HEC MONTRÉAL

Modèle d'estimation des revenus pour les clients d'une banque canadienne

par

Alexandre Gauro

Sciences de gestion

Projet supervisé présenté en vue de l'obtention du grade de maîtrise ès Sciences (M.Sc)

Décembre 2018

Table des matières

Liste des tableaux

Liste des figures

Liste des annexes

1. Introduction

Les formulaires d'application au crédit peuvent souvent paraître comme étant une formalité donnant accès à du crédit aux yeux d'un consommateur. Mais qu'en est-il de l'information recueillie à cet effet par l'institution financière? Outre les renseignements personnels permettant l'identification, le revenu annuel déclaré représente l'une des informations les plus importantes divulguées par le client. L'institution financière utilise cette information pour calculer son ratio d'endettement. Ce ratio, plus connu sous le nom de ATD (Amortissement Total Dette), permet d'évaluer la capacité d'un demandeur à contracter davantage de dettes via un prêt automobile, une consolidation de dettes, un prêt personnel ou encore une limite à octroyer sur une carte de crédit. Le ratio est calculé systématiquement sur chaque demande de crédit et plusieurs standards sont établis quant aux seuils d'acceptation. L'honnêteté et la validité des revenus déclarés sont essentiels pour obtenir des ratios représentatifs, sachant que l'institution financière se fie entièrement sur ce que son client lui déclare. Une combinaison d'un faible ratio d'endettement, un bureau de crédit[1] de qualité et de bonnes habitudes de paiement avec l'institution financière sont essentiels pour avoir accès au financement. Ce rapport de stage présente l'élaboration d'un modèle d'estimation des revenus annuels ayant pour but de confirmer la validité des revenus déclarés par les clients, mitiger le risque de crédit et ainsi éviter l'octroi de crédit basé sur de fausses déclarations.

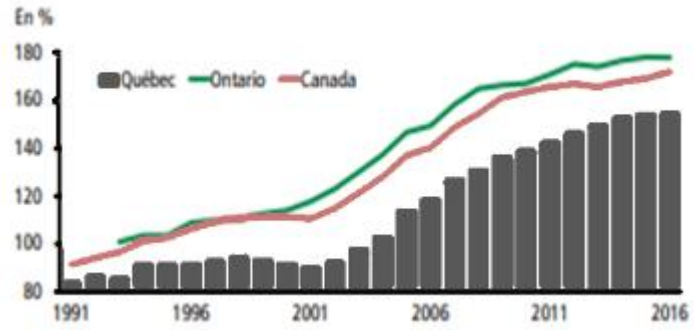
1.1 Évolution de la dette au Canada en lien avec l'estimation de revenus

Il est légitime de se questionner sur l'importance d'estimer correctement les revenus lors d'une demande de crédit. Or, la dette prend une place de plus en plus grande dans la vie des consommateurs canadiens. Celle-ci a augmenté de façon importante au Canada au cours des dernières années, sans pour autant que la croissance des revenus soit subséquente. Le phénomène de surconsommation que traverse la société actuelle en est la principale cause. Il est donc primordial, aujourd'hui, d'évaluer les revenus de façon

[1] Historique de l'utilisation du crédit du consommateur. Les principaux fournisseurs au Canada sont Equifax et TransUnion

précise sachant que le crédit continue d'être octroyé activement. On approche des seuils de capacités d'endettement maximal comme le démontre la Figure 1 alors que le ratio moyen d'endettement au Canada en 2016 était supérieur à 160% des revenus annuels bruts.

Figure 1 : Évolution du ratio crédit/revenus des consommateurs canadiens



Source : Institution financière canadienne

Ce sont les engagements mensuels sur ces dettes qui sont analysés par les institutions financières pour déterminer la capacité d'un client à contracter davantage de financement. Pour chaque demande de crédit, un ratio d'endettement est automatiquement calculé tel que :

$$Ratio\ Endettement\ (ATD) = \frac{Revenus\ mensuels\ bruts}{Engagements\ mensuels\ au\ crédit + Charge\ de\ loyer}$$

Par la suite, des standards sont établis selon chaque institution financière à savoir si le client se qualifie pour contracter plus de dettes. Un client ayant un ratio inférieur à 40% est généralement perçu comme étant en mesure d'augmenter son niveau d'endettement si la combinaison avec d'autres facteurs tels que la qualité de son bureau de crédit et ses habitudes de paiement sont bonnes. En ayant accès au bureau de crédit du demandeur, on peut s'assurer de la validité de notre dénominateur du ratio d'endettement. En effet, la charge de loyer représente souvent la mensualité sur l'hypothèque. Toutefois, la validité du numérateur ne peut être confirmée puisque l'institution financière se fie sur la

déclaration du demandeur. Comme ce ratio est l'une des principales mesures utilisées dans le processus d'octroi de crédit, la légitimité d'estimer les revenus annuels bruts prend tout son sens.

1.2 Présentation du projet et mandat

Ce mandat fut réalisé pour le département de gestion des risques d'une institution financière canadienne qui est actuellement dépendante des revenus fournis par sa clientèle. Notre but ultime était de développer un modèle permettant d'estimer les revenus des clients et de faire une gestion des risques efficiente en optimisant l'octroi de crédit. Le modèle permettrait à l'institution financière de connaître le profil d'un client en temps continu en s'ajustant de façon dynamique. On aimerait ainsi éviter des situations de surinvestissement, soit lorsque du crédit de « mauvaise qualité » est octroyé. À noter qu'on définit le crédit de mauvaise qualité comme étant du crédit octroyé à un demandeur alors qu'il ne se qualifiait pas. Actuellement, l'entreprise est à risque d'une situation telle qu'un jeune consommateur, déclarant un salaire annuel ne correspondant pas à son profil, pour lequel un faible ratio d'endettement serait calculé et qui se verrait attribuer un financement auquel il ne se qualifiait pas. Ce type de fausses déclarations s'apparente à de la fraude lors des déclarations d'informations personnelles. L'élaboration d'un modèle d'estimation des revenus est donc primordiale pour réduire cette dépendance à l'information déclarée par le consommateur. Ce projet est d'une importance capitale puisque l'institution financière n'a pas l'intention d'effectuer davantage de vérifications manuelles concernant les revenus déclarés en raison de l'investissement en coût et en temps requis. Le département de gestion des risques a pu me fournir une base de données d'environ 590 000 clients sur lesquels travailler. Des techniques mathématiques telles que les régressions linéaires, l'optimisation sous contraintes et le « backtesting » furent utilisées pour mener à terme le projet.

1.3 Objectifs

L'objectif initial du stage était de développer un modèle global pouvant estimer les revenus des clients et non-clients de l'institution financière qui appliquent pour du crédit.

Toutefois, selon le type de relation d'affaires existante avec chaque client, l'institution financière est en possession d'information différente sur chacun d'eux. Plus un client est fidèle et possède des produits à travers l'institution, plus la banque est en possession d'information lui permettant d'estimer les revenus avec une meilleure précision. Pour bénéficier de ce surplus d'information, nous avons décidé de séparer le modèle initial en 4 sous-modèles. Le client se répartit alors parmi les 4 scénarios suivants : ses revenus sont déposés via dépôt direct à l'institution (DI), le client a déjà fait une mise à jour de ses revenus dans le passé (MAJ), le client possède une carte de crédit avec l'institution (PMT) et le client n'est pas inclus dans les précédents modèles (SOCIO). Les modèles ne devraient évidemment pas tous performer de la même façon et un client peut être inclus dans plusieurs modèles en fonction de ce qu'il possède avec l'institution. Les ajustements à effectuer pour les clients pouvant être inclus dans plusieurs modèles sont discutés à la section 5.

Une base de données recueillant toutes ces informations était nécessaire pour permettre la formation des modèles. Le département de gestion des risques m'a alors créé une base de données sur mesure contenant toute l'information nécessaire à la réalisation du projet. Une autre partie de l'objectif fut donc de faire un nettoyage dans la base de données et de créer toutes les variables pertinentes pour les différents modèles. Suite aux attentes de l'institution financière quant à la précision des revenus estimés en fonction de la base de données fournie, nous avons convenu que si l'estimation des revenus se chiffrait à ± 20% ou ± 10 000$ des revenus réels annuels bruts, on pouvait qualifier l'estimation de bonne. Il est important de mentionner que le but de l'étude n'est pas d'observer de petites déviations entre les revenus déclarés et les revenus réels, mais bien d'être en mesure de trouver les écarts extrêmes. Ce seuil sera discuté aux sections 4 et 5 du rapport.

Le reste du rapport sera présenté comme suit : la section 2 contient une revue de la littérature sur l'estimation de revenus. La section 3 présente la base de données ainsi que les hypothèses faites pour créer les différentes variables. La section 4 détaille les différents modèles et résultats obtenus. La section 5 explique le phénomène d'optimisation à travers la combinaison de sous-modèles. Finalement, les sections 6 et 7 détaillent les applications possibles du projet, les recommandations et une conclusion.

2. Revue de la littérature

2.1 Risque lié à l'information fournie

Aujourd'hui, il serait trop coûteux pour une banque d'investiguer sur chaque demande de crédit et d'exiger une preuve de revenus en raison du volume élevé de demandes de financement. Les institutions financières préfèrent assumer les pertes liées à de fausses déclarations faute d'avoir une meilleure solution que des investigations manuelles. Il existe pourtant une asymétrie d'information entre le prêteur et l'emprunteur qui force les institutions à renforcer leur seuil d'octroi (Calem et al, 2011). Un emprunteur connaît ses véritables revenus annuels alors que le prêteur a besoin de cette information pour correctement évaluer l'engagement et l'endettement du demandeur. Le demandeur a donc intérêt à mentir sur son application puisqu'il n'y a pas de conséquences négatives liées à une fausse déclaration. Au contraire, il améliore ses chances que du crédit lui soit octroyé si son ratio d'endettement devient conforme au seuil d'acceptation. La sélection adverse contribue donc à maintenir des standards élevés sur la gestion des risques dans les emprunts bancaires en raison du principe d'asymétrie d'information existant entre les différents acteurs. Les standards plus élevés ne sont pas causés par un appétit pour le risque moins élevé, mais par un risque associé au consommateur plus grand en réalité que celui observé/divulgué.

L'étude réalisée par Blackburn et al (2012) prouve que la surestimation des revenus a été un phénomène observable sur les demandes de prêts du marché hypothécaire au début des années 2000. Les formulaires d'application au crédit qui permettent aux revenus et actifs d'être déclarés par le consommateur sans autres vérifications sont biaisés. Le manque de vérification de la part des institutions financières a fait en sorte que les revenus étaient en moyenne de 15 % à 20% surestimés. Ils ont également découvert une corrélation positive entre la surestimation des revenus et la délinquance sur prêt, soit la probabilité qu'un prêt se retrouve en situation de défaut. Les institutions financières se sont ajustées depuis au marché hypothécaire en exigeant une copie de la déclaration de revenus faite à l'impôt pour chaque demande de prêts hypothécaires. Cependant, il n'y a pas de tel mécanisme de vérification mis en place sur les prêts personnels ou encore le crédit rotatif. Le volume de ce type de prêts est beaucoup plus élevé et les montants octroyés sont généralement de

moindre importance. La quête de profit et d'automatisation des banques les expose à de nouveaux risques que les modèles traditionnels de « relationship banking » pouvaient contrôler. En automatisant le processus d'octroi de crédit, les banques négligent les vérifications manuelles qu'un prêteur se devait d'effectuer.
Avery et al (2004) en sont arrivés à la même conclusion, soit que des données erronées dans l'application au crédit engendrent des coûts et pertes monétaires. L'institution financière doit minimiser les refus inutiles et les acceptations trop généreuses.

Une solution proposée par Miller (2015) était alors d'augmenter le nombre de questions posées sur un formulaire d'application en espérant que les informations supplémentaires se traduisent en amélioration de la performance sur les prêts. Les résultats obtenus par l'auteure démontrent qu'un prêteur améliore sa performance de 0.12% en diminuant la probabilité de défaut sur prêts lorsque de la nouvelle information lui est disponible.

Or, augmenter le nombre de questions sur ces formulaires ne serait pas plus pertinent puisque tel que mentionné par Thomas et al (2002): « Plus le formulaire est long, plus la probabilité qu'un client le remplisse adéquatement est faible ». La qualité de l'information reçue serait amoindrie si le nombre de questions sur les formulaires devait augmenter. Idéalement, il faudrait développer un modèle sur les niveaux de revenus et dépenses d'un consommateur pour mieux définir les risques de défaut et éviter les erreurs de déclarations trompeuses des revenus. Ce modèle peut être créé avec l'information que possède l'institution financière à propos de son client (Thomas et al, 2009).

