithme LDA, c'est l'algorithme qui a le moins de score AUC sur tous les modèles et sur toutes les années.

Année	AUC Gradient Boosting	AUC Réseau de neurone	AUC Régression logistique	AUC LDA
2011	0.797	0.7432	0.7161	0.6473
2012	0.76.9	0.7174	0.6860	0.6338
2013	0.693	0.6087	0.6038	0.5772
2014	0.801	0.7274	0.7212	0.6389
2017	0.72	0.7340	0.7215	0.6221

TABLE 2.1 – Score AUC des modèles par année

Le modèle choisi par la BNI est détaillé ci-dessous, les détails des autres modèles non retenus peuvent être consultés dans l'annexe.

Le modèle retenu utilise l'algorithme gradient Boosting Classifieur doté de

- 2000 arbres dont la hauteur maximum d'un arbre est trois(3)
- deux pas d'apprentissage de 0.01.

Les variables du modèle retenu

Les variables de données fournies par la BNI ne sont pas assez discriminantes pour avoir une bonne classification. On procède alors par une phase appelée Feature engineering. Cette phase consiste à transformer, et ou à synthétiser des variables à partir de combinaisons des variables existantes :

- art_mensualite : Sachant que les données de 2011 et 2012 ne contiennent pas la variable, on a créé une variable artificielle pour la redevance mensuelle.
- Salaire vivable : C'est le reste du salaire de l'emprunteur par mois.
- Argent journalier : C'est l'argent que la personne pourra dépenser journalièrement. Donc, c'est la variable défini par $\frac{\text{salaire vivable}}{30}$.
- Salaire des dossiers du même âge : C'est le salaire moyen des autres personnes de mêmes tranchées d'âge que l'emprunteur.
- Salaire des dossiers de la même catégorie employeur : C'est le salaire moyen des autres personnes de même catégorie d'emploi que l'emprunteur.
- Endettement des dossiers de même durée : C'est le taux d'endettement moyen des personnes de même tranche d'âge que l'emprunteur.
- Age fin : C'est l'âge de la personne à la fin de son prêt.

Désormais, le modèle comporte de nombreuses variables, pour réduire au plus la complexité du modèle, on doit réduire le nombre de variables, tout en essayant de garder la performance du modèle. Pour ce faire, on choisit les variables par son importance au modèle, illustré dans la figure 2.3

Les variables suivantes ont donc été retenues :

- L'age à la fin de prêts ('age_end'),
- Le salaire vivable ('salaire_vivable'),
- L'argent journalier ('daily_money'),
- Le salaire moyen de la même tranche d'âge ('age_catégorie_salaire'),
- Le nombre de salaires viré ('nb_salaires'),
- Le statut marital ('statut_marital'),

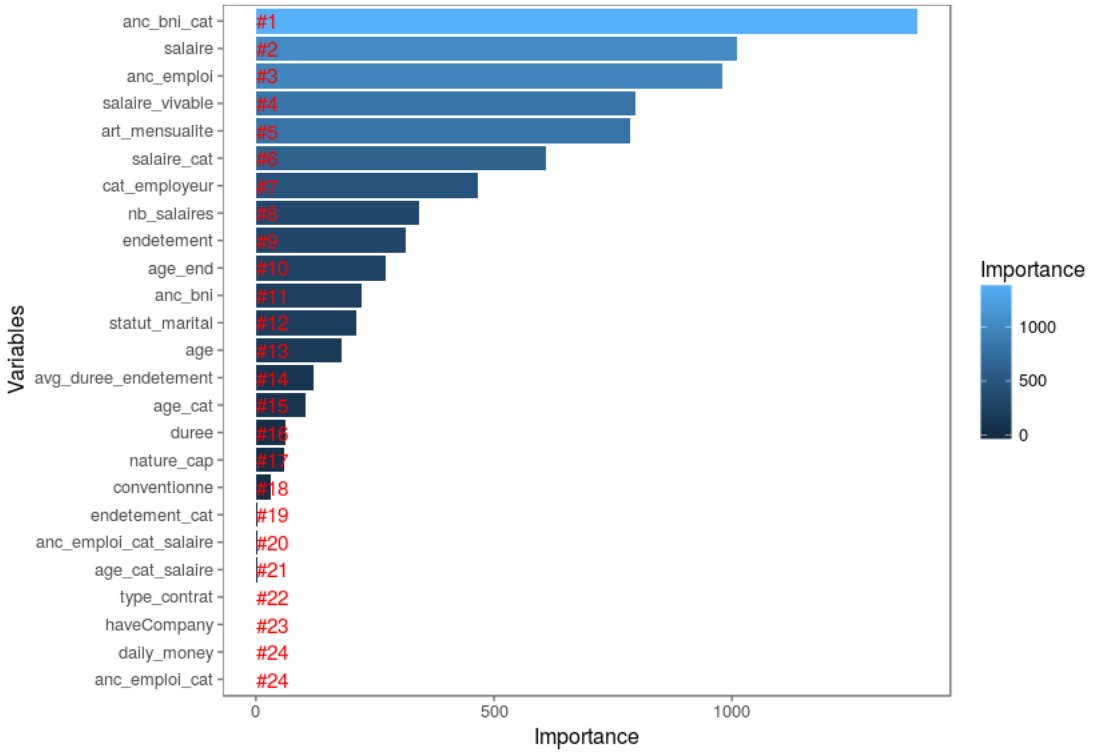


FIGURE 2.3 – Feature importances des variables

- L'endettement moyen pour les prêts de même durés ('avg_duree_endetement'),
- L' ancienneté BNI discrétisée ('anc_bni_cat'),
- L' ancienneté d'emploi discrétisée ('anc_emploi_cat'),
- Le salaire discrétisé ('salaire cat').

2.1.5.2 Répartition des scores et choix du seuil de décision

À terme de l'apprentissage, il est important de bien choisir un bon seuil, afin d'avoir une bonne balance entre faux positifs et faux négatifs. La façon simple est la visualisation de la distribution des scores des deux classes.

Répartition des scores

Nous construisons deux figures ci-dessous, à partir de la distribution des prédictions sur la donnée de test et de validation, pour voir la discrimination des deux catégories et avoir une approximation du seuil de décision :

