LACTU2310 - Statistical Learning Methods for Insurance

Prédiction de souscription d'assurance voyage avec des méthodes ensemblistes

Davy Romaric KAMGA BOPDA

UCLouvain - ISBA
Superviseur: Karim Barigou

24 juillet 2025

Plan de la présentation

Introduction et Contexte

Problématique

- **Contexte :** Agence de voyages proposant une assurance voyage avec couverture Covid-19 en 2019
- **Données :** 2000 clients ciblés, seulement 36% ont souscrit (1 client sur 3)
- **Objectif :** Prédire la probabilité de souscription future en utilisant les techniques de modélisation vu dans le cours.

Variables disponibles

- **Sociodémographiques :** Age, Employment Type, GraduateOrNot, AnnualIncome, FamilyMembers
- **Médicales :** ChronicDiseases
- **Voyage :** FrequentFlyer, EverTravelledAbroad
- **Cible :** TravelInsurance (Oui/Non)

Challenge

Classification binaire avec 1987 individus - Améliorer le ciblage commercial

Preprocessing des données

Transformations effectuées

- **ChronicDiseases :** Recodage en variable binaire (0/1 $\rightarrow$ Oui/Non)
- **FamilyMembers :** Traitement en variable discrète
- **TravelInsurance :** Conversion en facteur pour la modélisation
- **Données manquantes :** Aucune détectée dans l'échantillon

Division des données

- **Entraînement :** 80% des observations (1590 individus)
- **Test :** 20% des observations (397 individus)
- **Validation :** Validation croisée k-fold pour l'optimisation des hyperparamètres

Qualité des données

Jeu de données complet et bien structuré, prêt pour la modélisation

Analyse univariée - Variables quantitatives

Table – Statistiques descriptives

Variable	Médiane	Moyenne	Min.	Max.
Age	29.00	29.65	25.00	35.00
AnnualIncome	900k	932k	300k	1.8M

Variables qualitatives - Répartitions

- **Secteur d'emploi :** 71% secteur privé, 29% gouvernement
- **Diplôme universitaire :** 85% diplômés, 15% non-diplômés
- **Voyage fréquent :** 21% voyageurs fréquents, 79% occasionnels
- **Voyage à l'étranger :** 19% ont voyagé, 81% jamais
- **Maladies chroniques :** 28% en ont, 72% n'en ont pas