Cette sous-section permet de comprendre que l'asymétrie d'information existante entre l'emprunteur et le prêteur incite les emprunteurs à surestimer les revenus, phénomène qui fut observable par le passé. Cette surestimation engendre un taux de défaut sur prêt plus élevé que celui prévu originalement. Obtenir davantage d'informations peut aider à contrebalancer cet effet, mais il n'est pas recommandé d'allonger les formulaires actuels pour y parvenir. Un modèle d'estimation des revenus pourrait ainsi être une solution envisageable pour limiter les effets de la sélection adverse.

2.2 Engagement du consommateur

Les experts s'entendent pour affirmer que l'octroi de crédit de mauvaise qualité est négatif pour une banque puisque les termes du financement ne sont pas représentatifs du profil de risque de l'emprunteur. Par exemple, un demandeur de crédit, ayant déclaré de fausses informations sur son application, pourrait se voir octroyer un financement à 10% de taux d'intérêt annuels alors que le réel risque encouru pour ce client aurait nécessité un taux d'intérêt annuel de 15%. La gestion des risques de la banque devient ainsi moins efficiente en raison de ses actifs (prêts) qui comportent davantage de risques. Différentes propositions furent soumises au courant des dernières années pour tenter de tarifer plus efficacement le risque associé à un emprunteur et sa capacité d'engagement maximale.

2.2.1 Credit Scoring

Évaluer la qualité d'une demande de crédit à l'aide d'un score interne est l'une des pratiques émergentes dans l'industrie pour améliorer le processus d'octroi de crédit. Les prêteurs devraient continuer d'investir dans le développement de systèmes qui améliorent le « Credit scoring » pour éviter des situations d'engagement trop élevé (Finlay,2006). Le problème des demandes traditionnelles de crédit est que le profil de risque d'un individu peut fortement varier d'une application à l'autre en fonction du comportement de l'individu et de ce qu'il déclare. La capacité du demandeur à contracter une dette est donc intimement liée à sa déclaration, soit les revenus divulgués.

Certes, l'investissement dans des modèles de type « Credit scoring » permettrait d'améliorer le processus d'octroi, toutefois, l'évaluation des revenus n'est pas actuellement effectuée par ce type de modèles.

2.2.2 Capacité à payer

Un modèle pouvant déterminer la capacité à payer d'un consommateur permettrait de réduire le nombre de questions posées sur les formulaires d'application au crédit et de mieux tarifer le risque associé à un emprunteur. L'institution financière serait moins dépendante de l'information fournie par le demandeur, moins exposée à l'asymétrie

d'information et pourrait mieux évaluer la capacité du client à contracter davantage de dettes. Diboune (2008) a proposé un tel modèle représentant ainsi une alternative au ratio d'endettement typique utilisé par les institutions financières (ATD). L'auteure suppose ainsi que l'ATD n'est plus une mesure fiable puisqu'il ne prend pas en considération les conditions et le niveau de vie de chaque client comparativement à la mesure de Capacité qui inclue une variable « dépenses estimées ».

$$Capacité = revenu + tangible_r - dépenses\ estimées$$

La capacité à payer du demandeur varie alors selon ses revenus, ses actifs et ses dépenses courantes. L'étude réalisée par Diboune (2008) utilise le modèle présenté par Finlay (2006) pour estimer les dépenses d'un individu selon son âge, sexe, revenus, son statut d'habitation (Locataire/Propriétaire), ses actifs et d'autres variables. Elle tient pour acquis que les revenus et actifs déclarés par le consommateur sont valides et que ce sont plutôt les dépenses qui seraient faussées. La variable *Capacité* représente la capacité à payer d'un demandeur et découle de l'estimation des dépenses fournie par le modèle. L'une des hypothèses de base concernant l'utilisation d'une telle mesure de risque est que la banque n'a pas accès aux informations sur les dépenses de façon fiable. Or, « une bonne mesure de la capacité à payer doit inclure des données fiables sur les revenus » tels qu'avancé par Diboune (2008).

Nous allons estimer les revenus plutôt qu'estimer la capacité à payer d'un client (via les dépenses) tel que présenté par Diboune (2008). Plusieurs variables utilisées pour déterminer les dépenses seront reprises dans notre modèle d'estimation des revenus. L'ATD est encore aujourd'hui la mesure de risque la plus couramment utilisée. Le but de ce rapport n'est pas de remplacer l'ATD, mais plutôt de l'optimiser en ayant des revenus plus fiables avec lesquels calculer le ratio d'endettement du consommateur.

2.3 Estimation des revenus

Kibekbaev et al (2016) ont tenté de créer un modèle prédictif des revenus sur les clients des banques turques. En Turquie, le « Single Limit Law » est une loi qui prétend que la limite sur une carte de crédit ne peut excéder 4 fois les revenus mensuels du client. Leur étude avait pour but d'aider ces banques dans leur problème de prédiction des revenus en

offrant un processus décisionnel rapide et efficace pour maintenir une forte capacité d'octroi.

Les auteurs ont utilisé 18 différentes techniques de régression à l'aide de modèles linéaires et non linéaires. Leur conclusion affirme que la régression linéaire traditionnelle démontre des résultats similaires aux autres modèles de régression non linéaires plus sophistiqués. Le R^2 ajusté permet d'évaluer efficacement le pouvoir explicatif d'un modèle, toutefois, ils suggèrent de jumeler cette mesure de performance avec d'autres indicateurs pour mieux évaluer la performance d'un modèle.

En conclusion, un modèle prédictif des revenus permettra d'avoir un portrait plus juste de la relation rendement-risque encouru avec chaque consommateur. L'âge d'un demandeur, son statut d'habitation, ses actifs et plusieurs autres variables seront utilisées et découlent des hypothèses avancées par Diboune (2008) du modèle d'estimation des dépenses d'un individu. Les régressions linéaires, le R^2 ajusté et la mesure de performance des revenus (±20% ou ±10 000$) sont les principaux indicateurs et techniques qui seront utilisés pour la création du modèle. À noter que nous allons plutôt créer un modèle estimatif des revenus qu'un modèle prédictif puisque notre but est de détecter les fausses déclarations actuellement effectuées.

3. Données

3.1 Base de données

Tel que mentionné précédemment, il n'y avait pas de base de données existante qui combinait les revenus annuels déclarés par les demandeurs ainsi que d'autres variables provenant des formulaires d'application et des produits détenus à la banque. Le département de gestion des risques de la banque a donc créé une base de données unique à partir de plusieurs segments d'affaires pour l'élaboration du projet.

La base de données initiale fournie par l'institution financière contenait donc 1 020 403 données et plus de 60 variables. Chaque donnée provenait d'un mix parmi des demandes de carte de crédit, prêt automobile, prêt personnel ou encore simplement d'informations générales que la banque possédait sur ses clients existants. Le revenu déclaré s'étendait de janvier 2017 à mars 2018 et représentait la variable de performance (revenu réel) dans chaque régression linéaire. Ce revenu provenait de déclarations faites sur des demandes de crédit. Il fut toutefois nécessaire de procéder à un nettoyage des données en raison de nombreux doublons et problèmes techniques lors de l'assemblage des différents segments.

3.2 Nettoyage et hypothèses

Après avoir éliminé les doublons, les applications faites par des clients d'âge mineur et les erreurs d'assemblage, nous avons constaté que la base de données contenait toujours beaucoup de bruit (« White Noise ») et plusieurs déclarations de revenus réels illogiques. L'un des principaux problèmes lors de la réalisation du projet fut la base de données qui contenait beaucoup d'erreurs au niveau de la déclaration des revenus. Confusion de la part du consommateur entre revenus hebdomadaires-mensuels-annuels, difficulté d'assemblage de la base de données et fausses déclarations ont créé un biais directement dans l'échantillon de données. En effet, nous n'avons pas pu avoir accès à une base de données contenant uniquement des revenus validés réels pour créer le modèle. L'échantillon contenait également des fausses déclarations contribuant ainsi à fausser les résultats.

Quelques hypothèses furent alors avancées pour tenter de réduire le bruit dans la base de données et éliminer le plus possible les fausses déclarations.

Hypothèse 1 : Exclus revenus inférieurs 100$ net/semaine = inférieurs à 6500$ brut/année

Hypothèse 2 : Exclus revenus supérieurs à 200 000$ brut annuel

Hypothèse 3 : Exclus les données non incluses dans les modèles (voir section 4)

Les seuils de revenus minimums et maximums à pouvoir évaluer furent décidés conjointement avec les responsables du département de gestion des risques de la banque. L'intuition est la suivante : il est presque impossible d'estimer des revenus inférieurs à 100$ net/semaine (6500$ brut/annuel) sachant que selon Emploi Québec, même le revenu d'aide sociale au Québec en 2018 est supérieur à ce seuil (7600$ brut annuel). Également, il n'est pas pertinent d'inclure les revenus supérieurs à 200 000$ puisqu'ils sont généralement causés par une confusion entre revenu déclaré hebdomadaire mensuel annuel ou par de fausses déclarations. La 3^{e} hypothèse sera discutée plus en détail à la section 4 de ce rapport et représente la principale source de triage des données. Ces exclusions ont permis d'améliorer la précision des revenus déclarés dans la base de données en abaissant le nombre de données totales de 1 020 403 à 594 277 données de qualités. Les 594 277 données se séparent en 415 994 données utilisées pour les estimations et 178 283 données utilisées pour le « backtest », permettant ainsi de confirmer les résultats obtenus (proportion 70%-30%).

3.3 Variables créées

Malgré l'étendue d'informations disponibles originalement dans la base de données, ce sont finalement des variables créées manuellement à partir des variables fournies originales qui constituent les données utilisées lors des régressions linéaires. Comme mentionné à la section 1, la banque possède de l'information différente sur chaque client en fonction du nombre et type de produits qu'il possède avec elle. Les variables ne sont donc pas toutes disponibles pour chaque client. Le tableau 1 présente une description de chaque variable utilisée pour tenter d'approximer le plus possible les revenus bruts annuels d'un consommateur.

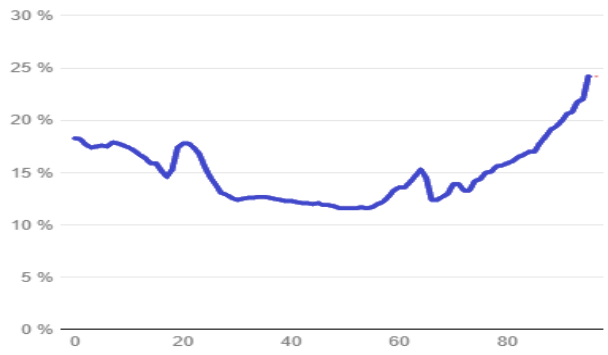
Tableau 1 : Description générale des variables utilisées

Variable	*Définition*
age	*Binaire sur l'âge*
ageRETRAITE	*Binaire si l'âge est de type retraite*
Age_OR	*Binaire si l'âge est de type OR*
BIN_NBRHYP	*Binaire sur nombre d'hypothèques*
NOUVEL_EMPLOI	*Binaire sur la croissance des revenus*
SEXE_Binaire	*Binaire sur le sexe*
Binaire_Epargne	*Binaire sur le montant $ en actif liquide*
NR_ANCIENS_REV	*Nombre de mois depuis la dernière déclaration de revenus*
wConversion_Brut	*Conversion des dépôts directs nets en revenus bruts mensuels winsorize*
wAnciens_REV	*Anciens revenus annuels déclarés winsorize*
wACHATS_ME	*Achats mensuels par carte de crédit winsorize*
logarit_limite	*Log de la limite sur la carte de crédit*
log_pmt_hyp	*Log du paiement hypothécaire mensuel*
logarithme_de	*Log des transactions par carte débit mensuel*

- **Âge**

Nous posons l'hypothèse qu'il existe une différence significative de revenus selon les classes d'âge. La figure 2 illustre les écarts existants parmi la population possédant un faible revenu. Nous avons préféré exclure les classes d'âge de 35 à 55 ans pour éviter des problèmes de colinéarité sachant que ces classes n'avaient pas de différences significatives de leurs revenus annuels moyens. La variable âge provient du formulaire d'application au crédit et sa décomposition est présentée au tableau 2.