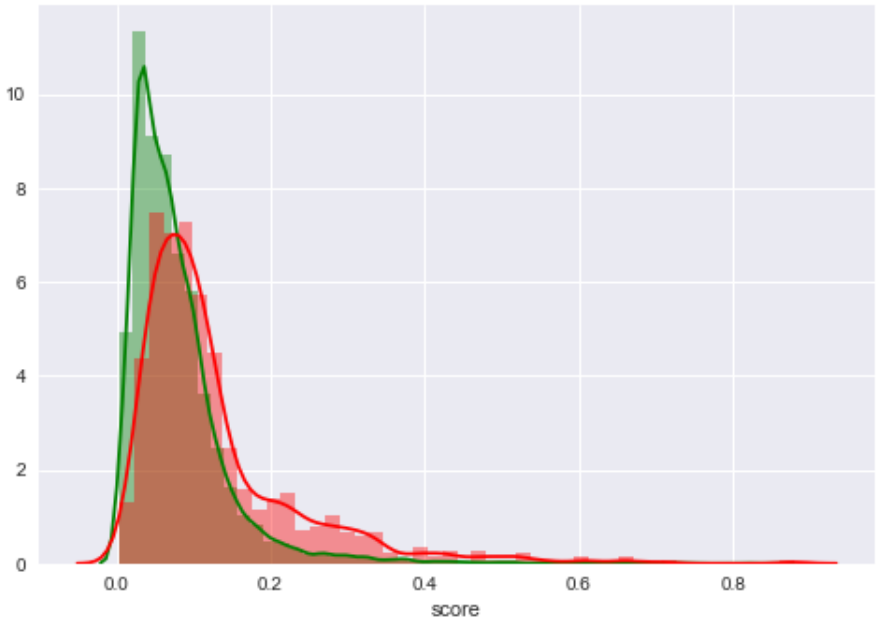


FIGURE 2.4 – Distribution des scores des données de test

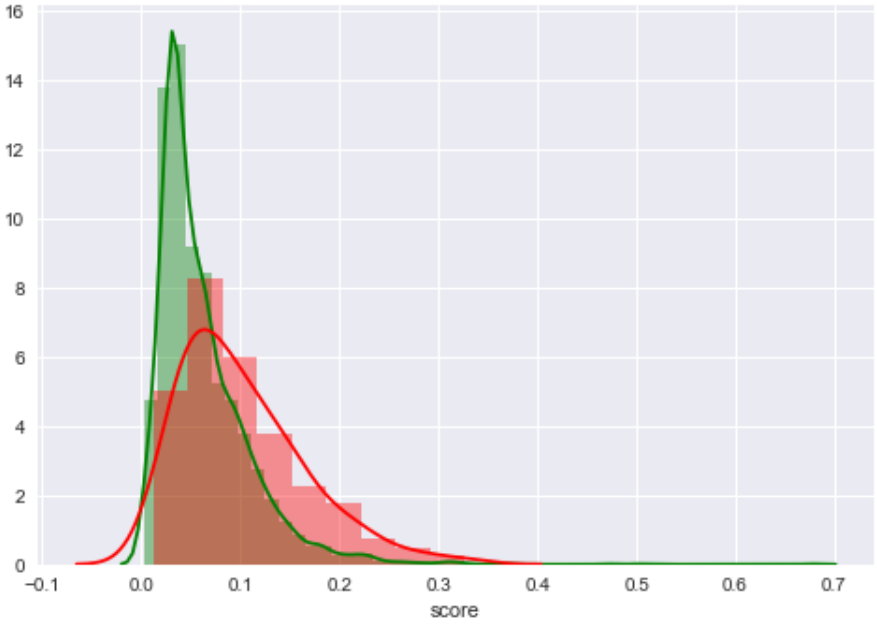


FIGURE 2.5 – Distribution des scores des données de validation

Seuil de décision

Le seuil de décision de 0.14 a été choisi par la BNI parcequ'en choisissant ce seuil, la BNI prévoyait de capturer 94.74% des dossiers sains. Néanmoins, En refusons 10% des dossiers totaux, nous ne refusons que 36.68% des dossiers sinitrés. La table de contingence reliée à ce seuil est dans le tableau ci-dessous.

		Classification prédite	
		Douteux	Sain
Classification réel	Douteux	824	2469
	Sain	1422	25642

TABLE 2.2 – Matrice contingence du Modèle Gradient Boosting

2.1.5.3 Incorporation du modèle scoring à l'organisation de la BNI

La communication du modèle avec la structure informatique de la BNI se fait par web service. En envoyant une requête http get, le module répond avec un fichier Json contenant les informations suivantes :

- les informations sur le dossier,
- le score basé sur la capacité de remboursement du projet. Ce score est obtenu par la formule :

$$Score = (1 - P(X)) * 100$$

  où P(X) est la probabilité de défaut prédit par le modèle
- la couleur correspondant à ce score :
  - vert : si le score est supérieur au score seuil
  - rouge : si le score est inférieur au score seuil

Chapitre 3

Proposition d'amélioration

Les données de la BNI ne comportent que peu d'informations sur le dossier du client et manque relativement de données antérieures sur les prêts et transactions du client. Nous n'avons pu tirer qu'une discrimination limitée de la variable cible. Nous allons voir dans ce chapitre un modèle de crédit scoring dont les données sont riches d'indicateurs dont les transactions bancaires, prêts antécédents et même de données personnelles intéressantes. Nous allons ainsi tirer de ce modèle quelques propositions d'amélioration que nous pourrons apporter au modèle de Crédit Scoring de la BNI.

3.1 Le Crédit scoring de Home Loan Default

3.1.1 Présentation de Home Loan Default.

Home Loan Default est une entreprise multinationale qui offre des prêts en vue d'achats immobiliers. Ses activités se rependent dans plus de dix pays différents. La prédiction de remboursement des clients un problème courant de l'organisation. Ainsi, l'organisation a proposé un concours dans le site web Kaggle en partageant une partie de ses données d'inscription de ses clients. L'organisation Home Loan default a donné trois mois aux participants pour tirer au mieux les particularités pour discerner les clients pouvant rembourser leurs prêts et ceux qui ne le peuvent pas.

3.1.2 Les données de Home loan Default

Le jeu de données de Home Loan comporte 8 tables [5], dont

- Application_test.csv : donnée principale pour évaluer le modéle par Kaggle
- Application_train.csv : donnée principale pour entrainer le modèle.

- Bureau.csv : donnée des prêts antécédents des clients dans d'autres organisations
- Bureau_balance.csv : donnée mensuel sur les prêts antécédents du clients dans d'autres organisations
- Credit_card_balance.csv : donnée mensuel du client sur sa carte de crédit.
- Installements_payments.csv : donnée sur les versement mensuel antécédent du client chez Home Loan
- POSH_CASH_balance.csv : donnée mensule sur les préts mensuel du clients en espèces.
- Previous_application.csv : donnée sur les prêts antécédents du client chez Home Loan.

Les tables citées précédemment sont accompagnées d'une table :

- HomeCredit_columns_description.csv : table contenant les informations et descriptions des champs de chacune des tables citées précédentes.

La figure suivante note les relations de chacun de ces tables avec la table application_train.csv :

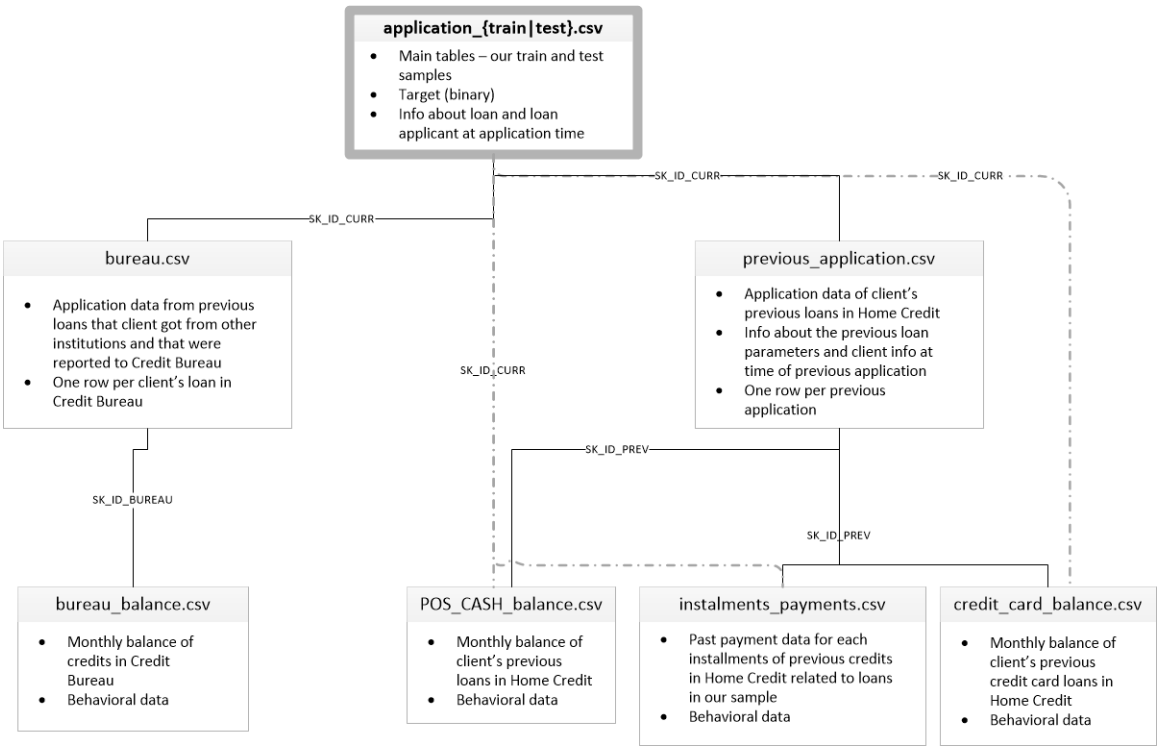


FIGURE 3.1 – Schéma relationnel des données de Home Loan [5]

3.1.3 Le meilleur modèle de crédit Scoring de Home Loan

Le meilleur modèle lors du concours Kaggle a obtenu un score AUC de 0.80570. Ce modèle est le résultat d'ensemble de modèles différents (Neural Network, XGBoost, LightGBM et un modèle linéaire Hill Climber) en utilisant la technique appelée Stacking.

3.1.4 Notre modèle de Crédit Scoring

Notre travail étant effectué en groupe, nous allons rendre compte sur la partie modélisation de la compétition.

