Université catholique de Louvain

Faculté des Sciences – Institut de statistique, biostatistique et sciences actuarielles (ISBA)

LACTU2310

Statistical learning methods for insurance

Travail réalisé par :

Davy Romaric KAMGA BOPDA

Superviseur académique :

Karim Barigou

Rapport de projet – 24 juillet 2025

Introduction

Contexte et Objectif du Projet

En 2019, une agence de voyages a proposé à sa clientèle une nouvelle assurance voyage intégrant une couverture Covid-19. Sur environ 2 000 clients ciblés, seuls 36 % ont effectivement souscrit au produit, soit un client sur trois. L'objectif de ce projet est d'exploiter les données collectées lors de cette campagne pour prédire la probabilité de souscription future, à partir des caractéristiques sociodémographiques, des habitudes de voyage et de l'historique médical des clients.

Concrètement, il s'agit de construire un modèle de classification binaire (souscription : Oui/Non) performant, permettant à l'entreprise de mieux cibler les clients les plus susceptibles d'accepter cette offre. Dans le cadre du cours, cette modélisation s'apparente à un exercice de type *claim occurrence modeling* avec estimation de la propension individuelle I_i, sur une base binomiale. Toutefois, contrairement aux contextes traditionnels d'assurance, aucune variable ne permet ici de mesurer explicitement l'exposition au risque.

Données et Variables

Le jeu de données contient 1 987 individus, avec les variables explicatives suivantes :

1. **Age** : âge du client,
2. **Employment Type** : type d'emploi (secteur privé ou indépendant vs fonction publique),
3. **GraduateOrNot** : niveau d'éducation (diplômé universitaire ou non),
4. **AnnualIncome** : revenu annuel (en roupies indiennes),
5. **FamilyMembers** : nombre de membres dans le foyer,
6. **ChronicDiseases** : présence d'une maladie chronique (Oui/Non),
7. **FrequentFlyer** : client voyageant fréquemment (Oui/Non),
8. **EverTravelledAbroad** : client ayant déjà voyagé à l'étranger (Oui/Non),
9. **TravelInsurance** : variable cible (souscription à l'assurance : Oui/Non).

Toutes ces variables sont potentiellement informatives pour prédire la décision de souscription. Une **analyse exploratoire des données** préliminaire, principalement réalisée avec le package `ggplot2`, permettra de visualiser les relations entre la variable cible et les covariables, de détecter d'éventuelles valeurs aberrantes, et de mieux comprendre la structure des dépendances présentes dans les données.

Le cadre de modélisation retenu, basé sur une régression logistique binomiale, s'appuie sur les travaux de référence suivants :
Frees, E. W. (2010). Regression Modeling with Actuarial and Financial Applications. Cambridge University Press.

Ce projet vise in fine à identifier le **modèle d'arbre optimal** permettant d'atteindre un bon compromis entre précision prédictive et capacité de généralisation sur de nouvelles données.

Pré-processing

Cette phase initiale a permis de vérifier le type et le format des variables disponibles. Certaines transformations ont été nécessaires. Par exemple, bien que codée numériquement, la variable **ChronicDiseases** représente une information binaire : 1 pour la présence d'une affection chronique, 0 pour son absence ; elle a donc été recodée comme telle. La variable **FamilyMembers** a été traitée comme variable discrète. La variable cible **TravelInsurance**, bien que numérique dans le fichier source, est de nature catégorielle binaire et sera donc convertie en facteur pour l'ajustement des modèles. L'analyse exploratoire n'a révélé aucune donnée manquante dans l'échantillon.

Analyse descriptive

Analyse univariée

Les résultats de cette analyse indiquent que 50% des individus ont un âge supérieur à 29 ans (médiane = 29), et 75 % ont un âge inférieur à 32 ans. De plus, l'âge moyen est de 29,65 ans, avec le plus âgé ayant 35 ans et le plus jeune, 25 ans. Concernant la variable **AnnualIncome**, les analyses révèlent que 50% des individus ont un revenu annuel inférieur à 900 000 roupies (médiane = 900 000), tandis que le revenu moyen s'élève à 932 763 roupies. Le revenu le plus élevé atteint 1 800 000 roupies.

TABLE 1 – Statistiques descriptives pour les variables quantitatives

Variable	Min.	1er Qu.	Médiane	Moyenne	3e Qu.	Max.
Age	25.00	28.00	29.00	29.65	32.00	35.00
AnnualIncome	300,000	600,000	900,000	932,763	1,250,000	1,800,000

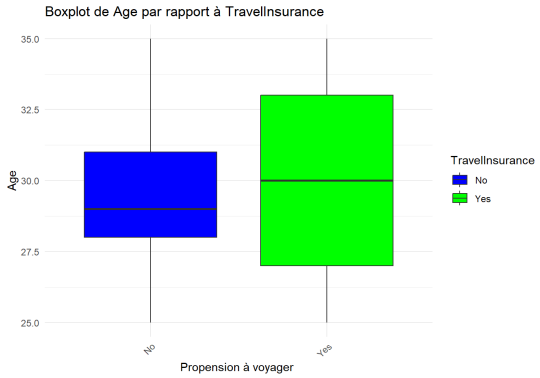
— **Secteur d'emploi** : 71% des individus travaillent dans le **secteur privé** ou sont **travailleurs indépendants**, contre 29% employés par le **gouvernement**.

— **Niveau d'études** : 85% des individus détiennent un **diplôme universitaire**, tandis que 15% n'en ont pas.
— **Fréquence de voyage** : 79% **ne voyagent pas fréquemment**, contre 21% qui **voyagent régulièrement**.
— **Voyage à l'étranger** : 81% **n'ont jamais voyagé à l'étranger**, tandis que 19% **l'ont déjà fait**.
— **Maladies chroniques** : 72% des individus **ne souffrent d'aucune maladie chronique**, contre 28% atteints d'au moins une.
— **Structure familiale** : Les familles de **4 membres** sont les plus fréquentes (**25,42%**), suivies de près par celles de **5 membres** (**21,44%**). Ces deux groupes représentent **près de 47%** de l'échantillon.
— **Souscription à l'assurance voyage** : Seuls **36% des clients** ont **souscrit** à l'assurance voyage, tandis que **64%** ne l'ont **pas fait**, ce qui suggère que **près de 2 clients sur 3** refusent l'offre.