Variable cible

Souscription : 36% Oui, 64% Non $\rightarrow$ Déséquilibre modéré des classes

Analyse bivariée - Relations avec la cible

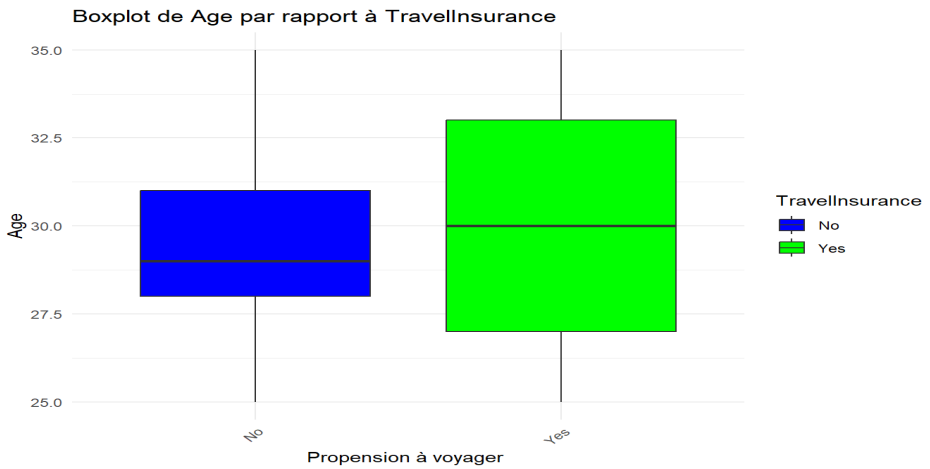


Figure – Age vs Souscription

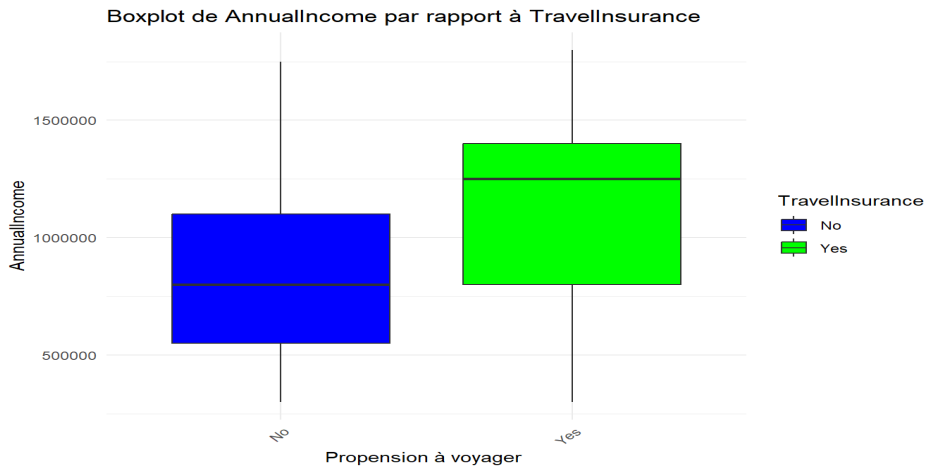


Figure – Revenu vs Souscription

Test du Khi² - Variables significatives ($p < 0.05$)

- **Employment Type :** p-value = 0.0000
- **FamilyMembers :** p-value = 0.0001
- **FrequentFlyer :** p-value = 0.0000
- **EverTravelledAbroad :** p-value = 0.0000

Arbre de décision - Optimisation du paramètre cp

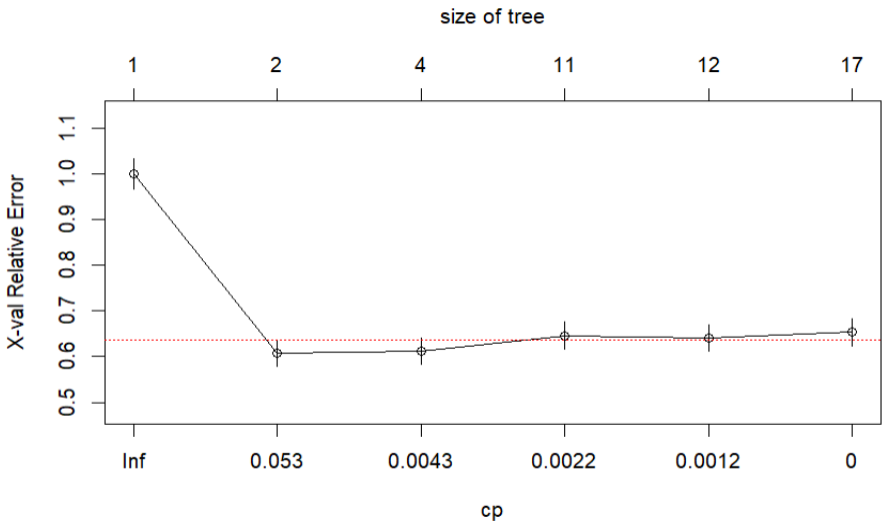


Figure – Courbe de validation croisée

Paramètres optimaux

- **cp optimal :** 0.007042254
- **Validation croisée :** 10-fold
- **Critère :** Minimisation erreur CV