3.1.4.1 Modèle Utilisé

Le modèle utilisé est un « gradient Boosting décision Tree », nommé Light gradient Boosting. Le gradient Boosting est une technique de machine Learning utilisée pour la régression et les problèmes de classification qui produit une prédiction composée à partir d'ensemble de prédictions de petits modèles :

- Prenant un modèle F1 qui va faire l'apprentissage et admettre un taux de vrai positif $TPR = \frac{TP}{TP+FN}$ en minimisant l'erreur quadratique moyenne.
- Le modèle F1 aura des erreurs résiduelles qui seront l'objet de l'apprentissage du modèle $h_1(x) = y - F_1(x)$
  La combinaison de ces deux modèles F_1 et h_1 crée un nouveau modèle. $F_2(x) = F_1(x) + h_1(x)$
- En répétant ces dernières étapes un nombre fini M fois, nous créons M modèles qui minimisent l'erreur de son prédécesseur. Nous avons alors un modèle F_M composé à partir d'ensemble de petits modèles.

Le Light gradient Boosting, développé par Microsoft, se différencie des autres gradients Boosting algorithme par une nouvelle technique d'échantillonnage appelé « Gradient-Based One-Side Sampling » (GOSS) qui consiste à garder toutes les instances avec de grands gradients, et sélectionne aléatoirement parmi les instances avec de petits gradients. La figure suivante est une illustration du principe d'échantillonnage cité auparavant :

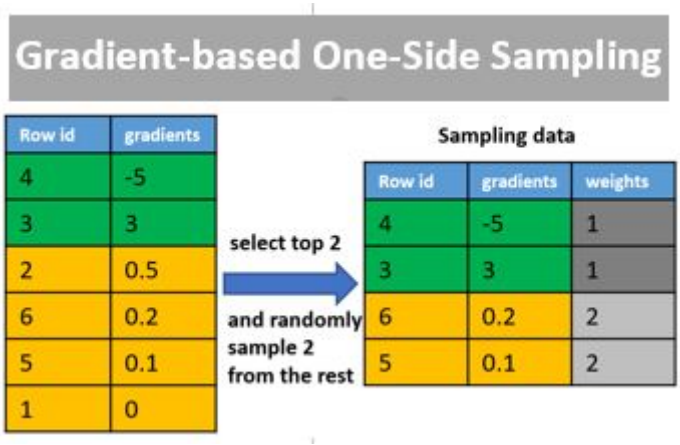


FIGURE 3.2 – Echantillonage du Light Gradient Boosting

3.1.4.2 Ingénierie des variables

Cette partie rend compte des transformations faites aux données avant l'apprentissage de l'algorithme. Nous avons créé les variables suivantes :

- credit_income_percent : amt_credit / amt_income_total
- annuity_income_percet : amt_annuity / amt_income_total
- credit_trem : amt_annuity / amt_credit
- days_employed_percent : days_employed / days_birth
- age = days_birth / -365
- age_bin : regroupement de la variable age par cinq ans.
- month_employed : days_emplloyed / -30
- umployement_bin : regroupement de la variable mount_employed par période de six mois.
- month_id_publish_bin : catégorisation de la (variable days_id_publish / -30) par groupe de 3.
- combinaisons polynomial des variables EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3, DAYS_BIRTH
- ratio_income_mean_org : amt_income_total / salaire moyen des individus de même catégorie d'organistion
- ratio income_mean_age : amt_income_total / salaire moyen des individus de même tranche de cinq ans d'âges.

- ratio_income_mean_age_org : ratio_income_mean_org * ratio_income_mean_âge

Enfin, nous avons modifié les valeurs 365243 de la variable days_employed par 30

3.1.4.3 Bilan

3.1.4.3.1 Problèmes rencontrés :

Les problèmes rencontrés sont nombreux principalement à cause de la grande taille des données :

- Problème d'agrégation : donnée volumineuse, coûteuse en temps et en mémoire vive
- Problème du calcul : le manque de mémoire vive conduit a de longue attente des résultats. Nous avons donc dû ajouter de la mémoire vive et étendre la mémoire SWAP

3.1.4.3.2 Score et classement au concours :

Le résultat calculé sur les données application_test.csv a rapporté un score AUC de 0.78891 qui nous donnait le rang de 3310 parmi 7198 équipes et participants.

3.1.4.4 Perspective d'évolution

En utilisant les données de home crédit, nous avons constaté une différence de performance entre les deux modèles. Le score AUC de donnée de validation de la BNI est de 0,72 contre un score AUC de 0,78 pour celui de donnée de classement de home crédit).

3.1.4.4.1 Perspective d'évolution selon l'importance des variables :

La feature importance des variables dans le modèle renseigne à quel point une variable spécifique a contribué pour la construction du modèle. La liste des 220 variables avec l'importance des variables peut être consultée dans les annexes. Par suite de ces variables ordonnées par feature importances, nous pouvons suggérer l'ajout des variables suivant au modèle de la BNI :

- L'âge de la voiture la plus récente que le client possède
- Versement restant sur le prêt précédent
- Le nombre de jours de changement de téléphone
- La densité de population du quartier du client (plus le nombre est élevé, plus le client vit dans une région plus peuplée)
- Le nombre de jours de retard lors des prêts précédents
- Montant en retard maximum sur les crédits antécédents

- Valeur immobilière de l'habitation du client
- Le niveau d'études du client

3.1.4.4.2 Perspective d'évolution selon la construction du modèle :

Devant d'aussi bons résultats du modèle conçu à partir de home crédit, nous avons tiré quelques idées de perspective d'évolution du modèle de la BNI : premièrement, certains prétraitements tels que la transformation polynomiale se sont avérés bénéfiques au modèle. Nous pourrions déjà ajouter ces prétraitements lorsque les variables suggérées sont disponibles. Enfin, pour atteindre le maximum de potentiel des données, nous suggérons aussi la technique Stacking. Le stacking est une technique d'assemblage de modèles, cette technique est utilisée pour combiner les informations de plusieurs modèles de prédictions. En utilisant la sortie de plusieurs modèles, on les combine avec diverses techniques en une seule. On intègre ainsi les informations de plusieurs modèles en une seule.

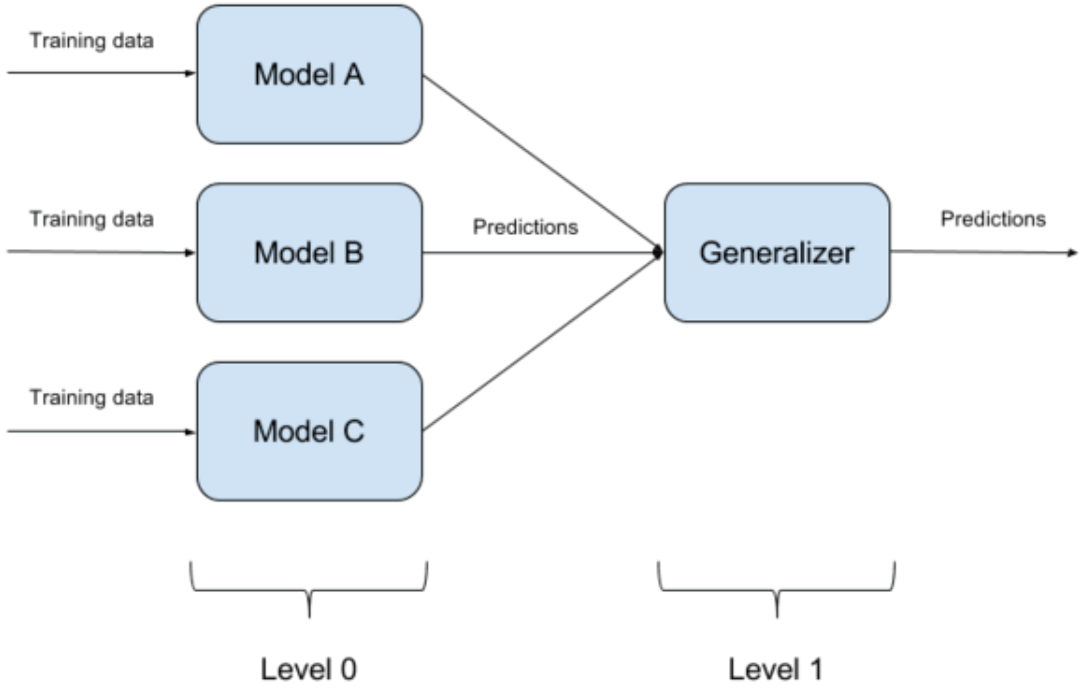


FIGURE 3.3 – Illustration du Stacking [6]

Conclusion

Face au problème important de trancher sur la décision d'octroi de crédit, nous avons construit au travers de ce travail un système, innovant l'ancien système de la BNI, d'attribution de points en vue d'une demande de prêt : un modèle d'apprentissage basé sur les données des clients de la BNI pour donner une probabilité de défaut de remboursement. Nous rappelons que dans le cadre de notre travail, nous avons analysé et étudié les existants de la BNI et des existants dans le domaine du Credit Scoring en général. Ensuite, nous avons construit quatre différents modèles à partir de quatre différents algorithmes d'apprentissage pour n'en choisir qu'un. En même temps, nous avons appliqué divers techniques de création et de transformation de variables. Enfin, nous nous sommes inspirés d'un concours organisé par Kaggle et HOME Credit sur le Credit Scoring pour proposer une amélioration sur ledit modèle de Credit Scoring de la BNI. L'objectif est de tirer au mieux le potentiel des données de la BNI afin d'avoir un modèle de prédiction performant.