Analyse bivariée

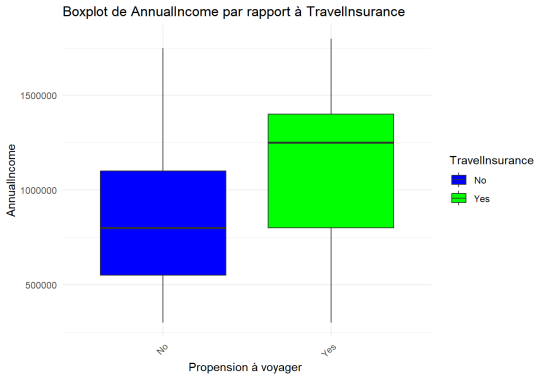
Le graphique des *boxplots* met en évidence des tendances marquées : L'**âge** semble jouer un rôle déterminant dans la souscription à l'assurance voyage : les individus **plus âgés** sont **davantage enclins à souscrire**, contrairement aux plus jeunes. Cette observation suggère une **corrélation positive** entre l'âge et la probabilité de souscription. Une tendance similaire est observée pour le **revenu annuel** : les personnes ayant un **revenu plus élevé** sont **significativement plus susceptibles** de souscrire. Cette variable apparaît donc comme un **facteur discriminant majeur** dans la décision.

(a) **Boxplot de la variable age vs target**



Source : Nos Travaux

(b) **Boxplot de la variable revenu vs target**



Source : Nos Travaux

Pour les variables qualitatives, un test d'indépendance du **Khi deux** a été réalisé afin d'évaluer leur association statistique avec la variable cible `TravelInsurance`. Les résultats sont présentés ci-dessous :

TABLE 2 – Résultats du test du Khi² entre les variables qualitatives et la souscription à l'assurance voyage

Variable	p-value	Association significative
Employment Type	**0.0000**	**Oui**
GraduateOrNot	0.4365	Non
FamilyMembers	**0.0001**	**Oui**
ChronicDiseases	0.4481	Non
FrequentFlyer	**0.0000**	**Oui**
EverTravelledAbroad	**0.0000**	**Oui**

Les variables **statistiquement liées à la souscription** à une assurance voyage (c'est-à-dire ayant une p-value $< 0{,}05$ au test du Khi deux) sont les suivantes :`Employment Type; FamilyMembers; FrequentFlyer et EverTravelledAbroad`. Ces quatre variables présentent une **association significative** avec la variable cible TravelInsurance, et peuvent donc être retenues pour les phases ultérieures de *modélisation prédictive*.

Méthodologie et Modèles Envisagés

Pour modéliser la variable cible TravelInsurance (Oui/Non), nous avons retenu des méthodes de classification supervisée fondées sur des arbres de décision et des techniques d'ensembles d'arbres, vues dans le cours à savoir : Un arbre de décision CART (Classification And Regression Tree) initial et son arbre élagué optimal, le Bagging (Bootstrap Aggregating) d'arbres de décision, la Forêt Aléatoire (Random Forest), le Gradient Boosting (arbres de décision boostés). Ces méthodes exploitent toutes la nature binaire de la cible (modèle binomial avec fonction logit) tout en offrant différents compromis entre complexité, variance et biais du modèle. L'idée est de comparer un arbre simple interprétable à des modèles d'ensemble plus complexes mais potentiellement plus précis.

Modélisation

Avant de procéder à l'ajustement des modèles, nous avons divisé notre base de données en deux sous-ensembles : **80% des observations ont été utilisées pour l'entraînement (train)** et **20% pour le test**. Cette séparation vise à évaluer la capacité de généralisation des modèles sur des données non vues.

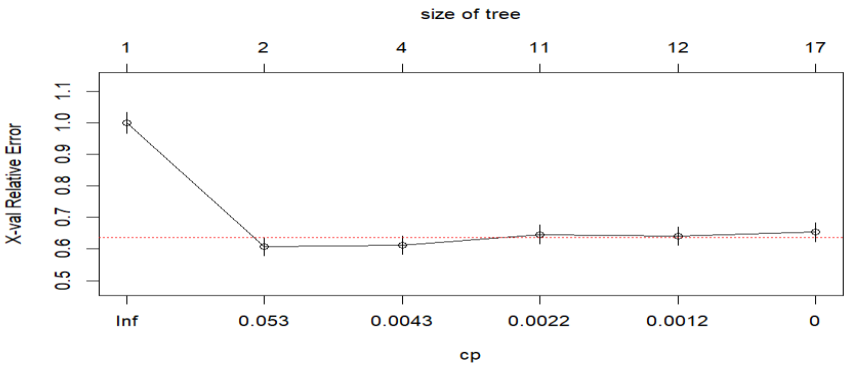
Le modèle de classification binaire utilisé ici vise à prédire la probabilité de souscription à l'assurance ($Y = 1$) conditionnellement aux variables explicatives $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. Sa formulation générale est la suivante : $P(Y = 1 \mid \mathbf{X}) = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$ où

g^{-1} est la fonction de lien inverse : Dans le cas des **modèles linéaires généralisés** (GLM), notamment la régression logistique, il s'agit de la **fonction logit inverse** : $g^{-1}(z) = \frac{1}{1+e^{-z}}$

Arbre de décision (CART) et élagage

Sélection du paramètre de complexité (cp) En ajustant le modèle d'arbre de décision avec la fonction `rpart`, nous avons activé la **validation croisée interne** via l'argument `xval`. Cela nous a permis de générer la courbe des valeurs du `cp` (complexity parameter), visualisée ci-dessous :

FIGURE 2 – Visualisation du paramètre de complexité (`cp`)



Source : Élaboration des auteurs à partir des données.

Cette courbe nous a permis d'identifier une première approximation du **cp optimal**, correspondant à la valeur minimisant l'erreur de validation croisée. Afin d'affiner ce choix, nous avons ensuite testé plusieurs valeurs de `cp` comprises entre cette valeur minimale, sa moitié et son carré. La valeur retenue, correspondant au meilleur compromis entre complexité et performance, a ensuite été utilisée pour **l'élagage final de l'arbre**.