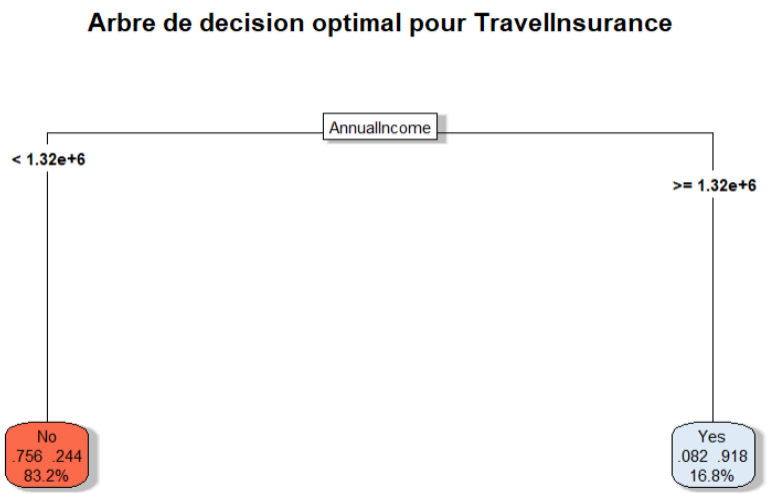
Résultats

- Convergence stable
- Compromis biais-variance optimal
- Arbre parcimonieux

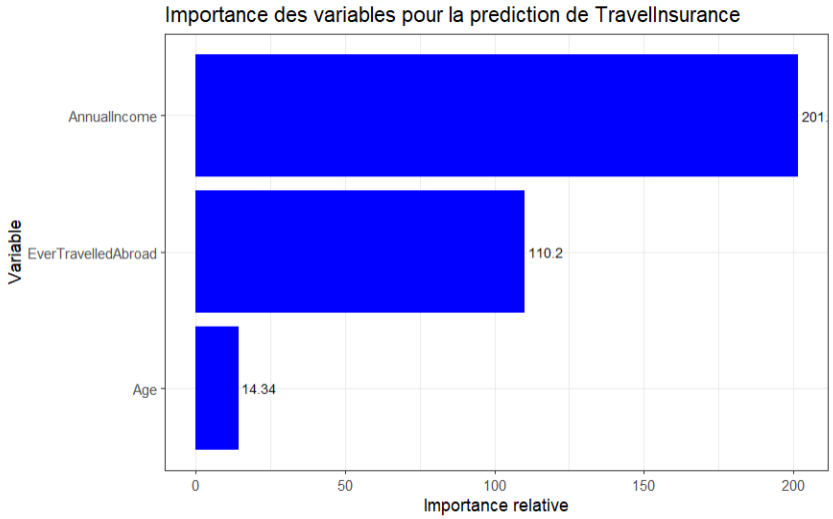
Cross-validation

La validation croisée confirme la robustesse du modèle avec une erreur stabilisée

Arbre élagué – Structure et importance



Arbre de décision élagué



Importance des variables

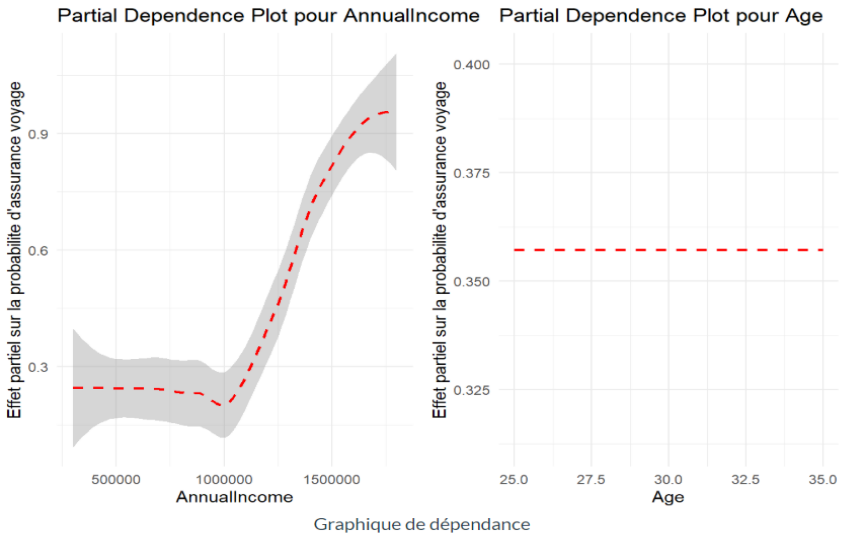
Variables clés

- **AnnualIncome** : 201 → facteur principal
- **EverTravelledAbroad** : 110.2 → facteur secondaire
- **Age** : 14.34 → effet marginal

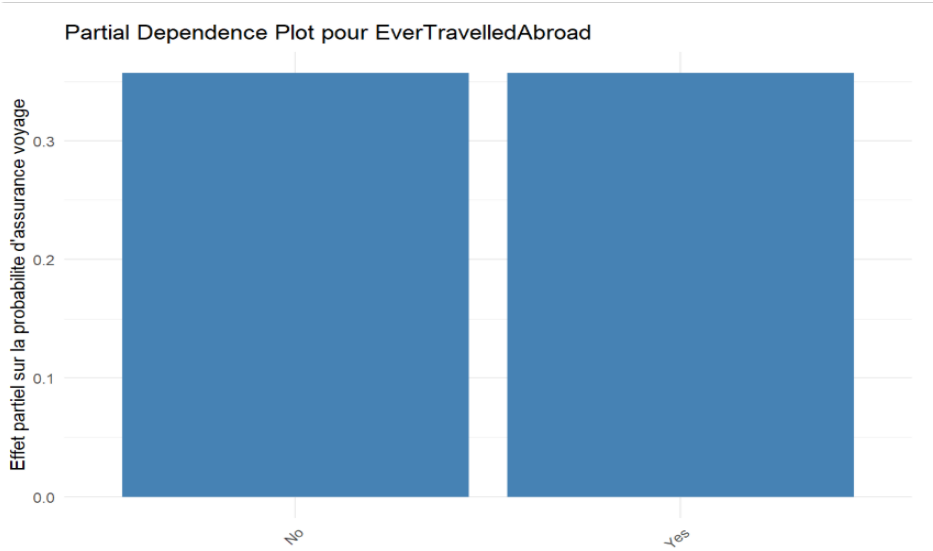
Interprétation

Le revenu constitue le critère de segmentation principal de l'arbre.