Nous avons dans un premier temps et d'une part exposé les solutions et pratiques existantes dans le domaine du Credit Scoring. Le sujet fait depuis quelques décennies le sujet de certains articles scientifiques. Ainsi, nous avons recherché les algorithmes adapté au Credit Scoring. Toutefois, il est accepté aujourd'hui que l'algorithme performant dépend beaucoup des données d'entraînements. Puis, d'autre part nous avons étudié le système antérieur de grille de score pour l'octroi de prêt de la BNI pour la comprendre.

Dans un second temps, nous avons construit le modèle d'apprentissage basé sur les données de la BNI. D'abord, nous avons exploré les données de la BNI en vue de corrélation avec la variable à prédire. Ensuite nous avons créé quatre modèles à partir de quatre algorithmes d'apprentissages différents. Afin de comparer ces quatre modèles nous avons utilisé l'AUC ROC. Nous avons enfin choisi un parmi les quatre modèles pour l'intégrer dans le système d'octroi de crédit de la BNI.

Dernièrement, pour l'amélioration du modèle de Credit scoring, nous avons participé à un concours organisé par Kaggle et Home Credit. En soumettant notre modèle fait avec les données de Home Credit, nous avons remarqué une amélioration de la discrimination du modèle. Nous avons donc noté des variables à récolter en vue d'améliorer le modèle de prédiction de la BNI.

Table des figures

Liste des tableaux

Bibliographie

[1] Hussein A Abdou and John Pointon. Credit scoring, statistical techniques and evaluation criteria : a review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 18(2-3) :59–88, 2011.

[2] Younes Boujelbène and Sihem Khemakhem. Prévision du risque de crédit : Une étude comparative entre l'analyse discriminante et l'approche neuronale. *arXiv preprint arXiv :1311.4266*, 2013.

[3] Maria Teresinha Arns Steiner Celso Carnieri. Pattern recognition in credit scoring analysis. 1999.

[4] Boubacar Diallo. un modèle de "crédit scoring" pour une institution de micro-finance africaine : le cas de nyesigiso au mali. 2006.

[5] Kaggle. Home credit default risk, 2018. [Online ; accessed Mai 03, 2019].

[6] Keshav Dhandhania, Marta Enesco,Savan Visalpara. Ensemble methods (part 3) : Meta-learning, stacking and mixture of experts, 2018. [Online ; accessed Mai 03, 2019].

[7] Hassen MATHLOUTHI. Cours de méthode de scoring - ecole supérieure de statistique et d'analyse de l'information – université de carthage, 2014.

[8] Nicolas Castelein Nikos Skantzos. Credit scoring – case study in data analytics, 2006.

[9] Pierre Louis GONZALEZ. Calcul d'un score (scoring) application de techniques de discrimination, 2019. [Online ; accessed Mai 03, 2019].

[10] Véronique Rougès. Gestion bancaire du risque de non-remboursement des crédits aux entreprises : une revue de la littérature. In *Identification et maîtrise des risques : enjeux pour l'audit, la comptabilité et le contrôle de gestion*, pages CD–Rom, 2003.

[11] Sarang Narkhede. Understanding auc - roc curve, 2018. [Online ; accessed Mai 03, 2019].

[12] Marcos de Moraes Sousa and Reginaldo Santana Figueiredo. Credit analysis using data mining : application in the case of a credit union. *JISTEM-Journal of Information Systems and Technology Management*, 11(2) :379–396, 2014.

Annexe A

Le crédit scoring de la BNI

A.1 Les modèles non choisis de la BNI

A.1.1 Le Réseau de neurone

Notre modèle Neural Network comporte 8 couches cachées et 6 couches en dropout 6 . La couche finale est une couche qui contient la probabilité de « sinistre ». Pour le neural network, nous n'avons pas fait de k-folds 10, parce que l'apprentissage du neural network utilise déjà la Stochastique gradient Descent [1] qui est déjà un K-fold aléatoire.

La figure A.1 note la courbe ROC du modèle de réseau de neurone ;

A.1.2 Le modèle de Régression logistique

Le modèle de Régression logistique est entraîné à partir de 2856 observations de la donnée d'entraînements de l'ensemble de données 2011 à 2014 dont la proportion de clients sains et celle des clients sinistrés sont égale, après un ré-échantillonnage (resampling) aléatoire, on a donc ;

- 1428 observations de clients sains
- 1428 observations de clients sinistrés

Nous avons validé le modèle sur K-Fold où K=10 dont la table A.1 résume les 10 sous populations. Puis, la figure A.2 note la courbe ROC du modèle de de regression linéaire.