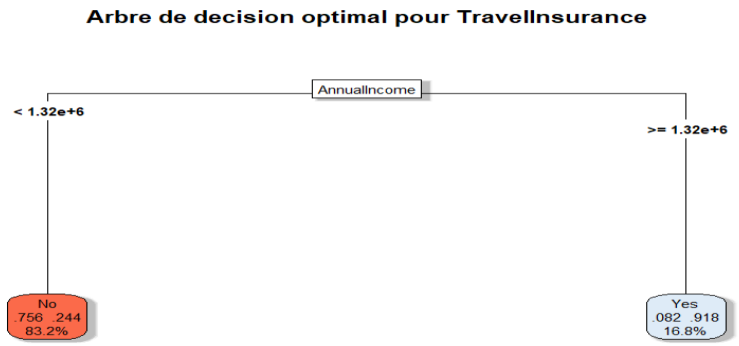
La valeur optimale du paramètre de complexité, estimée à **0,007042254**, a été utilisée pour procéder à l'élagage de l'arbre. L'arbre de décision ainsi obtenu, simplifié à l'aide de cette valeur de `cp`, est présenté ci-dessous :

Importance des variables issue de l'arbre de décision

Comme l'illustre la figure ci-dessous, les variables les plus déterminantes dans la construction de l'arbre sont **AnnualIncome** et **EverTravelledAbroad**.

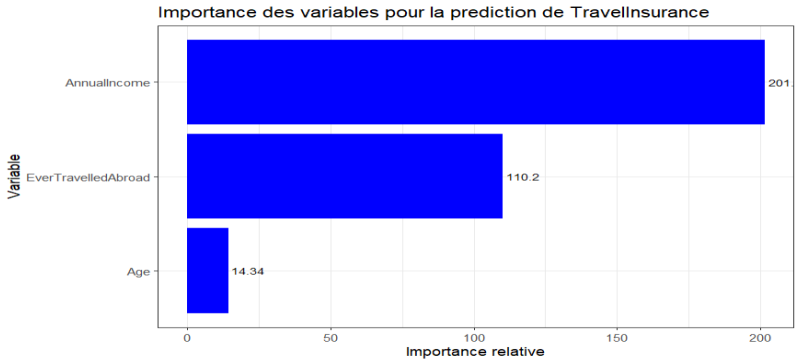
La variable **AnnualIncome** se distingue nettement avec une importance relative de **201**, ce qui signifie qu'elle intervient très fréquemment dans les critères de séparation de l'arbre. Elle constitue le principal facteur de décision pour prédire la souscription à

FIGURE 3 – Arbre de décision élagué



Source : Élaboration des auteurs à partir des données.

FIGURE 4 – Importance des variables dans l'arbre de décision



Source : Élaboration des auteurs à partir des données.

l'assurance voyage. En deuxième position, la variable **EverTravelledAbroad** affiche une importance de **110,2**, indiquant un rôle également significatif dans la modélisation. Elle contribue probablement à affiner la décision après franchissement d'un certain seuil de revenu. À l'inverse, la variable **Age**, avec une importance bien plus faible (**14,34**), exerce un effet marginal. Elle pourrait intervenir dans quelques feuilles terminales, mais n'apparaît pas comme un critère discriminant majeur dans la structure globale de l'arbre.

Graphiques de dépendance partielle (PDP)

Les graphiques de dépendance partielle (PDP) permettent d'évaluer l'effet marginal de chaque variable sur la probabilité de souscription, toutes choses égales par ailleurs.

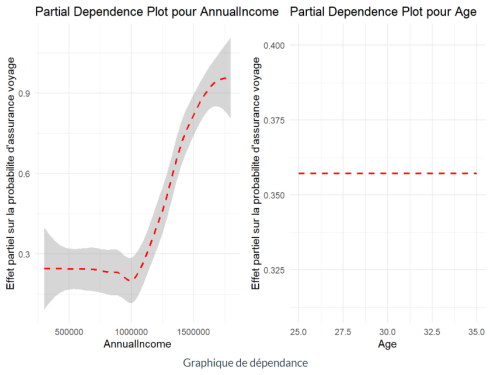
— **Revenu annuel** : la probabilité de souscription reste faible et stable (~30%) tant que le revenu est inférieur à 1 million. Une hausse marquée est observée entre 1 et 1,3 million, avec une probabilité dépassant 90% au-delà de 1,5 million. Cela

suggère un **effet seuil** fort du revenu sur la décision.

— **Âge** : aucun effet significatif n'est détecté. La probabilité de souscription demeure constante quel que soit l'âge, indiquant une absence d'influence directe dans ce modèle.

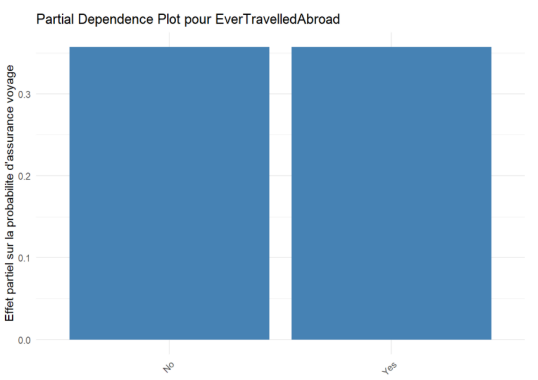
— **EverTravelledAbroad** : les deux modalités (Oui/Non) présentent des effets très similaires, ce qui révèle une **influence marginale** de cette variable dans la prédiction.

(a) Effet du revenu annuel et age sur la probabilité de souscription



Source : Élaboration des auteurs

(b) Effets de l'âge et du voyage à l'étranger



Source : Élaboration des auteurs

Méthodes d'Ensemble : Bagging

Implémentation du bagging avec des arbres de décision

Dans cette étude, nous avons implémenté la méthode de **Bagging** (Bootstrap Aggregating), en combinant plusieurs arbres de décision construits sur des échantillons bootstrap issus des données d'entraînement. L'objectif principal de cette méthode est de **réduire la variance** associée à un modèle unique, afin d'améliorer la stabilité et la robustesse des prédictions.