Figure 2 : Évolution du pourcentage de la population à faibles revenus selon l'âge



Source : Statistique Canada 2016 (cité par Radio-Canada)

Tableau 2 : Description des variables « Âge »

Variable	*Description*
age25	*Client âgé de 18 à 25 ans*
age35	*Client âgé de 26 à 35 ans*
age65	*Client âgé de 55 à 65 ans*
Age_OR	*Client âgé de plus de 55 ans*
age_RETRAITE	*Client âgé de plus de 65 ans*

- **Nombre hypothèques**

Selon Statistique Québec, le revenu moyen d'un ménage étant propriétaire en 2011 se chiffrait à 69 600$ comparativement à 37 700$ pour un ménage étant locataire. Nous avons voulu exprimer cet impact possible sur les revenus à travers une variable binaire. Cette variable nous permet également de mieux approximer les clients possédant un revenu locatif. Le nombre d'hypothèques a été limité à 2 ou plus puisque l'échantillon de clients possédant plus de 2 hypothèques était seulement de 1.93%. La variable provient du formulaire d'application au crédit, plus précisément du bureau de crédit du demandeur.

$$BIN\_NBR\_HYP1 = \begin{cases} 1 & \text{le client possède 1 hypothèque au bureau de crédit} \\ 0 & \text{sinon} \end{cases}$$

$$BIN\_NBR\_HYP2 = \begin{cases} 1 & \text{le client possède 2 hypothèques ou plus au bureau de crédit} \\ 0 & \text{sinon} \end{cases}$$

- **Nouvel emploi**

En observant la distribution des dépôts directs dans les comptes débit des clients, nous avons convenu qu'il fallait rajouter une variable pour exprimer la croissance des dépôts des salaires au fil du temps. Une croissance supérieure à 40% entre le dépôt direct moyen mensuel 3 mois et le dépôt direct moyen mensuel 6 mois nous laisse croire que le client a obtenu un nouvel emploi récemment. Elle provient de l'information générale que la banque possède sur son client.

$$NOUVEL\_EMPLOI = \begin{cases} 1 & \text{la croissance du DI (Dépôt direct) moyen 3 mois est supérieure à 40\% comparativement au DI moyen 6 mois} \\ 0 & \text{sinon} \end{cases}$$

- **Sexe**

Cette variable permet d'observer si les hommes de notre échantillon déclarent des revenus annuels plus élevés que les femmes. Elle provient du formulaire d'application au crédit.

$$SEXE\_Binaire = \begin{cases} 1 & \text{le client est un homme} \\ 0 & \text{sinon} \end{cases}$$

- **Épargne**

L'hypothèse amenée est la suivante : Plus un client à des actifs liquides sous forme de placement avec la banque, plus son revenu devrait être élevé. Nous avons ainsi créé 4 différentes classes en fonction des actifs moyens à la banque lors des 6 derniers mois. La variable prend la forme d'une variable binaire et sépare les classes selon le conseiller financier attitré au client. En effet, il existe une hiérarchie à la banque en fonction de l'actif que l'on possède alors que différents conseillers peuvent nous être attitrés. La variable provient de l'information générale que la banque possède sur son client et sa décomposition est présentée au tableau 3. La variable « Binaire_Epargne1 » fut exclue des régressions pour éviter des problèmes de colinéarité.

Tableau 3 : Description des variables « Binaire_Epargne »

Variable	*Description*
Binaire_Epargne1	*Actif liquide de 0 à 3000$*
Binaire_Epargne2	*Actif liquide de 3000$ à 25 000$*
Binaire_Epargne3	*Actif liquide de 25 000$ à 100 000$*
Binaire_Epargne4	*Actif liquide supérieur à 100 000$*

- **Dépôt Direct du salaire (wConversion_Brut)**

Cette variable, de type continu, effectue la conversion des DI (Dépôts directs) du compte débit du client en revenus bruts mensuels. Le revenu net utilisé est le DI moyen mensuel des 6 derniers mois dans le compte chèque de la banque. En d'autres termes, on fait une moyenne des dépôts de paie mensuels en analysant les 6 derniers mois dans le compte chèque de la banque. La variable provient de l'information générale que la banque possède sur son client. Selon la méthodologie utilisée actuellement par l'institution financière, il existe 2 paliers pour convertir les revenus nets en revenus bruts. Un revenu inférieur à 2500$ net/mois est multiplié par 1.25 alors qu'un revenu supérieur à 2500$ net/mois est multiplié par 1.35. Cette différence de conversion est causée par le taux d'imposition qui varie selon les revenus.

Finalement, la technique de « Winsorize » à 2% est appliquée sur cette variable. Rappelons ici que « Winsorize » signifie l'ajustement de valeurs extrêmes en ajustant les percentiles 0, 1, 99 et 100 au 2^e et 98^e percentile. Cet ajustement nous permet de mieux capter la masse et sera discuté à la section 4 du rapport.

- **Nombre de mois depuis la dernière déclaration (NR_ANCIENS_REV)**

Comme mentionné précédemment, les données utilisées pour notre variable de performance sont celles de janvier 2017 à mars 2018. Toutefois, pour plusieurs clients, nous étions également en possession d'une ancienne déclaration de revenus pour une application faite entre janvier 2013 et février 2018.

Cette variable est donc une variable continue qui correspond au nombre de mois séparant les 2 déclarations de revenus. Elle provient de l'information générale que la banque possède sur son client.

- **Anciens revenus déclarés (wAnciens_REV)**

Variable continue qui représente une ancienne déclaration de revenus entre janvier 2013 et février 2018. Cette variable est disponible lorsqu'un demandeur effectue plusieurs applications au crédit avec la même institution financière dans un intervalle de 5 ans. Nous avons accès aux derniers revenus déclarés et pouvons les comparer avec ceux déclarés aujourd'hui. Si plusieurs demandes ont été faites dans les 5 dernières années pour un même client, l'avant-dernière demande est donc utilisée pour créer la variable. Cette variable provient des anciens formulaires d'application au crédit. La technique de « Winsorize » à 2% est également appliquée sur cette variable.

- **Achats mensuels moyens via la carte de crédit (wACHATS_ME)**

Cette variable correspond au volume moyen d'achats mensuels porté au compte de la carte de crédit de l'institution financière. Elle ne peut évidemment pas être calculée pour tous les clients, mais seulement pour ceux détenant une carte de crédit avec la banque de l'étude. Nous supposons que plus les dépenses sur une carte de crédit sont élevées, plus le demandeur devrait déclarer des revenus élevés. La moyenne mensuelle est établie selon les 12 derniers mois d'utilisation de la carte. La technique de « Winsorize » à 2% est également appliquée sur cette variable. Cette variable est obtenue grâce à l'information que possède la banque sur son client.

- **Limite sur la carte de crédit (logarit_limite) &**
  Paiement mensuel hypothèque (log_pmt_hyp) &
  Transactions moyennes mensuelles via la carte débit (logarithme_de)

Suite à l'analyse de la base de données, nous avons déterminé qu'il existait une relation de type logarithmique entre ces variables et les revenus d'un consommateur. Ces variables

proviennent de l'information que possède la banque sur son client. La carte de crédit, l'hypothèque et les transactions via le compte débit sont propres à l'institution financière de l'étude. C'est donc dire que ces variables ne sont pas calculées pour les clients ayant, par exemple, leur hypothèque dans une autre banque.

Les conclusions avancées par Finlay (2006) sont à l'origine de la création de ces variables. L'auteur fait un lien entre les dépenses mensuelles d'un client et les revenus qu'il gagne. Il prétend que plus le revenu d'un client est élevé, plus son train de vie sera élevé, tout comme ses dépenses.

Pour que la variable logarit_limite soit créée, le client doit posséder une carte de crédit avec une limite minimale de 300$ à l'institution financière à l'étude.

Pour que la variable log_pmt_hyp soit créée, le client doit posséder une hypothèque avec un paiement hypothécaire minimum mensuel de 200$ à l'institution financière à l'étude.

Pour que la variable logarithme_de soit créée, le client doit posséder un compte débit avec des dépenses mensuelles moyennes variant entre 500$ et 22 500$ à l'institution financière à l'étude.

Les différents seuils furent déterminés en observant les distributions des différentes variables. Toutes les variables présentées à la section 3.3 se retrouvent dans les régressions linéaires à la section 4.

4. Méthodologie et modèles

Tel que mentionné à la section 1.3, le modèle principal d'estimation des revenus se divisait en 4 sous-modèles : DI, MAJ, PMT et SOCIO. Chacun des sous-modèles comporte plusieurs conditions qu'il faut satisfaire pour que le modèle puisse évaluer les revenus du client. Ces conditions seront détaillées à la section 4.1. Un client peut alors se faire évaluer par 1 seul modèle, plusieurs modèles (voir ajustements section 5) ou encore n'être inclus dans aucun des modèles. Si le demandeur n'est pas évalué par l'un des 4 sous-modèles, il est exclu de la base de données telle que mentionnée précédemment à la section 3. En général, l'exclusion est causée par un manque de données pour évaluer le client. À noter que les sous-modèles MAJ, PMT et SOCIO se décomposent également en sous-modèles. Au total, 12 régressions linéaires sont utilisées pour créer le modèle final d'estimation.

Les sous-modèles DI, MAJ et PMT sont composés d'une variable principale et de quelques variables secondaires, alors que le modèle SOCIO est composé de plusieurs variables secondaires. Les variables principales expliquent la plupart des variations de revenus et ont été « Winsorize » telles que vues à la section 3, tout comme notre variable de performance des revenus déclarés. Ceci permet à la variable principale et la variable de performance d'être plus alignées. L'une des contraintes données par l'institution financière était que les modèles devaient utiliser le moins de variables possible pour éviter le phénomène de « Overfitting » des données. Ce phénomène se produit lorsque trop de variables plus ou moins pertinentes sont présentes dans une régression linéaire. Les résultats présentés à la section 4 sont donc les régressions linéaires optimisées avec les variables les plus pertinentes et significatives à inclure. De nombreux essais furent réalisés pour parvenir à la bonne combinaison de variables pour chaque régression.

Nous avons aussi conclu que le R^2 ajusté serait la principale observation pour juger de la qualité du modèle. De plus, une variable de mesure nommée « Efficacité du modèle » allait également permettre d'évaluer la performance de chaque modèle. Cette variable représente le pourcentage de clients pour lesquels nous avons effectué une bonne estimation de revenus. Pour que notre estimation soit bonne, elle doit être incluse dans un intervalle de ± 10 000$ ou ± 20% des revenus réels déclarés. Nous n'avons pas cherché à

observer si ces différences sommaient à 0 en raison de l'imprécision dans la base de données. À première vue, ces seuils peuvent paraître élevés. Toutefois, tel que mentionné précédemment, le but de l'étude n'est pas d'observer de petites variations entre les déclarations et les revenus réels, mais bien des écarts extrêmes qui sont dommageables pour la banque. Si un client déclare un revenu annuel de 50 000$ et que son revenu réel est de 42 000$, la décision de lui octroyer un financement ne devrait pas changer du tout au tout. Cependant, si un client déclare un revenu annuel de 50 000$ et que son revenu réel est de 20 000$, la décision d'octroi aurait fort probablement été différente. Ce sont ces écarts extrêmes qui sont davantage dommageables à la banque et qui causent une probabilité de défaut plus élevé sur les financements. Les seuils ont été fixés de sorte que pour un revenu réel de 50 000$ ou moins, nous sommes satisfaits d'être à ± 10 000$ avec notre modèle d'estimation. Pour un revenu réel supérieur à 50 000$, nous sommes satisfaits d'être à ± 20% avec notre estimation.

4.1 Régressions linéaires

4.1.1 Modèle DI (Dépôt Direct)

Le 1er modèle est un modèle qui estime les revenus mensuels bruts des clients possédant un dépôt direct à la banque. La variable wConversion_Brut agit à titre de variable principale pour expliquer les revenus réels, soit en les comparant aux DI moyens mensuels 6 mois convertis en revenus mensuels bruts. Voici les conditions à respecter pour que ce modèle évalue un client :

- *Le DI moyen 1 mois, 3 mois et 6 mois doivent être supérieurs à 430$ par mois (1)*
- *Le DI moyen 1 mois, 3 mois et 6 mois doivent être inférieurs à 13 000$ par mois (2)*
- *Le DI moyen 1 mois doit être inférieur à 2.5x le DI moyen 6 mois (3)*
- *Le DI moyen 6 mois doit être inférieur à 8x le DI moyen 1 mois (4)*

Les conditions sont utilisées sur les variables DI (provenant de la base de données originale). Les 2 premières conditions sont nécessaires pour aligner les revenus estimés avec les hypothèses 1 et 2 mentionnées à la section 3.2. Comme les revenus déclarés réels de l'échantillon sont bornés [6500$; 200 000$], il était nécessaire de borner la variable principale wConversion_Brut qui est créée à partir des variables DI. Les conditions 3 et 4

permettent d'éviter des variations extrêmes des revenus au courant des 6 derniers mois rendant impossible l'évaluation. La condition 3 permet entre autres d'éviter la situation d'un étudiant travaillant à temps partiel il y a 3 mois et qui change son statut à travailleur temps plein. Il serait impossible d'estimer ses revenus avec le modèle DI puisqu'empiriquement, la banque a observé que ce type de clients déclarait les revenus de son nouvel emploi plutôt que son revenu cumulatif de l'année courante. La condition 4 permet finalement d'éviter d'évaluer les pertes d'emplois durant la période observée.