Arbre élagué –Graphiques de dépendance partielle (PDP)



Arbre de décision élagué



Importance des variables

Interprétation

Revenu annuel : la probabilité de souscription reste faible (30%) en dessous de 1M, augmente fortement entre 1M et 1,3M, et dépasse 90% au-delà de 1,5M. *Effet seuil marqué.*

Âge : aucun effet significatif détecté, la probabilité reste stable quel que soit l'âge.

EverTravelledAbroad : effet marginal, les modalités Oui/Non ont un impact similaire.

Bagging - Implémentation et convergence

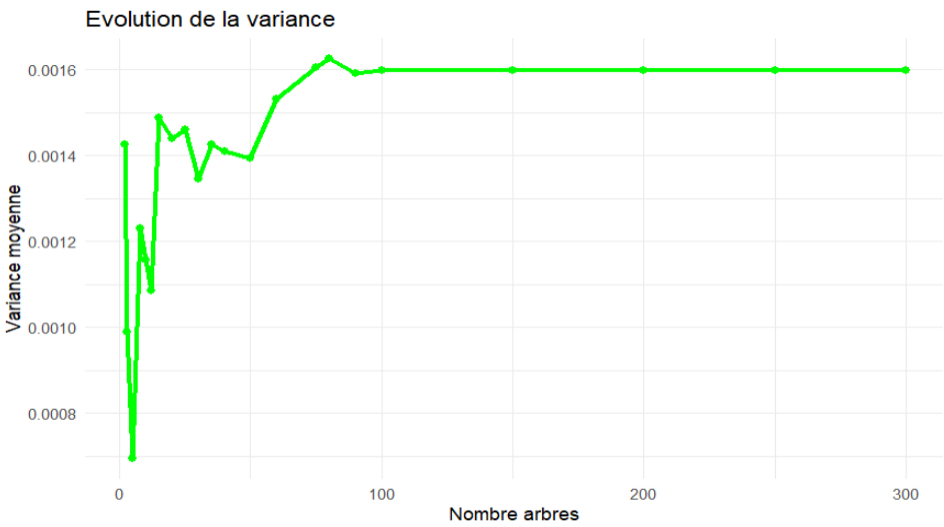


Figure – Évolution variance vs nb arbres

Configuration

- **Nombre d'arbres :** 100
- **Échantillons :** Bootstrap
- **Agrégation :** Moyenne probabiliste
- **cp :** Optimisé (arbre simple)

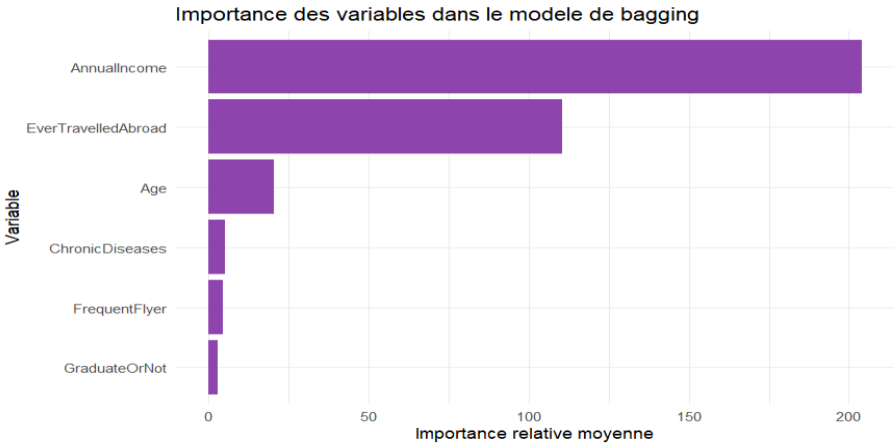
Performances

- **Accuracy :** 0.836 (stable)
- **Variance :** Diminution progressive
- **Stabilisation :** 30-50 arbres