Concrètement, nous avons généré **50 sous-échantillons bootstrap** à partir de l'échantillon d'apprentissage. Pour chacun de ces échantillons, un arbre de décision a été entraîné en utilisant les variables explicatives disponibles selon un cadre de classification.

Chaque arbre a été construit à l'aide d'un **paramètre de complexité** *cp* optimisé au préalable (issu de la phase de calibration de l'arbre unique), de manière à éviter les modèles surajustés ou trop simples. L'ensemble des arbres ainsi obtenus a ensuite été agrégé pour produire une prédiction finale :

— soit par **vote majoritaire** dans le cas d'une classification binaire stricte,

— soit par **moyenne des probabilités prédictives** lorsque l'on souhaite interpréter le score de souscription.

Cette approche permet d'augmenter la **capacité de généralisation du modèle** en neutralisant les erreurs spécifiques aux différents échantillons et en exploitant la diversité des arbres produits.

Bagging

Une implémentation manuelle du **Bagging** a été réalisée en générant 100 arbres de décision sur des échantillons bootstrap de l'échantillon d'apprentissage. Chaque arbre a été construit avec un cp issu du modèle d'arbre pour éviter le surapprentissage.

Les prédictions probabilistes issues de chaque arbre ont été agrégées pour produire une probabilité moyenne de souscription, permettant ensuite le calcul de différentes métriques (accuracy, AUC, variance).

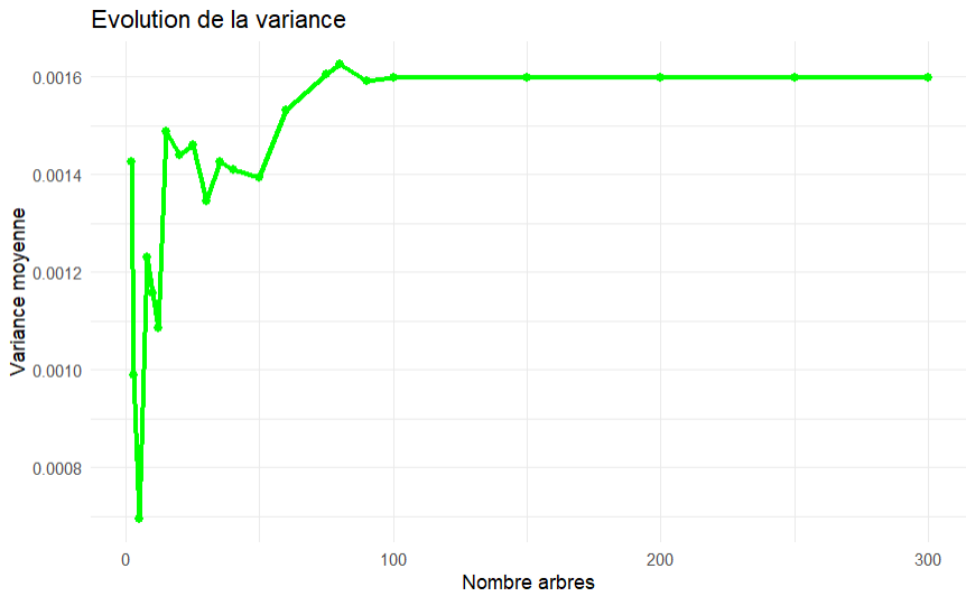


FIGURE 6 – Diminution de la variance des prédictions avec le nombre d'arbres

Analyse des performances du Bagging

Les résultats montrent une **stabilité remarquable de l'accuracy**, qui atteint **0,836** dès le premier arbre et reste constante jusqu'à 300 arbres. Cela suggère que le signal prédictif principal est capté très tôt, et que l'ajout d'arbres supplémentaires n'améliore pas directement cette métrique.

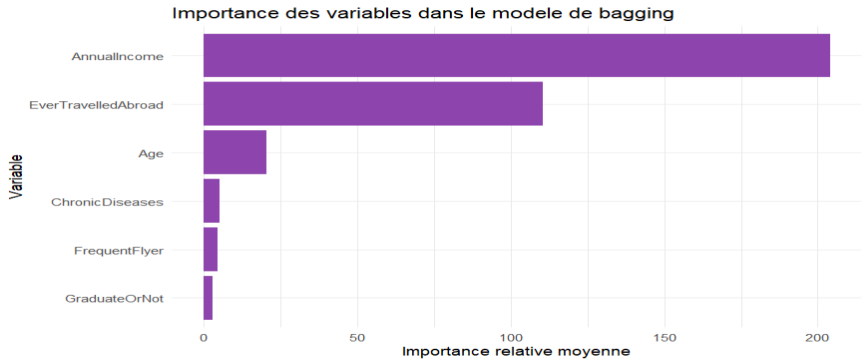
En revanche, le graphique de la variance moyenne met en évidence une **diminution progressive de la variance** des prédictions à mesure que le nombre d'arbres augmente. La

variance devient stable à partir de **30 à 50 arbres**, confirmant le rôle central du bagging dans la **réduction de la variance** des modèles instables comme les arbres de décision. Le bagging améliore la **robustesse** du modèle sans perte de performance. Il stabilise les prédictions en lissant les fluctuations liées aux arbres individuels, tout en maintenant une capacité discriminante élevée dès les premières itérations.

Importance des variables dans le modèle Bagging

Le graphique ci-dessous présente l'importance relative des variables dans le modèle Bagging. On constate que, comme pour l'arbre de décision simple, les variables **AnnualIncome** et **EverTravelledAbroad** sont les plus influentes, suivies par **Age**. Les autres variables ont un impact négligeable sur la prédiction.

FIGURE 7 – Importance des variables dans le modèle Bagging



Source : Élaboration des auteurs à partir des données.