1. https ://en.wikipedia.org/wiki/Stochastic_gradient_descent

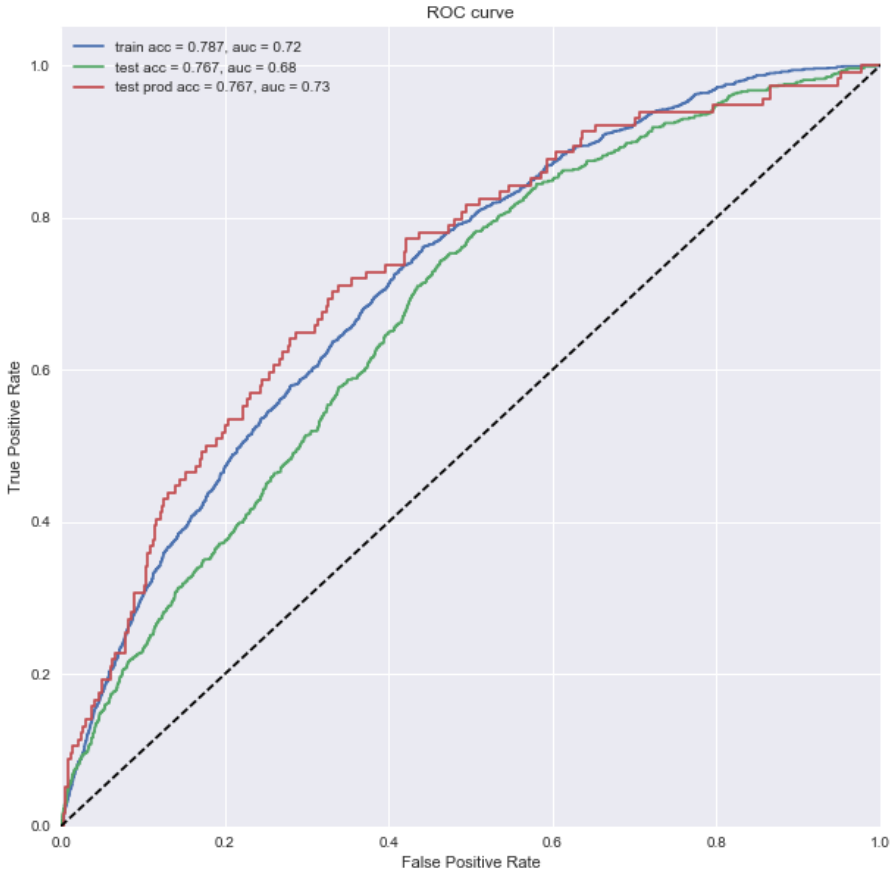


FIGURE A.1 – ROC curve du réseau de neurone

Moyenne	69.1945484619799
Min	63.68369
Max	72.32359
Médiane	69.14

TABLE A.1 – Résumé du Score AUC K-fold de la Régression Logistique

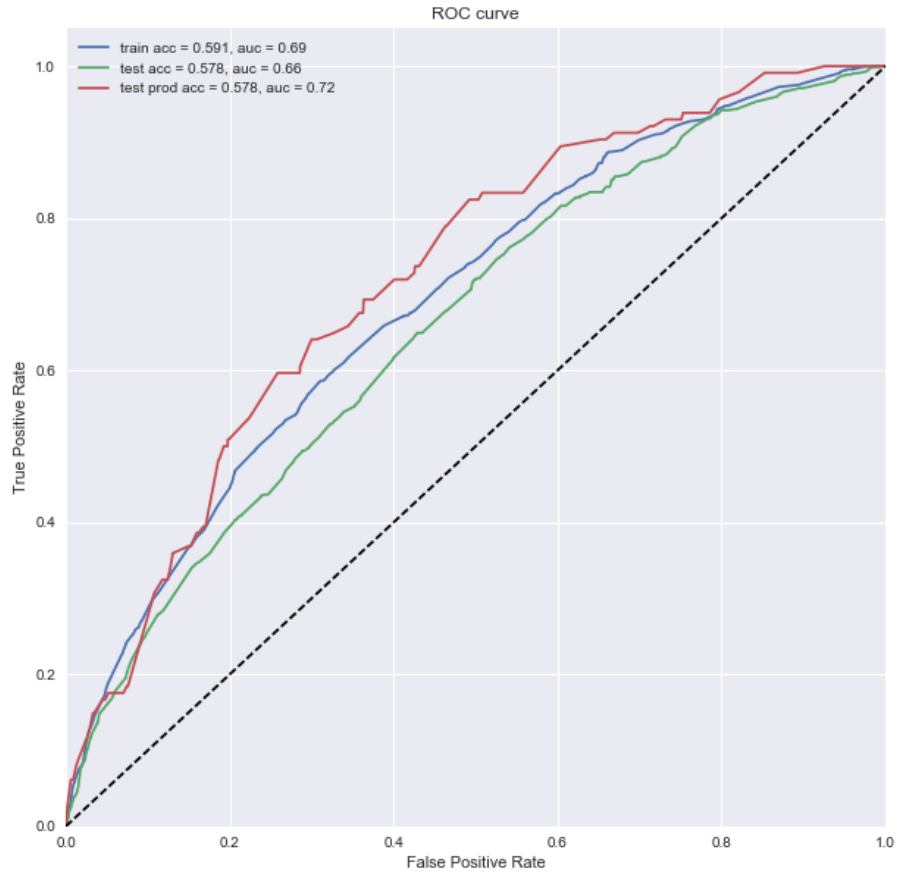


FIGURE A.2 – ROC curve du modèle régression logistique

A.1.3 Le modèle d'Analyse discriminante Linéaire (LDA)

Nous avons validé le modèle sur K-Fold où K = 10 dont voici le résumé des 10 sous populations ;

Moyenne	0.624
Min	0.603
Max	0.647
Médiane	0.62

Puis la table suivante rapporte les performances du modèle pour un seuil de 0.5

Année	AUC	Seuil	Sensitivité %	Spécificité %	Efficacité %	Taux d'erreur %
2014	0,63	0,5	60,96	66,82	61,15	66,82
2013	0,57	0,5	57,45	58,12	57,51	42,49
2012	0,63	0,5	51,41	75,37	53,11	46,89
2011	0,64	0,5	56,75	72,72	58,69	41,31
2017	0,62	0,5	71,8	52,63	71,22	28,77

Annexe B

Information Limite et proposition d'amélioration

B.1 Informations sur le concours et les données de Kaggle

B.1.1 HOME DESCRIPTION

Beaucoup de gens luttent pour obtenir des prêts en raison d'antécédents de crédit insuffisants ou inexistants. Et, malheureusement, cette population est souvent exploitée par des prêteurs indignes de confiance.

Home crédit s'efforce d'élargir l'inclusion financière pour la population non bancarisée en fournissant une expérience d'emprunt positive et sûre. Afin de s'assurer que cette population mal desservie a une expérience de prêt positive, Home crédit utilise une variété de données alternatives - y compris des informations téléphoniques et transactionnelles - pour prédire les capacités de remboursement de leurs clients.

Bien que Home crédit utilise actuellement diverses méthodes statistiques et d'apprentissage automatique pour faire ces prédictions, Home Credit défie les membres de la communauté Kaggle pour aider à exploiter le plein potentiel de ses données. Cela permettra de s'assurer que les clients capables de rembourser ne sont pas rejetés et que les prêts sont accordés avec un capital, une échéance et un calendrier de remboursement qui permettront à leurs clients de réussir.

B.1.2 DATA DESCRIPTION

B.1.2.1 application_train|test.csv

C'est la table principale, divisée en deux fichiers pour Train (avec TARGET) et Test (sans TARGET). Données statiques pour toutes les applications. Une ligne représente un prêt dans notre échantillon de données.

B.1.2.2 bureau.csv

Les crédits antérieurs de tous les clients fournis par d'autres institutions financières qui ont été déclarés à Crédit Bureau (pour les clients qui ont un prêt dans notre échantillon). Pour chaque prêt de notre échantillon, il y a autant de lignes que de nombre de crédits que le client avait dans le Bureau de crédit avant la date de la demande.

B.1.2.3 bureau_balance.csv

Soldes mensuels des crédits précédents dans le bureau de crédit. Ce tableau comporte une ligne pour chaque mois d'historique de tous les crédits antérieurs rapportés au Bureau de crédit - i.e. le nombre de ligne de la table est de

$$N_1.N_2.N_3$$

où
N_1 est le nombre de prêt dans l'échantillon
N_2 est le nombre de crédits précédents relatifs
N_3 est le nombre de mois où nous avons un historique observable pour les crédits précédents

B.1.2.4 POS_CASH_balance.csv

Images instantanées des soldes mensuels des points de vente (points de vente) précédents et des prêts en espèces que le demandeur a obtenues avec home crédit.
Ce tableau comporte une ligne pour chaque mois d'historique de chaque crédit précédent en crédit à la consommation (crédits à la consommation et prêts en espèces) liée aux prêts de notre échantillon - i.e. le nombre de ligne dans le tableau est de

$$N_1.N_2.N_3$$

où
N_1 est le nombre de prêt dans l'échantillon
N_2 est le nombre de crédits relatifs précédents
N_3 est le nombre de mois dans lequel nous avons un historique observable pour les crédits précédents

B.1.2.5 credit_card_balance.csv

Instantanés de soldes mensuels des cartes de crédit précédentes que le demandeur a avec le crédit à domicile.
Ce tableau comporte une ligne pour chaque mois d'historique de chaque crédit précédent en crédit à la consommation (crédits à la consommation et prêts en espèces) liée aux prêts de notre échantillon.

B.1.2.6 previous_application.csv

Toutes les demandes précédentes pour des prêts de crédit à domicile des clients qui ont des prêts dans notre échantillon. Il y a une ligne pour chaque application précédente liée aux prêts dans notre échantillon de données.

B.1.2.7 installments_payments.csv

Historique des remboursements pour les crédits précédemment décaissés en HOME CREDIT liés aux prêts de notre échantillon. Il y a :
a) une ligne pour chaque paiement effectué plus
b) une ligne pour chaque paiement manqué.

B.2 Importance des variables du modèle

Le tableau suivant, qui est l'extrait des 220 variables utilisées par le modèle LightGBM lors du concours Kaggle, rapporte l'importance des variables.

Variables	Importance	Valeur normalisé
CREDIT_TERM	314,00	0,0265786355
EXT_SOURCE_1 EXT_SOURCE_2 EXT_SOURCE_3	164,20	0,0138987642
AMT_ANNUITY	154,60	0,0130861690
EXT_SOURCE_1_x	131,40	0,0111223972