Le modèle est alors présenté de sorte que :

$$Revenus\ Réels = B_0 + B_1 * wConversion\_Brut + B_2 * nouvel\_emploi + B_3 * age25 + B_4 * age35 + B_5 * BIN\_NBRHYP1 + B_6 * BIN\_NBRHYP2 + B_7 * Binaire\_Epargne2 + B_8 * Binaire\_Epargne3 + B_9 * Binaire\_Epargne4 + \varepsilon$$

Le tableau 4 présente les résultats d'estimations du modèle DI. À noter que pour des fins de simplification dans le triage des données, les résultats sont présentés en revenus bruts mensuels plutôt qu'annuels.

Tableau 4 : Régression linéaire avec modèle DI

Variable	***Coefficient***	***Erreur Standard***	***Valeur t***	***Pr > |t|***
(Intercept)	*763.9624*	*7.1544*	*106.78*	*<.0001*
nouvel_emploi	*893.2251*	*11.1200*	*80.33*	*<.0001*
age25	*-586.2377*	*7.3779*	*-79.46*	*<.0001*
age35	*-172.1085*	*5.5505*	*-31.01*	*<.0001*
BIN_NBRHYP1	*244.0811*	*5.5872*	*43.69*	*<.0001*
BIN_NBRHYP2	*428.4921*	*9.3851*	*45.66*	*<.0001*
Binaire_Epargne2	*125.1001*	*5.4854*	*22.81*	*<.0001*
Binaire_Epargne3	*249.7512*	*7.4806*	*33.39*	*<.0001*
Binaire_Epargne4	*522.5324*	*11.2912*	*46.28*	*<.0001*
wConversion_Brut	*0.9298*	*0.0015*	*621.91*	*<.0001*
R2 ajusté	*0.7577*			
Nombre d'observations	*205 426*			

Il est compréhensible d'observer toutes les variables significatives au seuil de 1% puisque le modèle présenté est un modèle optimisé utilisant uniquement les variables les plus pertinentes.

La variable principale wConversion_Brut obtient un coefficient positif de 0.9298. On peut interpréter cette valeur comme étant l'effet des revenus bruts mensuels moyens du client dans son compte bancaire sur ses revenus bruts mensuels réels. Une augmentation de 100$ des revenus bruts mensuels moyens observés dans le compte bancaire du client se traduit en une augmentation de 92.98$ de ses revenus bruts mensuels réels. Les revenus réels manquants sont donc observés par d'autres variables

Les hypothèses avancées à la section 3.3 sont confirmées. Plus l'épargne d'un client est élevée, plus ses revenus devraient l'être également. De plus, un demandeur possédant une hypothèque a en moyenne des revenus mensuels bruts plus élevé de 244$ qu'un demandeur étant locataire. Un demandeur possédant plus d'une hypothèque a des revenus mensuels moyens plus élevés de 428$, confirmant également l'hypothèse de présence de revenus locatifs.

L'âge du client est également important puisque les jeunes de 18 à 25 ans gagnent en moyenne 586$ de moins que les autres tranches d'âge, tout comme les jeunes de 25 à 35 ans qui gagnent en moyenne 172$ de moins. Ces résultats confirment également les résultats présentés à la figure 2. Finalement, avoir eu un nouvel emploi avec augmentation salariale de plus de 40% durant la période d'observation du dépôt direct contribue aussi à augmenter le revenu moyen mensuel de 893$.

Globalement, ce modèle présente des très bons résultats. Le R^2 ajusté est de 0.7577, prouvant la bonne performance du modèle pour estimer les revenus mensuels bruts. Malgré les conditions imposées, il est en mesure d'évaluer 205 426 clients (49% de notre échantillon). La mesure « Efficacité du modèle » indique que **73.65%** des revenus estimés se trouvent dans l'intervalle de ±10 000$ ou ± 20% des revenus réels. Rappelons ici que le segment de 26.35% qui n'est pas inclus dans l'intervalle n'est pas nécessairement causé par des lacunes du modèle. L'échantillon de 205 426 clients peut contenir de mauvaises déclarations de revenus réels ayant pour effet de présenter notre modèle comme étant moins efficace qu'il ne l'est réellement.

4.1.2 Modèle MAJ (Mise à jour)

Le second modèle détermine les revenus bruts annuels des clients à partir d'une ancienne déclaration survenue entre janvier 2013 et février 2018. Le modèle MAJ (Mise à jour) se divise en 5 sous-modèles en fonction de la variable NR_ANCIENS_REV. Les intervalles sont de 0 à 6 mois entre l'ancienne déclaration et la nouvelle, 7 à 12 mois, 13 à 24 mois, 25 à 36 mois et 37 mois ou plus. La variable wAnciens_REV agit donc à titre de variable principale en expliquant les revenus réels déclarés récemment par des revenus déclarés dans une autre demande de crédit. Nous avons conjointement décidé de ne pas créer un modèle pour les déclarations survenues dans l'intervalle 0 à 6 mois. Normalement, les revenus déclarés à travers une si courte période ne devraient pas varier à moins d'un changement d'emploi. Aux fins de simplification du modèle, nous avons préféré ne pas tenir compte d'un possible changement d'emploi lors de l'évaluation des revenus entre 2 déclarations. Un client peut ainsi être évalué par 1 seul des sous-modèles, soit en fonction du nombre de mois s'étant écoulé entre ses 2 déclarations. Voici les conditions à respecter pour qu'un client soit évalué selon le modèle MAJ :

- *Les anciens revenus déclarés doivent être supérieurs ou égaux à 6500$ annuel (1)*
- *Les anciens revenus déclarés doivent être inférieurs ou égaux à 200 000$ annuel (2)*

Les 2 conditions utilisées sont nécessaires pour aligner la variable principale wAnciens_REV avec la variable de performance des revenus réels sachant que cette dernière est bornée [6500$; 200 000$].

Chacun des sous-modèles est alors présenté de sorte que :

$$Revenus\ Réels = B_0 + B_1 * age25 + B_2 * Age\_OR + B_3 * wAnciens\_REV + \varepsilon$$

Le tableau 5 présente les résultats des 4 sous-modèles MAJ. À noter que les résultats pour les anciens revenus déclarés 0 à 6 mois sont présentés dans le tableau 6 tout comme la variable « Efficacité du modèle ».

Tableau 5 : Régression linéaire avec modèle MAJ

Variable	*Modèle MAJ 7 à 12 mois*	*Modèle MAJ 13 à 24 mois*	*Modèle MAJ 25 à 36 mois*	*Modèle MAJ 37 mois et plus*
(Intercept)	*6406 (0.0001)****	*12706 (0.0001)***	*15717 (0.0001)****	*18639 (0.0001)****
age25	*-2160.60 (0.0001)****	*-3736.17 (0.0001)****	*-3713.88 (0.0001)****	*-4412.53 (0.0001)****
Age_OR	*-1450.57 (0.0001)****	*-2996.41 (0.0001)****	*-4425.33 (0.0001)****	*-7909.12 (0.0001)****
wAnciens_REV	*0.9457 (0.0001)****	*0.8564 (0.0001)****	*0.8177 (0.0001)****	*0.80629 (0.0001)****
R2 ajusté	*0.8352*	*0.6968*	*0.6214*	*0.5642*
Nombre d'observations	*60 653*	*77 847*	*46 283*	*46 810*

Tableau 6 : Efficacité modèle MAJ

Variable	*0 à 6 mois* ***SANS MODÈLE***	*MAJ 7 mois à 12 mois*	*MAJ 13 à 24 mois*	*MAJ 25 mois à 36 mois*	*MAJ 37 mois et plus*
Efficacité du modèle	*83.17%*	*80.49%*	*72.40%*	*67.80%*	*61.51%*
Nombre d'observations	*114 859*	*60 653*	*77 847*	*46 283*	*46 810*

On peut affirmer que le modèle MAJ est plus simpliste que le modèle DI présenté à la section 4.1.1 et confirme les hypothèses avancées à la section 3.3. Ceci s'explique par le fait que les anciens revenus déclarés servent de bons estimateurs des revenus réels d'un consommateur et ne nécessite pas beaucoup de variables pour les accompagner. Encore une fois, seules les variables les plus significatives sont présentées.

On observe que le coefficient de wAnciens_REV diminue de 0.9457 à 0.80629 au fil des modèles. Ceci s'explique par le fait que les anciens revenus déclarés sont plus représentatifs des revenus d'aujourd'hui s'ils ont été déclarés à court terme (94 .57% des

revenus déclarés il y a 7 à 12 mois) plutôt qu'à long terme (80.629% des revenus déclarés il y a 37 mois et plus). La valeur du Intercept grossit pour contrebalancer le fait que les anciens revenus sont moins représentatifs. Comme le modèle MAJ est segmenté selon le temps via la variable NR_ANCIENS_REV, la valeur du intercept capte également les phénomènes d'inflation et d'augmentation salariale. Plus le temps s'étant écoulé entre les déclarations est élevé, plus il est probable que des augmentations salariales aient contribué à faire croître les revenus jusqu'à aujourd'hui. Il est donc légitime de voir le coefficient du Intercept augmenter autant pour contrebalancer le coefficient de wAnciens_REV, autant pour capter l'augmentation des salaires au fil du temps.

Les 2 variables d'âge présentent des coefficients significatifs. Comme vu précédemment, ces 2 groupes démontrent généralement des revenus moins élevés que le reste de la population. La valeur négative s'explique par le coefficient du intercept qui est trop élevé pour compenser celui de wAnciens_REV. Prenons l'exemple d'un jeune de 19 ans ayant déclaré des revenus de 10 000$ il y a 1 an. Sans la variable âge, le modèle supposerait que ses revenus aujourd'hui seraient de 0.9457*10 000$ + 6406$ = 15 953$. La croissance est trop élevée pour ce groupe d'âge et nécessite un ajustement de -2160.60$, diminuant ainsi les nouveaux revenus estimés à 13 792.40$ (15 953$ - 2160.60$).

Globalement, ce modèle présente de bons résultats. Évidemment, le R2 ajusté est plus élevé pour un modèle court terme (0.8352) qu'un modèle d'estimation long terme (0.5642). Il permet d'évaluer une grande proportion de l'échantillon, soit 346 452 clients correspondant à 83%. La mesure d'efficacité du modèle suit la tendance du coefficient R2 ajusté et présente un résultat total de **75.18%** pour le modèle MAJ. À titre informatif, si nous avions uniquement utilisé les anciens revenus déclarés sans modèle, la mesure d'efficacité aurait été de 72.67%. Segmenter à travers le temps en ajoutant la variable âge permet ainsi de gagner en efficacité. L'efficacité de 83.17% du segment 0 à 6 mois nous permet d'observer clairement le bruit présent à travers notre échantillon. Il est illogique de constater que 16.83% des revenus déclarés il y a moins de 6 mois ne sont pas dans un intervalle de ± 10 000$ ou ± 20% des revenus déclarés aujourd'hui. Certes, des pertes d'emplois, des promotions ou encore le changement d'un statut travailleur temps partiel à temps plein peut avoir causé d'aussi grandes variations. Toutefois, nous supposons

également qu'il y ait eu confusion dans plusieurs déclarations, contribuant ainsi à limiter la précision que l'on peut atteindre avec la base de données.