Analyse par dépendance partielle (PDP)

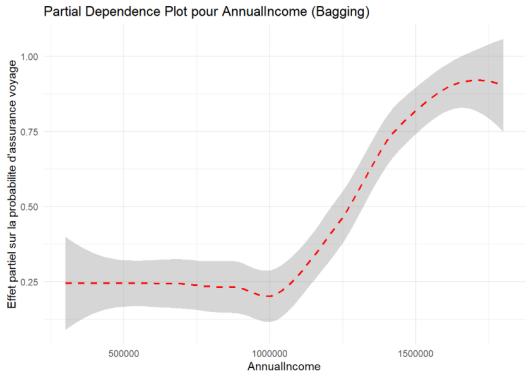
Les graphiques de dépendance partielle (PDP) permettent d'évaluer l'effet marginal des variables explicatives sur la probabilité de souscription, toutes choses égales par ailleurs. Les résultats issus du modèle Bagging révèlent les tendances suivantes :

— **AnnualIncome** : un effet seuil net est observé. La probabilité de souscription reste faible (environ 30%) en dessous d'1 million, puis augmente rapidement pour dépasser 90% au-delà de 1,5 million. Le revenu est ainsi confirmé comme **le principal facteur discriminant**.

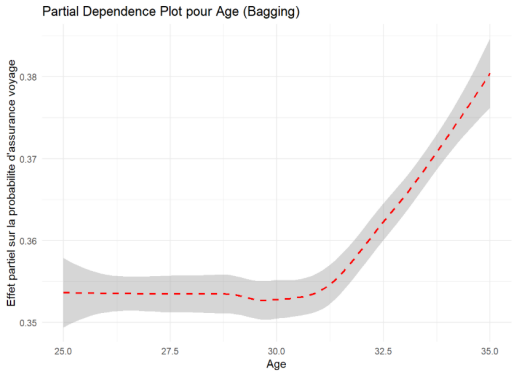
— **Age** : bien que secondaire, l'effet de l'âge est mieux capté dans le modèle Bagging. Une légère tendance croissante apparaît à partir de 30 ans, traduisant une influence modérée mais non négligeable.

— **EverTravelledAbroad** : les modalités Oui et Non présentent des effets très proches, indiquant une **contribution marginale** à la prédiction finale.

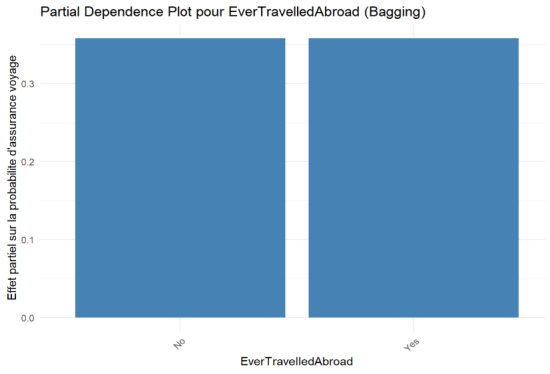
(a) Effet du revenu annuel



(b) Effet de l'âge



(c) Effet de EverTravelledAbroad



Source : Élaboration des auteurs à partir des données.

Modele Random Forest

Évaluation et Optimisation du Modèle Random Forest

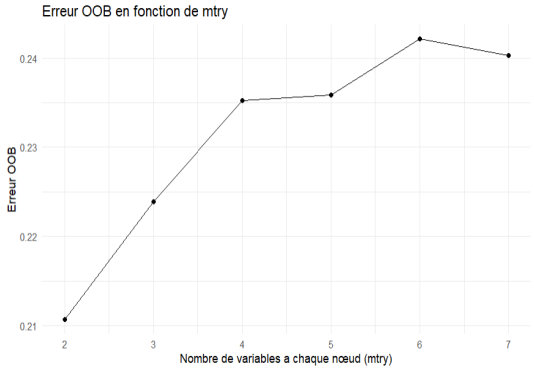
Cette section a pour objectif d'évaluer la performance du modèle **Random Forest** en fonction de trois hyperparamètres clés : `mtry` (nombre de variables candidates à chaque nœud), `ntree` (nombre d'arbres) et `nodesize` (taille minimale des nœuds terminaux). L'analyse s'appuie à la fois sur l'erreur Out-of-Bag (OOB) et sur la validation croisée, afin d'identifier la configuration optimale. En complément, des graphiques d'importance des variables et des courbes de dépendance partielle (PDP) seront fournis pour interpréter le modèle.

Un premier modèle Random Forest a été construit avec les paramètres suivants : `ntree = 50`, `mtry = 2`, `nodesize = 5`. L'erreur OOB observée se stabilise autour de **21,8%** dès 40 arbres, ce qui indique une bonne convergence du modèle. On remarque également que cette erreur est légèrement plus faible pour la classe majoritaire, ce qui

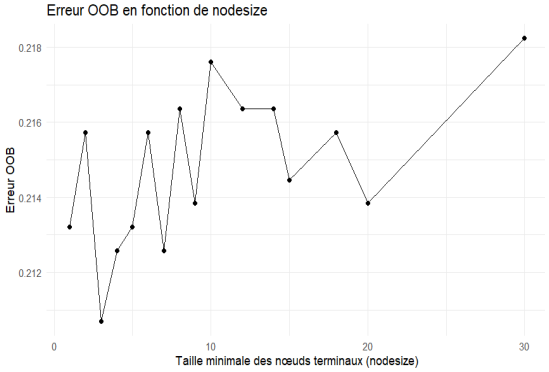
pourrait traduire un déséquilibre dans les données ou une prédiction plus aisée des non-souscriptions.

Pour affiner la configuration du modèle, nous avons fait varier le paramètre `mtry` entre 2 et 7. La valeur minimisant l'erreur OOB a ensuite été retenue pour explorer, dans un second temps, l'effet de `nodesize` (variant de 1 à 30). Cette démarche itérative nous a permis d'identifier les combinaisons d'hyperparamètres les plus performantes.