DAYS_CREDIT_max	118,00	0,0099881497
AMT_CREDIT	102,60	0,0086846115
AMT_GOODS_PRICE	97,60	0,0082613848
DAYS_CREDIT_ENDDATE_max	95,40	0,0080751651
DAYS_ENDDATE_FACT_max	92,80	0,0078550872
DAYS_ID_PUBLISH	92,00	0,0077873709
DAYS_EMPLOYED_PERCENT	87,20	0,0073810733
EXT_SOURCE_3_x	84,80	0,0071779245
OWN_CAR_AGE	83,60	0,0070763501
AMT_CREDIT_SUM_DEBT_mean	83,00	0,0070255629
ANNUITY_INCOME_PERCENT	79,40	0,0067208397
DAYS_REGISTRATION	74,20	0,0062806839
DAYS_BIRTH_x	70,40	0,0059590317
EXT_SOURCE_2 EXT_SOURCE_3 DAYS_BIRTH	68,00	0,0057558829
CNT_INSTALMENT_FUTURE_pos_mean	66,80	0,0056543084
EXT_SOURCE_2 EXT_SOURCE_3 2	66,60	0,0056373794
CREDIT_INCOME_PERCENT	66,40	0,0056204503
EXT_SOURCE_2 ^ 2 EXT_SOURCE_3	66,40	0,0056204503
DAYS_LAST_PHONE_CHANGE	65,00	0,0055019468
REGION_POPULATION_RELATIVE	64,80	0,0054850178
AMT_CREDIT_SUM_mean	61,60	0,0052141527
AMT_PAYMENT_installments_min_sum	60,20	0,0050956492
DAYS_CREDIT_mean	59,60	0,0050448620
AMT_CREDIT_SUM_max	58,00	0,0049094295
EXT_SOURCE_1 ^ 2 DAYS_BIRTH	57,40	0,0048586423
EXT_SOURCE_2_x	57,00	0,0048247842
EXT_SOURCE_1 EXT_SOURCE_3 DAYS_BIRTH	55,80	0,0047232098
CNT_PAYMENT_var	55,40	0,0046893516
cnt_late_days_payment_installments_max_sum	53,00	0,0044862028
AMT_CREDIT_SUM_min	52,80	0,0044692737

AMT_CREDIT_SUM_sum	52,20	0,0044184865
AMT_PAYMENT_installments_min_mean	51,80	0,0043846284
CNT_INSTALMENT_FUTURE_pos_mean_var	50,80	0,0042999831
cnt_late_days_payment_installments_max_mean	50,40	0,0042661249
DAYS_ENTRY_PAYMENT_installments_max_max	49,40	0,0041814796
AMT_CREDIT_MAX_OVERDUE_mean	48,00	0,0040629761
cnt_late_days_payment_installments_std_min	47,40	0,0040121889
DAYS_CREDIT_ENDDATE_sum	47,20	0,0039952599
CNT_PAYMENT_mean	47,00	0,0039783308
CNT_INSTALMENT_FUTURE_pos_min_mean	45,80	0,0038767564
EXT_SOURCE_2 EXT_SOURCE_3	45,60	0,0038598273
EXT_SOURCE_1 DAYS_BIRTH ^ 2	45,20	0,0038259692
EXT_SOURCE_1 EXT_SOURCE_2 ^ 2	45,20	0,0038259692
DAYS_EMPLOYED	44,80	0,0037921111
NAME_FAMILY_STATUS_Married	44,40	0,0037582529
cnt_late_days_payment_installments_sum_max	44,40	0,0037582529
age	43,00	0,0036397494
EXT_SOURCE_3 DAYS_BIRTH ^ 2	42,40	0,0035889622
nb_active	41,80	0,0035381750
AMT_PAYMENT_installments_min_min	41,40	0,0035043169
cnt_late_days_payment_installments_max_min	41,20	0,0034873878
CODE_GENDER_F	41,20	0,0034873878
DAYS_CREDIT_UPDATE_max	41,00	0,0034704588
DAYS_CREDIT_ENDDATE_mean	40,60	0,0034366006
EXT_SOURCE_2 DAYS_BIRTH ^ 2	40,00	0,0033858134
NAME_EDUCATION_TYPE_Higher education	39,80	0,0033688844
have_prev_app_refused	39,20	0,0033180972
EXT_SOURCE_1 ^ 2 EXT_SOURCE_2	39,00	0,0033011681
EXT_SOURCE_3 ^ 2 DAYS_BIRTH	39,00	0,0033011681
Unnamed : 0_y	39,00	0,0033011681

SK_ID_CURR__x	39,00	0,0033011681
DAYS_CREDIT_min	38,00	0,0032165228
EXT_SOURCE_1 EXT_SOURCE_2 DAYS_BIRTH	37,80	0,0031995937
cnt_late_days_payment_installments_sum_mean	37,80	0,0031995937
CNT_INSTALMENT_FUTURE_pos_min_sum	37,00	0,0031318774
CNT_PAYMENT_sum	36,80	0,0031149484
AMT_ANNUITY_mean_x	36,80	0,0031149484
EXT_SOURCE_1 DAYS_BIRTH	36,80	0,0031149484
EXT_SOURCE_3 DAYS_BIRTH	36,60	0,0030980193
month_employed	34,80	0,0029456577
DAYS_DECISION_var	34,80	0,0029456577
AMT_ANNUITY_min_x	34,40	0,0029117996
AMT_PAYMENT_installments_sum_mean	34,40	0,0029117996
EXT_SOURCE_2 ^ 2 DAYS_BIRTH	34,20	0,0028948705
AMT_CREDIT_SUM_DEBT_sum	34,20	0,0028948705
DAYS_ENDDATE_FACT_min	34,00	0,0028779414
cnt_late_days_payment_installments_mean_max	33,60	0,0028440833
AMT_INSTALMENT_installments_min_min	33,40	0,0028271542
AMT_CREDIT_MAX_OVERDUE_max	33,20	0,0028102252
DAYS_DECISION_max	32,80	0,0027763670
DAYS_DECISION_mean	32,80	0,0027763670
REGION_RATING_CLIENT_W_CITY	32,40	0,0027425089
CNT_INSTALMENT_FUTURE_pos_min_var	32,20	0,0027255798
AMT_ANNUITY_var	31,80	0,0026917217
cnt_late_days_payment_installments_min_var	31,80	0,0026917217
AMT_CREDIT_SUM_DEBT_max	31,40	0,0026578636
TOTALAREA_MODE	30,60	0,0025901473
INSTALMENT_VERSION_installments_summean	30,60	0,0025901473
DAYS_CREDIT_ENDDATE_min	30,40	0,0025732182
EXT_SOURCE_1 EXT_SOURCE_3 ^ 2	30,20	0,0025562891

AMT_INSTALMENT_installments_mean_min	30,20	0,0025562891
CODE_GENDER_M	30,00	0,0025393601
DAYS_CREDIT_UPDATE_mean	30,00	0,0025393601
cnt_late_days_payment_installments_mean_var	29,80	0,0025224310
DAYS_CREDIT_sum	29,80	0,0025224310
AMT_CREDIT_SUM_LIMIT_mean	29,80	0,0025224310
EXT_SOURCE_1 ^ 2 EXT_SOURCE_3	29,60	0,0025055019
EXT_SOURCE_1 EXT_SOURCE_3	29,40	0,0024885729
DAYS_ENDDATE_FACT_mean	29,00	0,0024547147
NAME_CONTRACT_TYPE_Cash loans	29,00	0,0024547147
cnt_late_days_payment_installments_sum_var	28,80	0,0024377857
AMT_CREDIT_MAX_OVERDUE_sum	28,60	0,0024208566
cnt_late_days_payment_installments_max_var	28,20	0,0023869985
DAYS_CREDIT_UPDATE_sum	28,20	0,0023869985
FLAG_DOCUMENT_3	28,20	0,0023869985
EXT_SOURCE_1 EXT_SOURCE_2	28,20	0,0023869985
ratio_income_mean_age	28,00	0,0023700694
cnt_late_days_payment_installments_sum_min	27,80	0,0023531403
cnt_installment_payment_difficult_mean	27,40	0,0023192822
DAYS_ENDDATE_FACT_sum	27,20	0,0023023531
missing_payment_installments_mean_mean	27,00	0,0022854241
DAYS_CREDIT_UPDATE_min	27,00	0,0022854241
cnt_late_days_payment_installments_mean_min	26,60	0,0022515659
AMT_APPLICATION_mean	26,40	0,0022346369
DAYS_LAST_DUE_var	26,20	0,0022177078
AMT_PAYMENT_installments_std_min	25,80	0,0021838497
EDUCATION_TYPE_Secondary secondary special	25,80	0,0021838497
MONTHS_BALANCE_pos_max_max	25,40	0,0021499915
cnt_late_days_payment_installments_var_var	25,00	0,0021161334
cnt_late_days_payment_installments_max_max	25,00	0,0021161334