4.1.3 Modèle PMT (Achats sur la carte de crédit)

Le troisième modèle permet l'estimation des revenus mensuels bruts des demandeurs à partir des achats mensuels moyens qu'ils effectuent sur leur carte de crédit de la banque à l'étude. À noter qu'un client peut posséder plusieurs cartes de crédit à la banque, mais c'est la carte la plus utilisée qui est retenue pour le modèle. Les cartes de crédit que possède le client à d'autres banques que celle à l'étude ne sont pas utilisés à travers ce modèle. La variable wACHATS_ME agit à titre de variable principale pour expliquer les revenus réels. Ce modèle d'évaluation de train de vie du consommateur se décompose également en 4 sous-modèles. Certains détenteurs de carte de crédit possèdent un compte bancaire à l'institution financière et/ou une hypothèque. La banque est donc en possession de plus d'informations sur le niveau de dépenses d'un individu, plus précisément ses transactions débits et ses paiements hypothécaires mensuels. Voici donc les conditions à satisfaire pour que le modèle PMT évalue le client :

- *Les achats mensuels moyens sur carte de crédit sont supérieurs à 200$ (1)*
- *Les achats mensuels moyens sur carte de crédit sont inférieurs à 10 000$ (2)*
- *La principale carte de crédit utilisée doit être de type régulière (3)*
- *Le client n'est pas un travailleur autonome (4)*

Les conditions 1 et 2 sont nécessaires pour borner la variable wACHATS_ME et l'aligner avec les revenus réels bornés [6500$; 200 000$]. On pose l'hypothèse qu'un client effectuant moins de 200$ en achats mensuels moyens sur sa carte de crédit doit probablement utiliser une carte de crédit d'une autre institution financière en tant que carte principale. À l'inverse, un client effectuant plus de 10 000$ en achats mensuels moyens sur sa carte démontre un rythme de dépenses trop élevé qui s'apparente à une utilisation commerciale de la carte de crédit. Il devient non-pertinent d'estimer des revenus personnels à partir d'une utilisation commerciale sur un produit de crédit. La condition 3 était nécessaire pour éliminer les cartes de crédits à usage restreint de l'échantillon puisque le niveau de dépenses est également non-représentatif. La condition 4 rejoint l'hypothèse avancée avec la condition 2, soit que les travailleurs autonomes ont tendance à démontrer

une utilisation commerciale de leur carte de crédit personnelle. La banque est au courant du statut d'emploi du client en raison des questions actuellement posées sur les formulaires d'application au crédit.

Voici les 4 sous-modèles qui constituent le modèle PMT. Les principales différences sont que le sous-modèle 1 évalue Carte-Hypothèque-Débit, le modèle 2 Carte-Hypothèque, le modèle 3 Carte-Débit et le modèle 4 la Carte. Le tableau 7 présente les résultats des régressions linéaires alors que le tableau 8 présente les mesures d'efficacité.

$$Revenus\ Réels\ (1) = B_0 + B_1 * log\_pmt\_hyp + B_2 * logarit\_limite + B_3 * logarithme\_de + B_4 * age25 + B_5 * age35 + B_6 * Age\_OR + B_7 * SEXE\_Binaire + B_8 * wACHATS\_ME + \varepsilon$$

$$Revenus\ Réels\ (2) = B_0 + B_1 * log\_pmt\_hyp + B_2 * logarit\_limite + B_3 * age25 + B_4 * age35 + B_5 * Age\_OR + B_6 * SEXE\_Binaire + B_7 * wACHATS\_ME + \varepsilon$$

$$Revenus\ Réels\ (3) = B_0 + B_1 * logarit\_limite + B_2 * logarithme\_de + B_3 * age25 + B_4 * age35 + B_5 * Age\_OR + B_6 * SEXE\_Binaire + B_7 * wACHATS\_ME + \varepsilon$$

$$Revenus\ Réels\ (4) = B_0 + B_1 * logarit\_limite + B_2 * age25 + B_3 * age35 + B_4 * Age\_OR + B_5 * SEXE\_Binaire + B_6 * wACHATS\_ME + \varepsilon$$

Tableau 7 : Régression linéaire avec modèle PMT

Variable	*Modèle 1 (HYP+DEBIT)*	*Modèle 2 (HYP)*	*Modèle 3 (DEBIT)*	*Modèle 4 (De base)*
(Intercept)	*-10702 (0.0001)****	*-6482 (0.0001)****	*-6950 (0.0001)****	*-781 (0.0001)****
log_pmt_hyp	*878.27 (0.0001)****	*1008.27 (0.0001)****	.	.
logarit_limite	*275.97 (0.0001)****	*473.05 (0.0001)****	*254.77 (0.0001)****	*582.26 (0.0001)****
logarithme_de	*711.78 (0.0001)****	.	*981.74 (0.0001)****	.
age25	*-677.01 (0.0001)****	*-1307.07 (0.0001)****	*-904.41 (0.0001)****	*-1729.65 (0.0001)****
age35	*-328.59 (0.0001)****	*-817.92 (0.0001)****	*-353.32 (0.0001)****	*-662.91 (0.0001)****
Age_OR	*-667.47 (0.0001)****	*-474.87 (0.0001)****	*-907.80 (0.0001)****	*-1184.53 (0.0001)****
SEXE_Binaire	*761.55 (0.0001)****	*852.14 (0.0001)****	*541.10 (0.0001)****	*853.03 (0.0001)****
wACHATS_ME	*0.4802 (0.0001)****	*0.3998 (0.0001)****	*0.5285 (0.0001)****	*0.5695 (0.0001)****
R2 ajusté	*0.3320*	*0.2886*	*0.4520*	*0.3340*
Nombre d'observations	*63 927*	*11 553*	*125 404*	*5 767*

Tableau 8 : Efficacité modèle PMT

Variable	*Hyp+Debit*	*Hyp*	*Debit*	*De base*
Efficacité du modèle	*47.4%*	*44.1%*	*50.5%*	*39.2%*
Nombre d'observations	*63 927*	*11 553*	*125 404*	*5 767*

Le modèle PMT est un peu plus complexe que les modèles précédents puisqu'il utilise plusieurs variables pour tenter d'approximer la consommation et les dépenses d'un client. Il y a une gradation dans les modèles alors qu'il est plus difficile d'estimer les revenus d'un client possédant uniquement une carte de crédit, plutôt qu'un client possédant son hypothèque, son compte débit et une carte de crédit avec la même banque. Encore une fois, ce sont des modèles optimisés qui sont présentés, donc toutes les variables sont pertinentes et significatives.

Le log du paiement hypothécaire, le log de la limite de carte de crédit et le log du volume mensuel des transactions débits ont tous des relations positives avec les revenus estimés. On observe également l'apparition de la variable SEXE_Binaire confirmant que les hommes déclarent des revenus plus élevés en moyenne que les femmes. Les coefficients des âges sont conformes aux résultats obtenus dans les autres régressions, soit que les classes extrêmes génèrent des revenus plus faibles que les autres classes d'âge. Lorsqu'on compare les modèles 1 et 3, qui évaluent le plus grand nombre de clients, on constate que le paiement hypothécaire est une variable importante dans le niveau de consommation d'un client. La formation de cette variable à travers le modèle 1 cause une baisse des coefficients des variables wACHATS_ME et logarithme_de du modèle 3. Même si le R2 ajusté du modèle 1 est plus petit, l'ajout du paiement hypothécaire agit comme variable de contrôle qui était précédemment incluse en partie dans d'autres variables évaluant les dépenses d'un client. On obtient ainsi un meilleur portrait de la consommation moyenne d'un client.

Globalement, le modèle PMT performe moins bien que ceux vu précédemment. Les R2 ajustés varient de 0.2886 à 0.4520, et sont bien inférieurs au 0.7577 du modèle DI par exemple. Il permet d'évaluer 206 651 clients, soit environ 50% de notre échantillon. L'efficacité du modèle semble également suivre la tendance du R2 ajusté alors qu'elle n'est que de **49%**. À titre informatif, un test sur les travailleurs autonomes exclus via la

condition 4 fut réalisé. L'efficacité de ce segment de travailleurs n'était que de 30%, justifiant ainsi leur exclusion du modèle PMT. Une des raisons pouvant expliquer la sous-performance du modèle PMT est que ce n'est pas tous les consommateurs qui utilisent la carte de crédit de la même façon. Certains l'utilisent comme substitut à la carte débit et d'autres comme complément. Quelques clients possèdent également des cartes de crédit dans plusieurs banques et répartissent leur consommation sur plusieurs cartes. D'autres vont dépenser plus que leurs revenus mensuels sur la carte de crédit et s'endetter progressivement. Finalement, quelques clients vont uniquement l'utiliser pour des achats sur Internet. La difficulté à connaître les préférences d'utilisation d'un consommateur justifie le R2 ajusté et la mesure d'efficacité moins élevés. Un portrait plus global de la consommation totale d'un client nous permettrait d'estimer ses revenus plus efficacement.

4.1.4 Modèle SOCIO (Socio-démographique)

Ce quatrième et dernier modèle devait originalement permettre d'évaluer les revenus de clients basés uniquement sur des variables socio-démographiques. Quelques ajustements furent toutefois nécessaires pour que le modèle soit pertinent et utile pour l'analyse. À titre d'élément différenciateur des autres modèles, celui-ci n'inclus pas de variables principales. Il se compose essentiellement de variables secondaires qui servent à estimer le revenu mensuel brut des clients pour lesquels la banque possède peu d'information à leur égard. Voici les conditions à respecter pour se faire évaluer par ce modèle :

- *Le client ne doit pas avoir été évalué par l'un des 3 autres modèles (1)*
- *Le client doit posséder un compte débit avec la banque à l'étude (2)*

La condition 1 est importante puisqu'elle est différente des principes appliqués précédemment. Un client peut se faire évaluer par plusieurs combinaisons de modèles parmi DI, MAJ et PMT. Or, s'il se fait évaluer par le modèle SOCIO, il ne peut pas avoir été évalué ailleurs. Ce modèle est utilisé en dernier recours pour les clients ne satisfaisant pas aux conditions des autres modèles. Nous posons l'hypothèse qu'évaluer un client principalement sur des variables socio-démographiques ne permet pas une analyse aussi précise que celle des autres modèles. Nous préférons ainsi garder une évaluation SOCIO uniquement pour les clients envers lesquels nous sommes en manque d'information plutôt qu'utiliser ce modèle à travers toute notre base de données. La condition 2 a été rajoutée

pour améliorer la performance du modèle SOCIO. Nous allons voir ci-bas qu'elle était nécessaire pour obtenir des résultats pertinents.

Le modèle SOCIO se divise donc en 2 sous modèles. Le premier utilise les transactions débits moyennes mensuelles d'un client, alors que le deuxième inclut son nombre d'hypothèques présents sur le bureau de crédit. Le tableau 9 présente le résultat des 2 sous-modèles de régressions et le tableau 10 l'efficacité des modèles.

$$Revenus\ Réels\ (1) = B_0 + B_1 * logarithme\_de + B_2 * Bin.Ep.2 + B_3 * Bin.Ep.3 + B_4 * Bin.Ep.4 + B_5 * age25 + B_6 * age35 + B_7 * age65 + B_8 * ageRETRAITE + B_9 * SEXE\_Binaire + \varepsilon$$

$$Revenus\ Réels\ (2) = B_0 + B_1 * Nbr\_hyp1 + B_2 * Nbr\_hyp2 + B_3 * Bin.Ep.3 + B_4 * Bin.Ep.4 + B_5 * age25 + B_6 * age35 + B_7 * age65 + B_8 * ageRETRAITE + B_9 * SEXE\_Binaire + \varepsilon$$

Le signe des coefficients obtenus concorde avec les résultats des modèles DI, MAJ, PMT et nos hypothèses. Le nombre d'hypothèques et l'épargne contribuent à augmenter le revenu estimé. À l'inverse, les jeunes de 35 ans et moins et les adultes de 55 ans et plus contribuent à diminuer l'estimation de revenu annuel. Encore une fois, le nombre d'hypothèques et le log des transactions mensuelles moyennes via carte de débit confirment l'hypothèse que le train de vie d'un consommateur (ses dépenses) augmente avec ses revenus puisque le coefficient de ces variables est positif. Le modèle SOCIO permet d'évaluer 28 872 clients (7% de l'échantillon). Il démontre le R2 ajusté le plus faible des différents modèles, soit variant de 0.2019 à 0.3114. Ceci s'explique entre autres par l'absence de variable principale pour estimer les revenus : ils sont estimés par des variables très générales et peu personnalisées. Le sexe et l'âge d'un client ne sont pas suffisants pour bien estimer les variations de revenus réels.

Le tableau 10 démontre toutefois une mesure d'efficacité moyenne de **52%,** soit comme étant plus performant que le modèle PMT. On constate une tendance parmi les clients ne s'étant pas fait évaluer par les autres modèles. C'est comme si les gens n'ayant pas satisfait aux conditions des autres modèles, mais réussi les conditions du modèle SOCIO, avaient un profil similaire. Après révisions de ce sous-segment, nous avons constaté qu'il était composé essentiellement d'adultes de plus de 65 ans à la retraite. L'estimation de leurs revenus (en raison de leur âge) ne nécessitait donc pas de bien expliquer les variations des

revenus réels via le R2 ajusté puisqu'ils gagnent presque tous un revenu similaire à la retraite.