(a) Erreur OOB en fonction de `mtry`



(b) Erreur OOB en fonction de `nodesize`



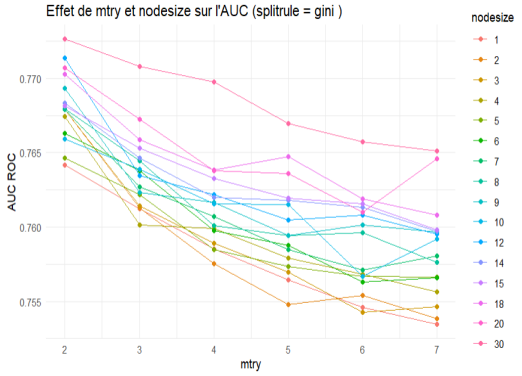
Source : Élaboration des auteurs à partir des données.

Le graphique de droite illustre l'évolution de l'erreur OOB en fonction de la taille minimale des nœuds terminaux. L'erreur varie de façon non monotone, mais les plus faibles valeurs sont obtenues lorsque `nodesize` est compris entre 1 et 5. Cela suggère que des feuilles plus petites permettent au modèle de mieux capturer la complexité des données. En revanche, au-delà de `nodesize = 10`, l'erreur tend à croître, ce qui peut indiquer une perte de précision liée à une sous-segmentation excessive de l'espace de décision.

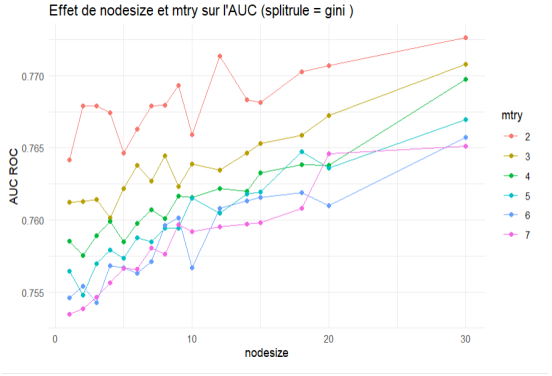
Recherche croisée des hyperparamètres optimaux par validation croisée

Dans un premier temps, nous avons effectué une validation croisée en faisant varier le nombre d'arbres (`ntree`) de 10 à 500 par pas de 20. L'objectif était d'identifier le nombre optimal d'arbres maximisant l'accuracy. La valeur optimale obtenue a ensuite été fixée pour affiner l'ajustement des deux autres hyperparamètres : `mtry` (variant de 2 à 7) et `nodesize` (variant de 1 à 30). L'ensemble de ces explorations a permis de construire une grille complète d'évaluation croisée, à partir de laquelle nous avons identifié les combinaisons de paramètres maximisant la performance du modèle, mesurée par l'AUC. Les résultats sont illustrés ci-dessous :

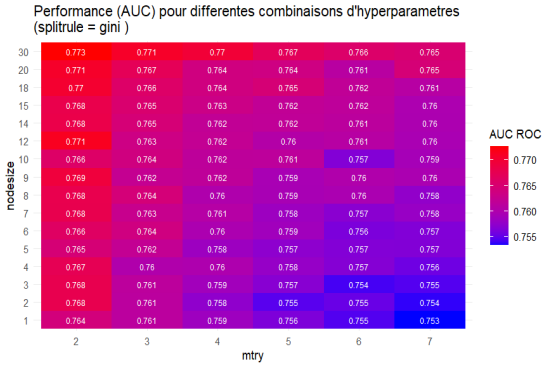
(a) Impact combiné de `mtry` et `nodesize` sur l'AUC



(b) Évolution de l'AUC selon `nodesize` pour différents `mtry`



(c) Effet croisé des paramètres sur la performance globale

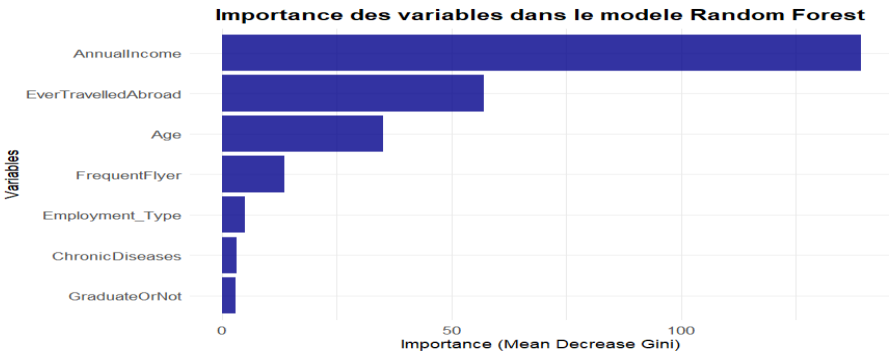


Source : Élaboration des auteurs à partir des données.

Le graphique 10a illustre que les meilleures performances sont atteintes avec un mtry faible (2 ou 3) et un nodesize relativement élevé (≥ 20). En revanche, le graphique 10b révèle qu'avec un nodesize fixé, une diminution de mtry s'accompagne d'une augmentation de la performance (AUC). Cela souligne l'importance des arbres moins complexes, mais plus diversifiés.

Avec un mtry faible, l'erreur OBB diminue, entraînant une réduction de l'erreur de validation croisée et une amélioration de la métrique AUC. Cela rend les arbres moins corrélés, ce qui réduit la variabilité. Cependant, la variation de nodesize fait fluctuer la courbe des erreurs OBB. Néanmoins, une évolution de celle-ci, pour un mtry fixé, permet de réduire progressivement l'erreur de validation croisée.

FIGURE 11 – Importance des variables dans le modèle Random Forest



Source : Élaboration des auteurs à partir des données.

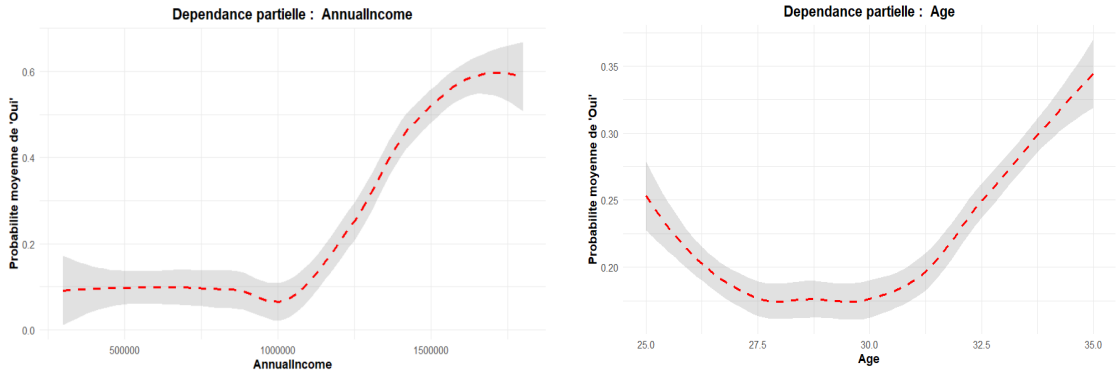
Importance des variables dans le modèle Random Forest

Le graphique illustre que les variables les plus influentes sont **AnnualIncome** et **EverTravelledAbroad**, suivies de près par **Age**.