DAYS_INSTALMENT_installments_max_max	25,00	0,0021161334
nb_Microloan	25,00	0,0021161334
cnt_late_days_payment_installments_min_mean	25,00	0,0021161334
AMT_PAYMENT_installments_sum_var	25,00	0,0021161334
HOUR_APPR_PROCESS_START	24,80	0,0020992043
ratio_income_mean	24,60	0,0020822753
AMT_PAYMENT_installments_sum_sum	24,60	0,0020822753
cnt_late_days_payment_installments_mean_mean	24,60	0,0020822753
AMT_INSTALMENT_installments_min_mean	24,60	0,0020822753
AMT_PAYMENT_installments_sum_max	24,40	0,0020653462
CNT_INSTALMENT_pos_std_mean	24,40	0,0020653462
CNT_INSTALMENT_FUTURE_pos_mean_max	24,20	0,0020484171
cnt_late_days_payment_installments_sum_sum	24,00	0,0020314881
cnt_late_days_payment_installments_mean_sum	23,80	0,0020145590
AMT_APPLICATION_var	23,80	0,0020145590
AMT_CREDIT_var	23,80	0,0020145590
cnt_late_days_payment_installments_std_var	23,60	0,0019976299
AMT_PAYMENT_installments_mean_min	23,60	0,0019976299
BASEMENTAREA_MODE	23,40	0,0019807009
AMT_INSTALMENT_installments_sum_min	23,40	0,0019807009
EXT_SOURCE_2 DAYS_BIRTH	23,20	0,0019637718
INSTALMENT_VERSION_installments_mean	23,00	0,0019468427
AMT_INSTALMENT_installments_min_var	22,80	0,0019299137
nb_Mortgage	22,60	0,0019129846
DAYS_ENTRY_PAYMENT_installments_std_min	22,60	0,0019129846
AMT_CREDIT_SUM_LIMIT_max	22,00	0,0018621974
cnt_late_days_payment_installments_var_mean	21,80	0,0018452683
cnt_late_days_payment_installments_std_mean	21,60	0,0018283393
MONTHS_BALANCE_pos_sum_var	21,60	0,0018283393
INSTALMENT_VERSION_installments_sum_var	21,60	0,0018283393

AMT_PAYMENT_installments_min_max	21,40	0,0018114102
DEF_30_CNT_SOCIAL_CIRCLE	21,40	0,0018114102
AMT_ANNUITY_max_x	21,40	0,0018114102
MONTHS_BALANCE_pos_sum_min	21,20	0,0017944811
AMT_CREDIT_SUM_DEBT_min	21,20	0,0017944811
nb_0_mean	21,00	0,0017775521
CNT_INSTALMENT_FUTURE_pos_std_var	20,60	0,0017436939
ratio_income_mean_age_org	20,60	0,0017436939
nb_0_sum	20,40	0,0017267649
cnt_late_days_payment_installments_std_max	20,40	0,0017267649
APARTMENTS_MODE	20,40	0,0017267649
cnt_late_days_payment_installments_min_min	20,20	0,0017098358
DAYS_LAST_DUE_1ST_VERSION_sum	20,00	0,0016929067
AMT_PAYMENT_installments_sum_min	20,00	0,0016929067
DAYS_LAST_DUE_sum	20,00	0,0016929067
DAYS_LAST_DUE_mean	19,60	0,0016590486
DAYS_DECISION_sum	19,60	0,0016590486
DEF_60_CNT_SOCIAL_CIRCLE	19,60	0,0016590486
AMT_ANNUITY_mean_y	19,40	0,0016421195
MONTHS_BALANCE_pos_sum_mean	19,40	0,0016421195
cnt_late_days_payment_installments_var_sum	19,20	0,0016251905
DAYS_LAST_DUE_1ST_VERSION_max	19,00	0,0016082614
missing_payment_installments_sum_mean	19,00	0,0016082614
AMT_INSTALMENT_installments_sum_var	18,80	0,0015913323
CNT_INSTALMENT_pos_std_var	18,80	0,0015913323
COMMONAREA_MODE	18,80	0,0015913323
APARTMENTS_AVG	18,60	0,0015744033
DAYS_LAST_DUE_1ST_VERSION_var	18,60	0,0015744033
LANDAREA_MODE	18,60	0,0015744033
AMT_CREDIT_max	18,40	0,0015574742

CNT_INSTALMENT_FUTURE_pos_min_max	18,40	0,0015574742
MONTHS_BALANCE_pos_max_var	18,40	0,0015574742
INSTALMENT_VERSION_installments_sum	18,20	0,0015405451
YEARS_BUILD_MODE	18,20	0,0015405451
CNT_DRAWINGS_ATM_CURRENT_credit_mean_min	18,20	0,0015405451
AMT_INSTALMENT_installments_max_min	18,20	0,0015405451
CNT_INSTALMENT_pos_min_var	18,20	0,0015405451
cnt_late_days_payment_installments_min_sum	18,20	0,0015405451
cnt_late_days_payment_installments_min_max	18,20	0,0015405451
AMT_CREDIT_mean	18,20	0,0015405451
missing_payment_installments_mean_sum	18,00	0,0015236160
MONTHS_BALANCE_pos_min_var	18,00	0,0015236160
missing_payment_installments_min_var	17,80	0,0015066870
DAYS_INSTALMENT_installments_sum_var	17,60	0,0014897579
AMT_INSTALMENT_installments_sum_mean	17,60	0,0014897579
MONTHS_BALANCE_pos_var_var	17,60	0,0014897579
LIVINGAREA_AVG	17,60	0,0014897579
AMT_ANNUITY_sum_x	17,40	0,0014728288
AMT_CREDIT_min	17,40	0,0014728288
CNT_PAYMENT_max	17,40	0,0014728288
nb_c_sum	17,40	0,0014728288
DAYS_FIRST_DUE_var	17,40	0,0014728288
LANDAREA_AVG	17,20	0,0014558998
CNT_INSTALMENT_pos_sum_var	17,20	0,0014558998
nb_x_mean	17,20	0,0014558998
CNT_INSTALMENT_FUTURE_pos_sum_var	17,00	0,0014389707
CNT_INSTALMENT_FUTURE_pos_mean_min	16,80	0,0014220416
INSTALMENT_VERSION_installments_mean_max	16,80	0,0014220416
AMT_PAYMENT_installments_mean_max	16,40	0,0013881835
OCCUPATION_TYPE_Core staff	16,40	0,0013881835

AMT_APPLICATION_max	16,40	0,0013881835
MONTHS_BALANCE_pos_sum_max	16,40	0,0013881835
missing_payment_installments_mean_min	16,20	0,0013712544
AMT_INSTALMENT_installments_min_max	16,20	0,0013712544
REG_CITY_NOT_LIVE_CITY	16,20	0,0013712544
cnt_installment_payment_difficult_var	16,20	0,0013712544
LIVINGAREA_MODE	16,20	0,0013712544
DAYS_DECISION_min	16,00	0,0013543254
AMT_ANNUITY_max_y	16,00	0,0013543254
DAYS_LAST_DUE_1ST_VERSION_mean	16,00	0,0013543254
INSTALMENT_VERSION_installments_var_var	16,00	0,0013543254
DAYS_FIRST_DUE_max	15,80	0,0013373963
CNT_INSTALMENT_FUTURE_pos_var_var	15,80	0,0013373963
EXT_SOURCE_2_y	15,60	0,0013204672
INSTALMENT_VERSION_installments_mean_sum	15,60	0,0013204672
LIVINGAPARTMENTS_MODE	15,60	0,0013204672
nb_Credit_card	15,60	0,0013204672
AMT_PAYMENT_installments_max_min	15,60	0,0013204672
DAYS_ENTRY_PAYMENT_installments_sum_max	15,40	0,0013035382
AMT_INSTALMENT_installments_sum_sum	15,40	0,0013035382
NONLIVINGAREA_AVG	15,20	0,0012866091
cnt_late_days_payment_installments_std_sum	15,20	0,0012866091
DAYS_INSTALMENT_installments_sum_max	15,20	0,0012866091
CNT_INSTALMENT_FUTURE_pos_std_min	15,20	0,0012866091
DAYS_LAST_DUE_1ST_VERSION_min	15,00	0,0012696800
NONLIVINGAREA_MODE	15,00	0,0012696800
AMT_APPLICATION_min	15,00	0,0012696800
AMT_PAYMENT_installments_min_var	15,00	0,0012696800
AMT_INSTALMENT_installments_std_min	15,00	0,0012696800
INSTALMENT_VERSION_installments_mean_var	15,00	0,0012696800