Finalement, la condition 2 énoncée précédemment n'était pas l'une des conditions originales imposées au modèle. Nous avons également tenté d'estimer les revenus des non-clients de l'institution financière qui appliquaient pour du crédit. Le tableau 10 démontre toutefois notre échec quant à la pertinence du modèle. Seulement 35.6% de nos estimations pour cet échantillon de 33 604 demandeurs se positionnaient dans l'intervalle de ± 10 000$ ou ± 20% des revenus réels. Suite à quelques discussions avec l'équipe de gestion des risques, nous avons préféré éliminer ces demandeurs de notre base de données. Ainsi, nous avons modifié l'utilité possible de notre modèle d'estimation des revenus en supposant qu'il était maintenant impossible d'approximer avec précision les revenus d'un demandeur de crédit étant non-client de l'institution financière. Cette exclusion est principalement causée par l'absence de données sur ces demandeurs, augmentant la difficulté à bien estimer leurs revenus.

Tableau 9 : Régression linéaire avec modèle SOCIO

Variable	Modèle 1(Débit)	Modèle 2 (Nbr Hyp)
(Intercept)	*-1232.94* *(0.0001)****	*3717.96* *(0.0001)****
BIN_NBRHYP1	.	*584.96* *(0.0001)****
BIN_NBRHYP2	.	*1554.09* *(0.0001)****
logarithme_debit	*497.59* *(0.0001)****	.
Binaire_Epargne2	*94.24* *(0.0001)****	.
Binaire_Epargne3	*137.37* *(0.0001)****	*489.89* *(0.0001)****
Binaire_Epargne4	*382.62* *(0.0001)****	*1628.74* *(0.0001)****
age25	*-1011.50* *(0.0001)****	*-2571.85* *(0.0001)****
age35	*-271.18* *(0.0001)****	*-721.73* *(0.0001)****
age65	*-271.84* *(0.0001)****	*-423.55* *(0.0001)****
ageRETRAITE	*-771.50* *(0.0001)****	*-1528.15* *(0.0001)****
SEXE_Binaire	*494.41* *(0.0001)****	*613.85* *(0.0001)****
R2 ajuté	0.2019	0.3114
Nombre d'observations	22 798	6 074

Tableau 10 : Efficacité modèle SOCIO

Variable	*Debit*	*Nbr Hyp*	*Rien*
Efficacité du modèle	*56.48%*	*47.74%*	*35.60%*
Nombre d'observations	*22 798*	*6 074*	*33 604*

4.2 Résumé de l'estimation des revenus selon le score des modèles

En résumé, le modèle d'estimation des revenus est essentiellement composé de 4 sous-modèles, qui se divisent également en plusieurs petits modèles. La figure 3 permet d'illustrer le schéma d'évaluation des revenus annuels pour les clients. Le demandeur peut se faire évaluer par 1, 2 ou 3 modèles (DI, MAJ, PMT) tout dépendant des conditions d'évaluation auxquels il satisfait et des produits qu'il détient avec la banque. S'il ne s'est pas fait tarifer par aucun des 3 modèles, il se fait évaluer par l'un des sous-modèles du modèle SOCIO à condition qu'il respecte les conditions de ce modèle. Il nous est toutefois impossible d'évaluer les revenus d'un nouveau client qui ne possède rien du tout avec l'institution financière puisque ce client ne satisfait à aucune des conditions mentionnées à la section 4.1. Il est important de rappeler que notre variable de revenus réels est bornée [6500$; 200 000$]. Lorsqu'une de nos estimations sortait de ces bornes, elle était automatiquement ramenée à la borne minimale/maximale la plus proche. Par exemple, si le modèle DI estimait un revenu annuel brut de 4000$, l'approximation était automatiquement ajustée à 6500$ pour être conforme à la borne sur la variable de performance.

Le tableau 11 résume également les mesures « Efficacité du modèle » trouvées à la section 4.1. L'efficacité globale obtenue de 72% est calculée en supposant aucune interaction entre les modèles. Elle est plutôt obtenue en hiérarchisant les modèles. Un client se fait évaluer en 1er par les modèles MAJ 0 à 6 mois et MAJ 7 à 12 mois (efficacité > 80%), si l'information n'est pas disponible, il se fait évaluer par le modèle DI (efficacité = 74%). Si l'information n'est toujours pas disponible, il se fait évaluer par le reste du modèle MAJ (efficacité < 73%), ensuite par le modèle PMT (efficacité = 49%) et finalement par le modèle SOCIO (efficacité = 52%). Ceci a pour but d'évaluer les revenus d'un client à partir des modèles les plus efficaces.

Nous sommes globalement satisfaits des résultats puisqu'à cette étape, notre modèle est en mesure d'estimer 72% des revenus annuels bruts des clients dans un intervalle de ± 10 000$ ou ± 20% de leurs revenus réels. La section 5 va présenter une technique d'optimisation pour tenter d'améliorer davantage cette mesure d'efficacité.

Figure 3 : Résumé des différents modèles d'évaluation

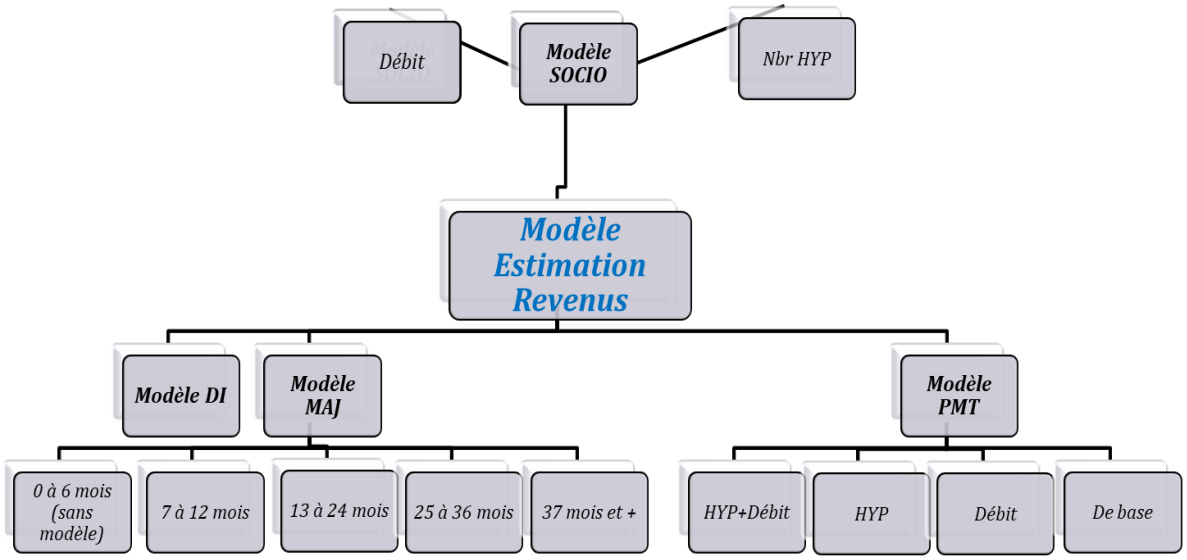


Tableau 11 : Résumé de « Efficacité du modèle »

Modèle	*Efficacité*
DI	*74%*
MAJ	*75%*
PMT	*49%*
SOCIO	*52%*
TOUS	***72%***

5. Optimisation

Les résultats présentés à la section 4.2 ont créé une hiérarchie selon l'efficacité des modèles pour déterminer lequel serait utilisé sur un client s'étant fait évaluer par plusieurs modèles. Cette technique se veut être un peu simpliste puisqu'aucune interaction est créée entre les modèles. La banque ne bénéficie pas de l'étendue des informations qu'elle possède sur son client. Or, une synergie entre les différents modèles pourrait permettre d'améliorer l'efficacité globale. Un plus grand nombre de consommateurs se retrouveraient dans l'intervalle ± 10 000$ ou ± 20% entre l'estimation et les revenus réels. Pour créer cette dynamique, nous avons eu recours à un modèle sous forme de plan cartésien. L'idée d'utiliser une moyenne entre les différents scénarios fut rejetée puisqu'ils ne performent pas tous de la même façon.

5.1 Plan cartésien

Le plan cartésien consiste à essayer toutes les combinaisons de scénarios possibles jusqu'à ce que 1 variable soit optimisée. Dans notre situation, la variable à optimiser était la mesure « Efficacité du modèle ». Nous voulions maximiser le nombre de clients pour lesquels nous effectuons une bonne approximation. Prenons l'exemple d'un client se faisant évaluer par les modèles DI et MAJ 0 à 6 mois. Le plan cartésien va générer toutes les combinaisons possibles de synergie entre les 2 modèles de sorte que :

$$Revenus\ estimés = 0.01 * Revenus\_Modele\_DI + 0.99 * Revenus\_Modele\_MAJ\_0À6\_MOIS$$

$$Revenus\ estimés = 0.02 * Revenus\_Modele\_DI + 0.99 * Revenus\_Modele\_MAJ\_0À6\_MOIS$$

$$\ldots$$

$$Revenus\ estimés = 0.99 * Revenus\_Modele\_DI + 0.01 * Revenus\_Modele\_MAJ\_0À6\_MOIS$$

Ces combinaisons sont testées jusqu'à ce que le plan cartésien détermine les bons poids à utiliser sur chaque modèle pour ainsi maximiser le nombre de bonnes estimations de revenus. La figure 4 présente un exemple concret du cheminement.

Figure 4 : Exemple de plan cartésien

	Revenus_Pr...	Revenus_Pr...	Revenus_Pr...	_FREQ_	variance1_S...	Good_Mean	stand_dev
1	0.3	0.69	0.01	8087	1.3543618E12	0.7437863237	12940
2	0.31	0.69	0	8087	1.353273E12	0.7435390132	12940
3	0.29	0.7	0.01	8087	1.3561344E12	0.743415358	12950
4	0.31	0.68	0.01	8087	1.352904E12	0.743415358	12930
5	0.34	0.66	0	8087	1.350257E12	0.743415358	12920
6	0.36	0.64	0	8087	1.3498198E12	0.7432917027	12920
7	0.3	0.7	0	8087	1.3549078E12	0.7431680475	12940
8	0.33	0.67	0	8087	1.3509476E12	0.7431680475	12920
9	0.35	0.65	0	8087	1.349881E12	0.7431680475	12920
10	0.29	0.68	0.03	8087	1.3558378E12	0.7430443922	12950
11	0.29	0.69	0.02	8087	1.3557946E12	0.7430443922	12950
12	0.29	0.71	0	8087	1.3568572E12	0.742920737	12950
13	0.39	0.6	0.01	8087	1.3525709E12	0.742920737	12930
14	0.39	0.61	0	8087	1.3515244E12	0.742920737	12930
15	0.3	0.68	0.02	8087	1.354199E12	0.7427970817	12940
16	0.35	0.64	0.01	8087	1.3502198E12	0.7427970817	12920
17	0.4	0.6	0	8087	1.352722E12	0.7426734265	12930
18	0.36	0.63	0.01	8087	1.3503355E12	0.7425497712	12920
19	0.32	0.68	0	8087	1.3519529E12	0.742426116	12930
20	0.33	0.64	0.03	8087	1.3520514E12	0.742426116	12930
21	0.38	0.62	0	8087	1.3506415E12	0.742426116	12920

Scénario : DI + MAJ 25 à 36 mois + PMT Débit

Le scénario d'un client se faisant évaluer par 3 modèles est présenté, soit par les modèles DI, MAJ 25 à 36 mois et PMT Débit. La variable Good_Mean représente notre « Efficacité du modèle ». Pour ce scénario précis, le plan cartésien suggère d'attribuer un poids de 30% au modèle DI, 69% au modèle MAJ et 1% au modèle PMT. Cette distribution maximise notre pourcentage de bonnes estimations à 74.38%. On observe ici un fait intéressant, ce n'est pas le même poids qui est attribué à chaque modèle. Comme vu précédemment, le modèle PMT semble sous-performer comparativement aux autres modèles. Lorsqu'on créé de la synergie avec les autres modèles, les effets de cette sous-performance sont d'autant plus visibles puisque le modèle est négligé en présence de modèles plus performants.