Analyse par dépendance partielle (PDP)

(a) Effet du revenu annuel

(b) Effet de l'âge



Source : Élaboration des auteurs à partir des données.

Revenu Annuel : Il existe une relation positive entre le revenu annuel et la probabilité de souscription à une assurance voyage. Pour des revenus inférieurs à 1 million, la probabilité est faible (10-15 %). Elle augmente rapidement à partir de 1,2 million, dépassant 60 % au-delà de 1,5 million. Ainsi, un revenu plus élevé est un facteur déterminant pour souscrire à cette assurance.

Âge : L'analyse des données montre que la probabilité de souscription à une assurance voyage est faible (17-20 %) chez les 25-28 ans. Elle commence à augmenter à partir de

30 ans, atteignant plus de 35 % au-delà de 33 ans. Cela indique une influence positive de l'âge sur la propension à souscrire, surtout à partir de la trentaine.

Gradient Boosting

Cette méthode repose sur l'utilisation d'une fonction de lien, en l'occurrence la fonction logit, définie par :

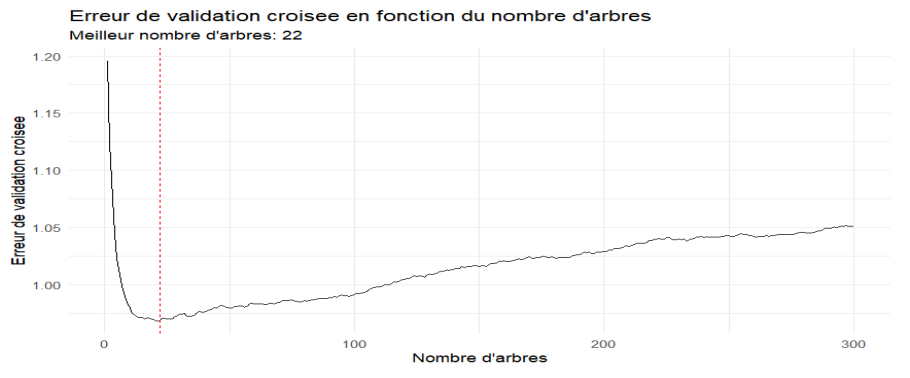
$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

Cette fonction est adaptée ici car la problématique relève d'un cas de classification binaire, analogue à l'estimation du nombre d'occurrences de sinistres en actuariat. Dans le modèle, la distribution binomiale est spécifiée pour refléter la nature binaire de la variable cible, et pour assurer la cohérence avec la fonction de lien logit.

Optimisation du modèle de Gradient Boosting

Nous avons fait varier les paramètres en ajustant le nombre d'arbres entre 10 et 500, tout en augmentant les hyperparamètres de 2 à 7. Le bag fraction a été testé à 0.5, 0.7 et 0.9. De plus, nous avons exploré un grid de shrinkage allant de 0.01 à 0.5. L'objectif était de déterminer les hyperparamètres optimaux en utilisant la distribution de Bernoulli afin d'identifier la meilleure configuration pour notre modèle.

FIGURE 13 – Nombre arbre vs Erreur de validation croisés



Source : Élaboration des auteurs à partir des données.

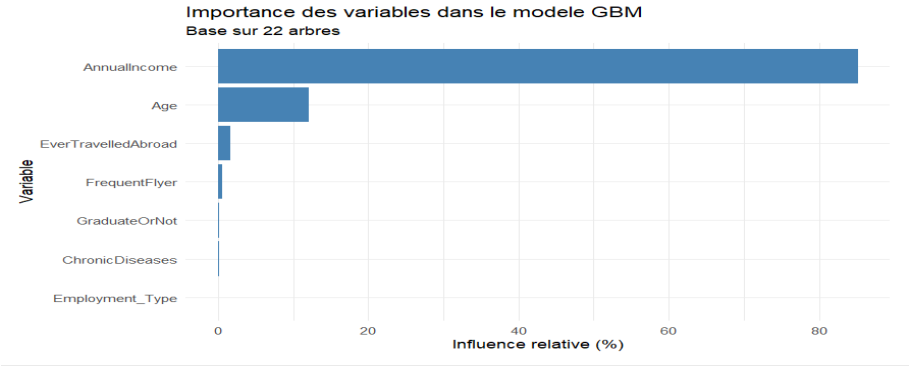
Après validation croisée et variation des hyperparamètres, nous avons obtenu les résultats suivants :

n_trees	interaction_depth	bag_fraction	shrinkage	best_iter	cv_error
22	3	0.9	0.2	21	0.9622771

Importance des variables dans le modèle Gradient Boosting

Nous obtenons des résultats similaires concernant les variables des modèles précédents. La variable la plus significative est le revenu, suivie de près par l'âge, puis par la variable EverTravelledAbroad.

FIGURE 14 – Importance des variables dans le modele de Gradient Boosting



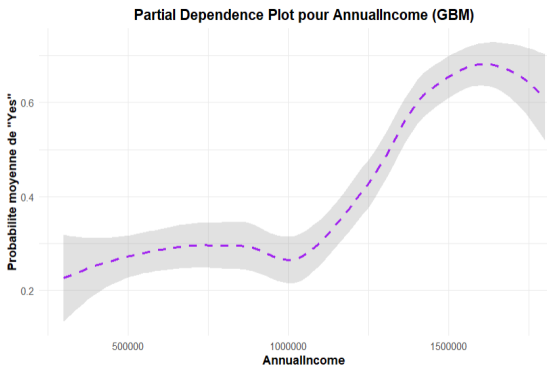
Source : Élaboration des auteurs à partir des données.