AMT_INCOME_TOTAL	14,80	0,0012527510
DAYS_INSTALMENT_installments_sum_min	14,80	0,0012527510
DAYS_INSTALMENT_installments_max_var	14,80	0,0012527510
INSTALMENT_VERSION_installments_std_mean	14,60	0,0012358219
INSTALMENT_VERSION_installments_std_var	14,60	0,0012358219
CNT_INSTALMENT_FUTURE_pos_std_mean	14,60	0,0012358219
nb_0_max	14,40	0,0012188928
DAYS_ENTRY_PAYMENT_installments_std_mean	14,40	0,0012188928
DAYS_ENTRY_PAYMENT_installments_sum_var	14,20	0,0012019638
CNT_INSTALMENT_pos_var_var	14,20	0,0012019638
AMT_INSTALMENT_installments_min_sum	14,20	0,0012019638
DAYS_ENTRY_PAYMENT_installments_mean_max	14,00	0,0011850347
MONTHS_BALANCE_pos_var_mean	14,00	0,0011850347
BASEMENTAREA_AVG	14,00	0,0011850347
MONTHS_BALANCE_pos_std_mean	14,00	0,0011850347
NAME_INCOME_TYPE_State servant	14,00	0,0011850347
CNT_INSTALMENT_FUTURE_pos_mean_sum	13,80	0,0011681056
AMT_APPLICATION_sum	13,80	0,0011681056
DAYS_INSTALMENT_installments_std_mean	13,80	0,0011681056
LIVINGAPARTMENTS_AVG	13,60	0,0011511766
AMT_INSTALMENT_installments_max_var	13,60	0,0011511766
CNT_INSTALMENT_pos_sum_min	13,60	0,0011511766
nb_pos_active_var	13,60	0,0011511766
DAYS_FIRST_DUE_mean	13,60	0,0011511766
DAYS_INSTALMENT_installments_sum_mean	13,60	0,0011511766
CNT_INSTALMENT_pos_min_mean	13,60	0,0011511766
CNT_INSTALMENT_FUTURE_pos_sum_min	13,60	0,0011511766
DAYS_INSTALMENT_installments_std_var	13,40	0,0011342475
AMT_CREDIT_sum	13,20	0,0011173184
COMMONAREA_AVG	13,20	0,0011173184

MONTHS_BALANCE_pos_mean_var	13,20	0,0011173184
APARTMENTS_MEDI	13,00	0,0011003894
DAYS_INSTALMENT_installments_min_max	13,00	0,0011003894
DAYS_ENTRY_PAYMENT_installments_std_var	13,00	0,0011003894
INSTALMENT_VERSION_installments_var_mean	12,80	0,0010834603
DAYS_FIRST_DUE_sum	12,80	0,0010834603
AMT_PAYMENT_installments_mean_var	12,80	0,0010834603
AMT_INSTALMENT_installments_mean_var	12,80	0,0010834603
missing_payment_installments_sum_sum	12,60	0,0010665312
FLOORSMAX_AVG	12,60	0,0010665312
CNT_INSTALMENT_pos_min_sum	12,60	0,0010665312
DAYS_ENTRY_PAYMENT_installments_var_mean	12,40	0,0010496022
LIVINGAREA_MEDI	12,20	0,0010326731
AMT_ANNUITY_min_y	12,20	0,0010326731
INSTALMENT_VERSION_installments_std_sum	12,20	0,0010326731
MONTHS_BALANCE_pos_var_sum	12,00	0,0010157440
ORGANIZATION_TYPE_Self-employed	12,00	0,0010157440
AMT_INSTALMENT_installments_max_mean	12,00	0,0010157440
DAYS_ENTRY_PAYMENT_installments_sum_min	12,00	0,0010157440
AMT_CREDIT_SUM_LIMIT_sum	11,80	0,0009988150
MONTH_ID_PUBLISH_bin	11,80	0,0009988150
nb_closed	11,80	0,0009988150
MONTHS_BALANCE_pos_std_var	11,80	0,0009988150
INSTALMENT_VERSION_installments_var_sum	11,80	0,0009988150
AMT_PAYMENT_installments_mean_mean	11,80	0,0009988150
CNT_INSTALMENT_FUTURE_pos_var_mean	11,80	0,0009988150
DAYS_INSTALMENT_installments_mean_max	11,60	0,0009818859
FLAG_WORK_PHONE	11,60	0,0009818859
missing_payment_installments_min_mean	11,60	0,0009818859
MONTHS_BALANCE_pos_mean_max	11,40	0,0009649568

AMT_CREDIT_SUM_OVERDUE_max	11,40	0,0009649568
DAYS_ENTRY_PAYMENT_installments_var_sum	11,40	0,0009649568
INSTALMENT_NUMBER_installments_mean_var	11,40	0,0009649568
DAYS_INSTALMENT_installments_std_sum	11,40	0,0009649568
missing_payment_installments_min_min	11,40	0,0009649568
CNT_DRAWINGS_ATM_CURRENT_credit_mean_max	11,40	0,0009649568
NONLIVINGAREA_MEDI	11,40	0,0009649568
CNT_INSTALMENT_pos_sum_mean	11,20	0,0009480278
nb_pos_active_mean	11,20	0,0009480278
CNT_INSTALMENT_FUTURE_pos_std_max	11,20	0,0009480278
AMT_ANNUITY_sum_y	11,20	0,0009480278
AMT_PAYMENT_installments_std_var	11,00	0,0009310987
MONTHS_BALANCE_pos_std_sum	11,00	0,0009310987
SK_DPD_DEF_pos_max_mean	11,00	0,0009310987
DAYS_ENTRY_PAYMENT_installments_std_max	11,00	0,0009310987
nb_c_mean	11,00	0,0009310987
CNT_INSTALMENT_pos_mean_var	11,00	0,0009310987
missing_payment_installments_sum_var	11,00	0,0009310987
CNT_INSTALMENT_pos_var_mean	11,00	0,0009310987
CNT_INSTALMENT_pos_mean_mean	10,80	0,0009141696
CNT_INSTALMENT_FUTURE_pos_sum_mean	10,80	0,0009141696
INSTALMENT_VERSION_installments_std_max	10,80	0,0009141696
LIVINGAPARTMENTS_MEDI	10,80	0,0009141696
AMT_INSTALMENT_installments_sum_max	10,80	0,0009141696
AMT_INSTALMENT_installments_std_mean	10,60	0,0008972406
INSTALMENT_NUMBER_installments_sum_var	10,60	0,0008972406
ORGANIZATION_TYPE_Industry : type 9	10,60	0,0008972406
MONTHS_BALANCE_pos_max_mean	10,60	0,0008972406
YEARS_BUILD_AVG	10,40	0,0008803115
nb_pos_active_sum	10,40	0,0008803115

LANDAREA_MEDI	10,40	0,0008803115
ENTRANCES_AVG	10,40	0,0008803115
AMT_PAYMENT_installments_var_var	10,40	0,0008803115
COMMONAREA_MEDI	10,40	0,0008803115
DAYS_ENTRY_PAYMENT_installments_sum_mean	10,20	0,0008633824
DAYS_ENTRY_PAYMENT_installments_max_mean	10,20	0,0008633824
AMT_PAYMENT_installments_mean_sum	10,20	0,0008633824
MONTHS_BALANCE_pos_sum_sum	10,20	0,0008633824
CNT_INSTALMENT_pos_sum_max	10,20	0,0008633824
INSTALMENT_NUMBER_installments_var_var	10,20	0,0008633824
SK_DPD_DEF_pos_std_mean	10,20	0,0008633824
ORGANIZATION_TYPE_Military	10,20	0,0008633824

Résumé

L'octroi de crédits a toujours été un problème de décision fondé sur la capacité de remboursement des emprunteurs. Auparavant, basée sur une grille de critères, la banque BNI avait consenti des prêts avec un taux élevé de non-remboursement et un taux inconnu de prêts qui auraient dû être accordés.

À la fin des recherches et des études sur la création de modèles de credit scoring, les travaux ont abouti au choix d'un modèle d'apprentissage parmi les quatre construits pour l'appréciation de crédit de la BNI. Puis à partir de notre participation au concours de Kaggle sur le credit scoring, nous avons déduit une proposition pour améliorer le modèle de credit scoring de la BNI.

Abstract

Credit granting has always been a decision-making problem based on the repayment capacity of borrowers. Previously based on a criteria grid, the BNI bank granted loans with a high rate of non-repayment and an unknown rate of loans that should have been granted.

At the end of research and studies on the creation of credit scoring models, the work culminated in the choice of one from four Machine Learning model for BNI scoring credit and a model improvement perspective based on the credit Scoring model during our participation in the Kaggle competition.

Titre : Mise en place d'un système d'attribution d'une demande de prêts
Auteur : ANDRIANIRINA Sandratra Herimanana (herysandratra@gmail.com)
Encadreur : Randrianarivony Arthur (arthur.randrianarivony@gmail.com)
Co-encadreur : Rasoanaivo Andry (r.andry.rasoanaivo@gmail.com)