Au total, 12 sous modèles sont utilisés pour créer le modèle final d'estimation des revenus. On utilise 1 modèle de DI, 5 sous-modèle de MAJ, 4 sous-modèle de PMT et 2 sous-modèles de SOCIO tel qu'illustré à la figure 3. Il existe alors 53 différentes combinaisons/scénarios parmi lesquelles un client peut avoir été évalué. L'annexe 1 présente les poids obtenus pour les 53 différents scénarios. Il est intéressant de constater que certains scénarios servent à évaluer beaucoup plus de clients que d'autres.

5.2 Synergie entre les modèles

Suite à l'optimisation des poids à attribuer sur chaque modèle, il fut possible de constater que certaines combinaisons de modèles généraient des efficacités beaucoup plus élevées que celles observées précédemment. Le figure 5 reproduit les résultats présentés à l'annexe 1 en démontrant les combinaisons ayant tendance à surperformer. On observe ici que nous sommes en mesure d'effectuer une bonne estimation pour 87.8% des clients se faisant évaluer par 3 modèles (DI, MAJ 0 à 6 mois et PMT). À l'opposé, l'échantillon de clients se faisant uniquement évaluer par le modèle PMT (efficacité 38.9%) et par le modèle SOCIO (efficacité 54.8%) performent moins bien. Nous pouvons aussi constater que le temps écoulé entre 2 mises à jour de revenus est très pertinent pour approximer les revenus réels. Toutes les combinaisons qui incluent une ancienne mise à jour de revenus survenue il y a moins de 12 mois surperforment comparativement aux autres combinaisons.

Ces résultats confirment que plus une banque est en possession d'informations de qualité sur son client, plus elle est en mesure de bien estimer ses revenus réels. Finalement, la synergie créée entre les différents modèles permet d'augmenter l'efficacité globale de notre modèle d'estimation des revenus de 72% à **74%**.

Figure 5 : Efficacité selon la synergie entre les modèles

Nom des modèles	Efficacité	Nbr de données	Efficacité	Nom des modèles	Nbr de données
3 Modèles : DI ; MAJ 6 mois; PMT	87,8%	43502	69,1%	DI et PMT	11570
3 Modèles : DI ; MAJ 12 mois; PMT	85,7%	22288	74,7%	DI SEUL	16085
3 Modèles : DI ; MAJ 24 mois; PMT	79,9%	25029	82,2%	MAJ SEUL0à6	30089
3 Modèles : DI ; MAJ 36 mois; PMT	76,7%	13162	65,8%	MAJ SEUL +6	71591
3 Modèles : DI ; MAJ 37 mois; PMT	74,7%	11082	38,9%	PMT SEUL	13015
			54,8%	Socio Seul	28872
2 modèles DI et MAJ 6 mois	87,5%	20 045			
DI et MAJ 12 mois	85,2%	10962			
DI et MAJ 24 mois	79,5%	14473			
DI et MAJ 36 mois	78,6%	8848			
DI et MAJ 37 mois	74,9%	8380			
PMT ET MAJ 6 MOIS	82,1%	21223			
PMT ET MAJ 12 MOIS	76,8%	11 902			
PMT ET MAJ 24 MOIS	66,7%	15 450			
PMT ET MAJ 36 MOIS	60,3%	9095			
PMT ET MAJ 37 MOIS	55,1%	9331			

6. Applications et recommandations

L'objectif premier de ce projet était que la banque puisse avoir sa propre évaluation des revenus du demandeur sur une application au crédit. De cette façon, lorsque l'écart entre le revenu estimé par la banque et celui déclaré par le client serait trop élevé, davantage de vérifications seraient faites avant d'octroyer du crédit. De plus, comme chaque modèle se base sur des variables principales étant régulièrement mises à jour pour l'institution financière à l'étude, nous sommes en mesure d'évaluer les revenus de façon continue. Il n'est pas nécessaire de recevoir une application pour du crédit afin d'évaluer les revenus d'un client. Une des solutions proposées pour évaluer l'écart entre le revenu estimé et le revenu réel consiste à créer un intervalle de confiance à l'entour des revenus estimés tel qu'il sera discuté à la section 6.1.

À noter qu'un « backtest » a été effectué sur l'échantillon de 178 283 données préparées à cet effet via la division de la base de données originale 70%-30%. La robustesse des résultats est confirmée alors que nous obtenons des mesures d'efficacité similaires en employant les coefficients des régressions définis à la section 4.1. Pour démontrer la stabilité de nos modèles, les prochains résultats présentés seront ceux associés aux données provenant du « backtest ».

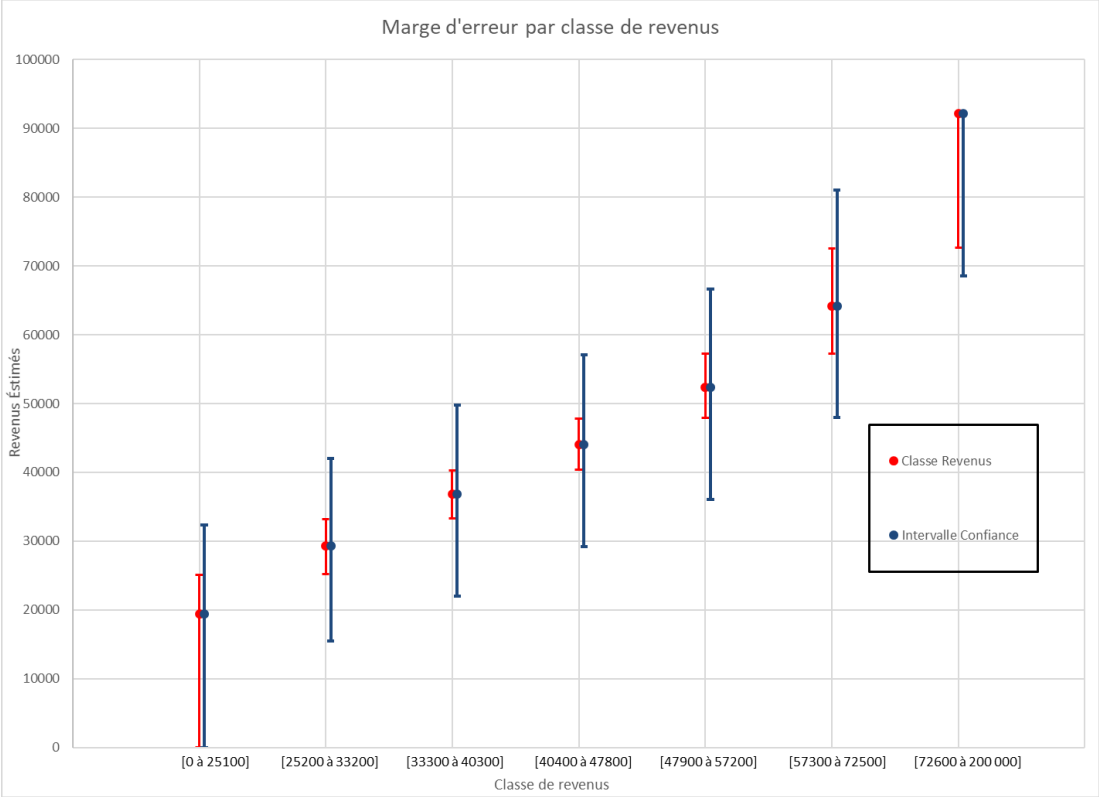
6.1 Estimation par intervalle de confiance

Nous avons opté pour la création d'un intervalle de confiance variant selon différentes classes de revenus. Le tableau 12 présente les résultats obtenus et la figure 5 en fait une représentation graphique.

Tableau 12 : Intervalle de confiance selon la classe de revenus

Intervalle de prédiction	***Inférieur à 25 100 $***	***25 200$ à 33 200$***	***33 300$ à 40 300$***	***40 400$ à 47 800$***	***47 900$ à 57 200$***	***57 300$ à 72 500$***	***Supérieur à 72 500$***
Revenus estimés moyens	*19 400$*	*29 300$*	*36 800$*	*44 000$*	*52 300$*	*64 200$*	*92 100$*
Revenus déclarés moyens	*19 800$*	*28 700$*	*35 900$*	*43 200$*	*51 880$*	*65 300$*	*94 600$*
80% des revenus réels	***Inférieur à 32 300$***	***15 500$ à 42 000$***	***22 000$ à 49 800$***	***29 200$ à 57 100$***	***36 100$ à 66 600$***	***48 000$ à 81 000$***	***Supérieur à 68 500$***

Figure 6 : Représentation graphique de l'intervalle de confiance



Prenons l'exemple de la classe de revenus [25 200$; 33 200$]. Cette classe est composée des revenus que notre modèle d'estimation a approximé. Instaurer un intervalle de confiance autour de cette classe permet d'affirmer que le revenu réel de 80% des clients se trouve dans la borne [15 500$; 42 000$]. Cette statistique est très importante puisqu'elle définit à partir de quand une déclaration sera considérée comme étant fausse. Par exemple, un client se faisant évaluer par notre modèle à 45 000$ brut annuel et qui en déclarerait 60 000$ serait considéré comme ayant fait une fausse déclaration (supérieur à 57 100$) et des vérifications manuelles seraient nécessaires. La dernière classe de revenus est la plus intrigante puisqu'elle suppose qu'un client se faisant évaluer à 75 000$ brut annuel pourrait déclarer des revenus de 200 00$ sans autres vérifications. Nous justifions cette classe en mentionnant que la proportion de clients de la banque à l'étude déclarant des revenus supérieurs à 75 000$ est faible. Le montant déclaré est secondaire pour ce type de clients puisqu'il est rare qu'ils se fassent évaluer en fonction de leurs revenus. En

général, ce type de clients est davantage évalué par la qualité de son bureau de crédit et ses habitudes de paiement.

6.2 Force de modèle

L'intervalle de confiance présenté précédemment est une bonne mesure pour déterminer les seuils correspondant à une fausse déclaration. Toutefois, cette mesure n'incorpore pas les résultats présentés à la figue 5 sur l'efficacité de chaque modèle. Notre modèle peut estimer un revenu pour le client A de 40 000$ et un revenu de 40 000$ pour le client B. Le client A peut avoir été évalué selon les modèles DI, MAJ 0 à 6 mois et PMT alors que le client B s'est fait évaluer uniquement par le modèle SOCIO. L'origine de l'évaluation est importante puisque l'estimation faite sur le client A est associée à un taux d'efficacité de 87.8% comparativement à 54.8% pour celle du client B.

Nous proposons alors d'ajouter une variable « Force du modèle » qui serait affiliée à chaque approximation de revenus en caractérisant l'efficacité sur une échelle de 1 à 10. Voici un exemple d'informations que la banque devrait recevoir pour chaque approximation :

Revenus Estimés : 50 000$
Revenus Déclarés : 78 000$
Intervalle acceptable : [36 100$; 66 600$] ⟶ *Fausse déclaration possible*
Force du modèle : 9

Le tableau 13 présente une suggestion de valeur pour la variable « Force du modèle » en fonction de l'efficacité avec laquelle le revenu d'un client a été estimé.

Tableau 13 : Proposition variable « Force du modèle »

Force du modèle	*Précision*
10	*>= 85%*
9	*>= 80%*
8	*>= 75%*
7	*>= 70%*
6	*>= 65%*
5	*>= 60%*
4	*>= 55%*
3	*>= 50%*
2	*>= 45%*
1	*<45%*

6.3 Recommandations

Évidemment, nous recommandons l'implantation du modèle d'estimation des revenus. Cette implantation pourrait s'effectuer par automatisation du processus d'octroi de crédit. Une intervention humaine serait uniquement nécessaire lorsque les revenus sortent de l'intervalle de confiance. La force du modèle permettrait de traduire les différentes mesures d'efficacité et les différents modèles utilisés en termes plus concrets à un analyste de crédit. Il serait libre à l'analyste de se fier ou non au modèle d'estimation en fonction de la force du modèle, l'écart entre les revenus estimés/réels et en fonction du montant demandé en $. En effet, nous ne recommandons pas l'utilisation de ce modèle lors de l'octroi de petits prêts personnels de 500$ ou encore d'une carte de crédit avec une faible limite. Le coût d'investigation serait plus élevé que l'espérance de pertes sur ce type de prêts. Or, si le prêt demandé dépasse par exemple 5000$, nous recommandons fortement d'investiguer sur les revenus potentiellement faux.

Pour pouvoir évaluer davantage de clients, il serait intéressant d'avoir accès à plus de variables sur les non-clients de l'institution financière faisant une application au crédit. Certaines compagnies externes se spécialisent dans la compilation de données provenant du bureau de crédit. Posséder des données outres celles de type socio-démographique permettrait de ne pas éliminer les 33 604 clients non-évalué de la base de données, tel que mentionné à la section 4.1.4.