Analyse par dépendance partielle (PDP)

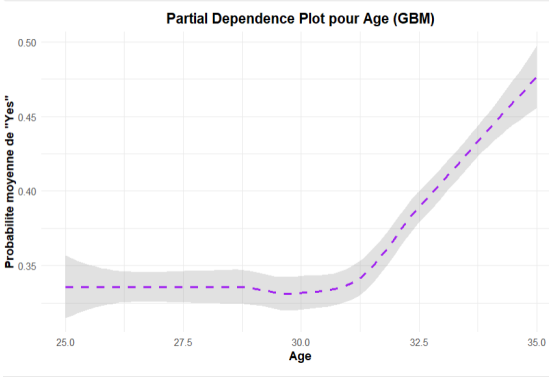
Revenu annuel (AnnualIncome) : La probabilité de souscription augmente nettement à partir d'un revenu de 1 000 000, jusqu'à se stabiliser vers 1 600 000. Les clients à haut revenu sont significativement plus susceptibles de souscrire.

Âge (Age) : La probabilité est faible autour de 30 ans, mais croît régulièrement au-delà. Les clients plus âgés manifestent une propension accrue à souscrire, probablement par prudence accrue ou habitudes d'achat différentes.

(a) Effet du revenu annuel



(b) Effet de l'âge



Source : Élaboration des auteurs à partir des données.

Comparaison des performances et choix du meilleur modèle

Afin d'identifier le modèle le plus performant pour prédire la souscription à une assurance voyage, nous avons comparé quatre approches supervisées : **Gradient Boosting**, **Bagging**, **Random Forest** et **Arbre élagué**. Les performances ont été évaluées sur l'échantillon test selon deux métriques complémentaires :

— **Accuracy** (précision globale)

— **AUC** (Area Under the ROC Curve), qui mesure la capacité de classement du modèle.

Analyse comparative

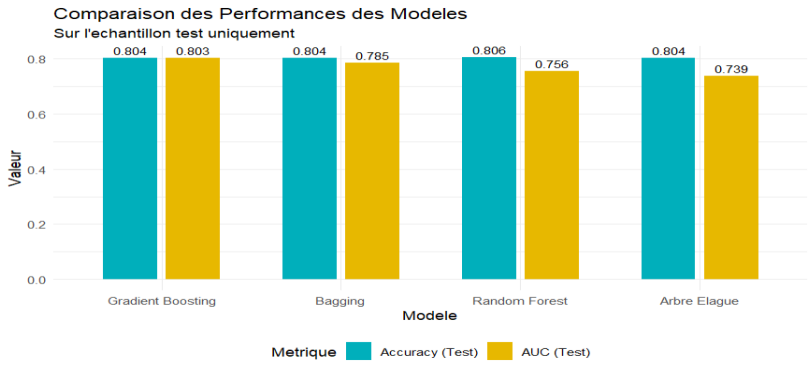
— Tous les modèles atteignent une **précision test très proche**, ce qui indique une bonne performance globale dans la classification binaire.

— La **meilleure performance globale** est généralement obtenue par les **méthodes ensemblistes** (Bagging, Random Forest, Gradient Boosting), qui surpassent l'arbre de décision simple.

— Le **Random Forest** et le **Gradient Boosting** montrent des performances particulièrement compétitives, confirmant l'intérêt des méthodes d'ensemble pour ce type de problématique actuarielle.

— Le modèle **Arbre élagué**, bien que plus interprétable, présente une performance légèrement inférieure, ce qui justifie l'utilisation des méthodes ensemblistes pour améliorer la capacité prédictive.

FIGURE 16 – Performance des différents modeles



Source : Élaboration des auteurs à partir des données.

Synthèse des résultats

L'**analyse exploratoire** des données a révélé que le **revenu annuel** (`AnnualIncome`), l'**expérience de voyage à l'étranger** (`EverTravelledAbroad`) et l'âge (`age`) constituent les variables les plus discriminantes pour prédire la souscription. Ces observations ont été confirmées par tous les modèles testés.

Concernant les **performances comparatives** :

— Les **méthodes ensemblistes** (Bagging, Random Forest, Gradient Boosting) surpassent systématiquement l'arbre de décision simple, confirmant l'intérêt de l'agrégation de modèles pour améliorer la robustesse et la généralisation.

— L'**arbre de décision élagué**, bien que moins performant, présente l'avantage d'une **interprétabilité maximale** en se concentrant uniquement sur le revenu annuel comme critère de segmentation principal.

— Le **Random Forest** et le **Gradient Boosting** offrent les meilleures performances globales, avec une capacité accrue à capturer les interactions complexes entre variables.

Implications actuarielles

Du point de vue actuariel, ces résultats suggèrent que :

1. Le **profil socio-économique** (revenu) est le facteur déterminant dans la décision de souscription, permettant une **segmentation claire de la clientèle**.
2. L'**expérience de voyage** constitue un indicateur comportemental pertinent pour affiner le ciblage marketing.
3. La **simplicité de l'arbre de décision** (basé principalement sur le revenu) offre un outil de communication efficace avec les équipes commerciales et marketing.

Perspectives d'amélioration

Plusieurs axes d'amélioration pourraient être explorés :

— **Enrichissement des données** : intégration de variables comportementales supplémentaires (historique d'achat, données de navigation web, saisonnalité des voyages).

— **Techniques avancées** : test de méthodes plus récentes (XGBoost, LightGBM, réseaux de neurones).

— **Optimisation métier** : calibrage des seuils de décision en fonction des objectifs business (optimisation du profit vs volume, coût d'acquisition client).

Conclusion

Ce travail illustre l'efficacité des méthodes d'ensemble en actuariat et fournit une base solide pour le développement d'outils d'aide à la décision en assurance voyage. La convergence des résultats entre les différents modèles concernant l'importance du revenu annuel renforce la robustesse des conclusions et offre une direction claire pour les stratégies de ciblage commercial.