La principale lacune de notre modèle réside dans la base de données alors que les fausses déclarations y sont actuellement inclues. Tel que vu à la section 4.1.2, le bruit présent dans les données nous oblige à choisir une mesure d'efficacité large (±10 000$ ou ±20%). La principale amélioration possible serait de refaire les modèles avec des revenus réels vérifiés. Répéter l'expérience sur une base de données plus petite, mais validée, confirmerait la robustesse des résultats et augmenterait l'efficacité lors de l'estimation.

Enfin, nous recommandons également de calculer les revenus de façon continue et non uniquement lorsqu'une application pour du crédit est effectuée. Ceci permettrait d'avoir un meilleur portrait du client et d'observer plus facilement les chocs entre revenus déclarés et revenus estimés.

7. Conclusion

À travers ce stage en entreprise, nous avons réussi à créer un modèle estimatif des revenus pour la clientèle de la banque. Ce modèle se sépare en 4 principaux modèles, soit DI, MAJ, PMT et SOCIO. Ces principaux modèles se divisent à leur tour en sous-modèles, permettant ainsi d'observer au total 12 mesures d'évaluation des revenus. Les divisions de modèles furent nécessaires pour maximiser l'utilisation de toute information pertinente disponible sur chaque client. Des variables telles que le dépôt direct moyen mensuel, les achats mensuels sur la carte de crédit, une mise à jour récente des revenus, l'âge du demandeur, son niveau d'épargne, etc… se sont avérées significatives. À l'inverse, de nombreux tests sur d'autres variables se sont avérés non pertinents et n'ont pas été présentés à travers ce rapport. À titre informatif, les scores de crédit, la province dans laquelle le client habite, le solde mensuel laissé sur une carte de crédit, la plus haute limite sur une carte de crédit au bureau de crédit et d'autres variables se sont toutes avérées être non-significatives.

Pour que notre estimation soit considérée comme étant réussie, elle devait être à ± 10 000$ ou ± 20% des revenus réels. L'évaluation faite par les différents modèles a démontré une efficacité de 72%, que nous avons par la suite réussit à optimiser à 74% en créant de la synergie entre les modèles. En d'autres termes, nous sommes capables d'estimer le revenu de 74% de nos clients dans un intervalle de ± 10 000$ ou ± 20% de leurs revenus réels. Lorsque l'écart entre notre approximation et la déclaration du client est trop élevé, nous recommandons de faire des investigations manuelles pour déterminer si le client a fait une fausse déclaration. Ceci permet de réduire le risque d'octroyer du crédit de mauvaise qualité (un octroi alors que le demandeur ne se qualifie pas). Il fut toutefois nécessaire d'exclure environ 7% de la base de données, soit les clients qui ne possèdent rien avec la banque et qui font une demande de crédit. La faible précision de ce segment a nécessité le réalignement des objectifs et de l'utilisation possible du modèle en le consacrant uniquement aux clients existants.

Finalement, la robustesse de nos résultats a été confirmée à l'aide d'un « backtest ». La prochaine étape serait de réaliser une analyse coût-bénéfice pour définir à partir de quel

montant demandé en $ par le client serait-il nécessaire d'effectuer des vérifications manuelles sur les revenus. Certes, ce modèle démontre une grande utilité pour la gestion des risques lors du processus d'octroi de crédit. Il peut également avoir une utilisation Marketing puisque les revenus des clients de la banque sont maintenant connus en temps continu. La banque peut ainsi faire de la sollicitation personnalisée sur chacun de ses clients. Plutôt que d'envoyer une proposition d'une carte de crédit régulière, un client pour lesquels de gros revenus seraient estimés se verrait recevoir une sollicitation pour une carte de crédit de type élite.

Une ouverture de recherche serait de proposer un intervalle de confiance personnalisé à chaque client. Selon notre modèle, les intervalles de confiances sont personnalisés par classe de revenus estimés. La force de modèle sert d'indicateur sur notre confiance à l'égard de l'estimation. L'intervalle personnalisé ferait en sorte que plus la banque est confiante dans son estimation, plus l'intervalle de confiance serait petit autour des revenus estimés. De plus, nous aurions aimé pouvoir tester certaines variables qui n'étaient pas présentes dans la base de données de la banque. Une reformulation des questions à poser sur le formulaire d'application au crédit pour inclure des questions sur le plus haut niveau de scolarité atteint, par exemple, pourrait également s'avérer pertinent.

Nous recommandons alors à chaque banque de se munir de son propre modèle d'estimation de revenus pour prévenir les fausses déclarations de revenus et limiter l'asymétrie d'information entre le prêteur et l'emprunteur.

Bibliographie

AVERY ET AL., (2004). «*Consumer credit scoring: Do situational circumstances matter?»,* Journal of Banking & Finance, no° 28, p. 835-856.

BLACKBURN M. ET T.VERMILYEA., (2012). «The prevalence and impact of misstated incomes on mortgage loan applications*»,* Journal of Housing Economics, no° 21, p. 151-168.

CALEM P. ET AL., (2011), «*Credit Cycle and Adverse Selection Effects in Consumer Credit Markets - Evidence from the HELOC Market*», FRB of Philadelphia, no° 11-13.

DIBOUNE, Hind., (2008). «*La prise en compte de la capacité à payer dans l'évaluation du risque de crédit des particuliers*», mémoire de maîtrise, HEC Montréal, Chaire de recherche au Canada en gestion des risques. Récupéré le 8 décembre 2018 de http://chairegestiondesrisques.hec.ca/wp-content /uploads/pdf/equipe/Diboune-memoire.pdf

FINLAY, SM., (2006). «*Predictive models of expenditure and over-indebtedness for assessing the affordability of new consumer credit applications*», Journal of the Operational Research Society, n°57, p. 655-669.

GRAVEL, MARIE-ANDRÉE., (2016). *Données sociodémographiques en bref.* Institut de la statistique du Québec. Récupéré le 8 décembre 2018 de http://www.stat.gouv.qc.ca/statistiques/conditions-vie-societe/bulletins/sociodemo-vol20-no2.pdf

KIBEKBAEV A. ET E.DUMAN., (2016). «Benchmarking regression algorithms for income prediction modeling», Elsevier, Book Information Systems, p. 40-52.

MELOCHE-HOLUBOWSKI, Mélanie., (2017). *Des revenus à la hausse, mais pas pour tous les canadiens*. Radio Canada. Récupéré le 8 décembre 2018 de https://ici.radio-canada.ca/nouvelle/1055308/recensement-revenu-salaire-statistique-canada-pauvrete-richesse

MILLER, Sarah., (2015). «*Information and default in consumer credit markets: Evidence from a natural experiment*», Journal of Financial Intermediation, no° 24, p. 45-70.

THOMAS ET AL., (2002). «Credit Scoring and its Applications», SIAM : Philadelphia

THOMAS, LC., (2009) «*Modelling the Credit Risk for Portfolios of Consumer Loans : Analogies with the corporate loan models* », School of Management, University of Southampton, Southampton.

Travail, Emploi et Solidarité sociale Québec (2018). *Programme d'aide sociale et Programme de solidarité sociale.* Gouvernement du Québec. Récupéré le 8 décembre 2018 de http://www.emploiquebec.gouv.qc.ca/fileadmin/fichiers/pdf/Publications/00_nouv-montants-prestation_2018.pdf

Annexe 1 : Poids que l'on attribue à chaque modèle selon les différents scénarios

Numéro du scénario	*Nom du scénario*	*DI*	*MAJ*	*PMT*	*SOCIO*	*NB de clients dans cette situation*	*Efficacité*
1	DI-MAJ0à6- PMT Hyp & Débit	0.67	0.25	0.08	.	13472	89,39%
2	DI-MAJ0à6- PMT Hyp	0.74	0.14	0.12	.	4206	88,54%
3	DI-MAJ0à6-PMT Debit	0.61	0.28	0.11	.	24530	86,84%
4	DI-MAJ0à6-PMT de base	0.66	0.2	0.14	.	1294	88,18%
5	DI-MAJ7à12-PMT Hyp & Debit	0.71	0.28	0.01	.	7900	87,11%
6	DI-MAJ7à12- PMT Hyp	0.7	0.28	0.02	.	1851	87,14%
7	DI-MAJ7à12-PMT Debit	0.67	0.25	0.08	.	12146	84,63%
8	DI-MAJ7à12-PMT de base	0.61	0.33	0.06	.	391	82,35%
9	DI-MAJ13à24-PMT Hyp & Debit	0,74	0.26	0	.	10094	83,31%
10	DI-MAJ13à24- PMT Hyp	0,8	0.2	0	.	534	82,96%
11	DI-MAJ13à24-PMT Debit	0,49	0.48	0.03	.	14185	77,28%
12	DI-MAJ13à24-PMT de base	0,73	0.17	0.1	.	216	78,7%
13	DI-MAJ25à36- PMT Hyp & Debit	0.69	0.3	0.01	.	4818	80,68%
14	DI-MAJ25à36- PMT Hyp	0.71	0.29	0	.	168	77,98%
15	DI-MAJ25à36- PMT Debit	0.3	0.69	0.01	.	8087	74,38%
16	DI-MAJ25à36- PMT de base	0.12	0.78	0.1	.	89	69,66%
17	DI-MAJ37+ - PMT Hyp & Debit	0.62	0.38	0	.	2745	79,24%
18	DI-MAJ37+ - PMT Hyp	0.29	0.59	0.12	.	100	79%
19	DI-MAJ37+ - PMT Debit	0.16	0.81	0.03	.	8175	73,19%
20	DI-MAJ37+ - PMT de base	0.47	0.53	0	.	62	72,58%
21	DI-MAJ0à6	0.64	0.36	.	.	20045	87,53%
22	DI-MAJ7à12	0.67	0.33	.	.	10962	85,15%
23	DI-MAJ13à24	0.61	0.39	.	.	14473	79,49%
24	DI-MAJ25à36	0.52	0.48	.	.	8848	79,14%
25	DI-MAJ37+	0.34	0.66	.	.	8380	75,38%
26	MAJ0à6-PMT Hyp & Debit	0.85	.	0.15	.	6822	82,89%
27	MAJ0à6-PMT Hyp	0.84	.	0.16	.	2409	81,45%
28	MAJ0à6-PMT Debit	0.82	.	0.18	.	10696	81,98%
29	MAJ0à6-PMT de base	0.84	.	0.16	.	1296	80,32%
30	MAJ7à12-PMT Hyp & Debit	0.92	.	0.08	.	4087	78,59%
31	MAJ7à12-PMT Hyp	0.93	.	0.07	.	1102	76,77%
32	MAJ7à12-PMT Debit	0.94	.	0.06	.	6223	75,96%
33	MAJ7à12-PMT de base	0.94	.	0.06	.	490	73,06%
34	MAJ13à24- PMT Hyp & Debit	1	.	0	.	6012	69,91%
35	MAJ13à24- PMT Hyp	0.99	.	0.01	.	455	70,77%
36	MAJ13à24- PMT Debit	1	.	0	.	8485	64,44%
37	MAJ13à24- PMT de base	0.97	.	0.03	.	498	62,65%
38	MAJ25à36- PMT Hyp & Debit	1	.	0	.	3169	65,13%
39	MAJ25à36- PMT Hyp	1	.	0	.	193	60,62%
40	MAJ25à36- PMT Debit	0.96	.	0.04	.	5445	57,39%
41	MAJ25à36- PMT de base	0.92	.	0.08	.	288	61,46%
42	MAJ37+ - PMT Hyp & Debit	1	.	0	.	2278	59,31%
43	MAJ37+ - PMT Hyp	0.77	.	0.23	.	127	50,39%
44	MAJ37+ - PMT Debit	0.93	.	0.07	.	6511	53,79%
45	MAJ37+ - PMT de base	1	.	0	.	415	53,01%
46	DI-PMT Hyp & Debit	.	0.96	0.04	.	655	61,68%
47	DI-PMT Hyp	.	0.62	0.38	.	82	53,66%
48	DI-PMT Debit	.	0.93	0.07	.	10752	69,78%
49	DI-PMT de base	.	0.84	0.16	.	81	59,26%
50	MAJ SEUL	.	1	.	.	101680	70,65%
51	DI SEUL	1	.	.	.	16085	74,70%
52	PMT SEUL	.	.	1	.	13015	38,90%
53	SOCIO SEUL	.	.	.	1	28872	54,80%