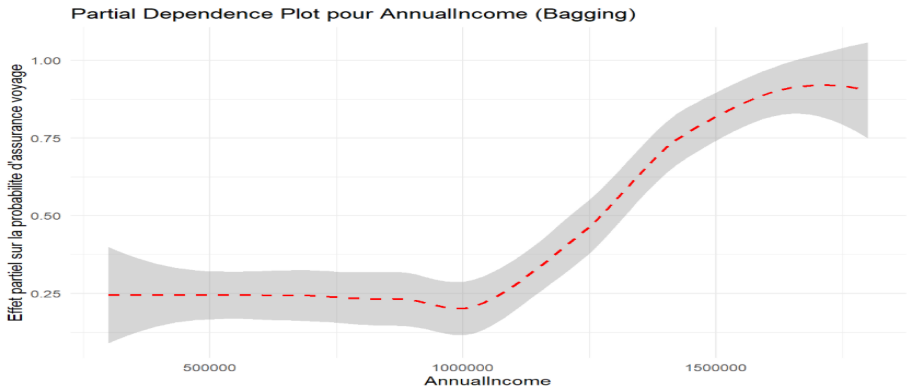
Cross-validation

Le bagging réduit efficacement la variance sans perte de performance

Bagging – Importance et dépendances partielles



Importance des variables



PDP – Effet du revenu

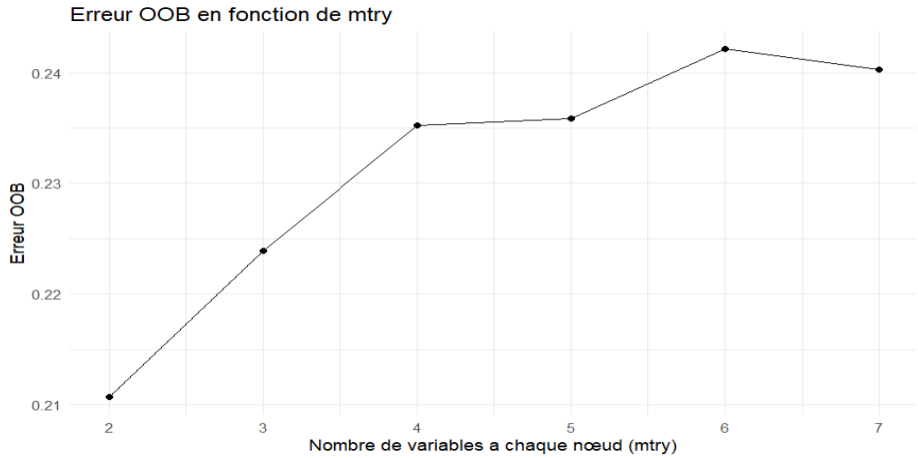
Analyse PDP

- **AnnualIncome** : effet seuil à 1M (30% $\rightarrow$ 90%)
- **Age** : effet croissant modéré après 30 ans
- **EverTravelledAbroad** : influence marginale

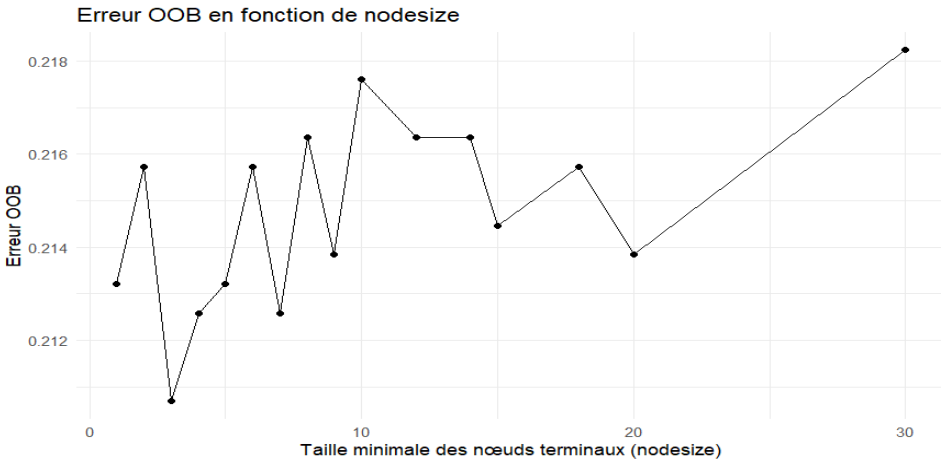
Validation croisée

Le modèle présente une **stabilité remarquable** : performance constante dès 30 arbres et réduction notable de la variance.

Random Forest – Optimisation des hyperparamètres



Erreur OOB vs mtry



Erreur OOB vs nodesize

Grille d'optimisation

- **ntree** : 10 à 500 (optimal 100)
- **mtry** : 2 à 7 (optimal = 2 ou 3)
- **nodesize** : 1 à 30 (optimal = 1 à 5)
- **Critère** : maximisation de l'AUC

Validation croisée

Performances optimales obtenues avec un **mtry faible** et **nodesize réduit** : meilleure diversité des arbres, erreur OOB minimale.

Random Forest - Résultats et interprétation

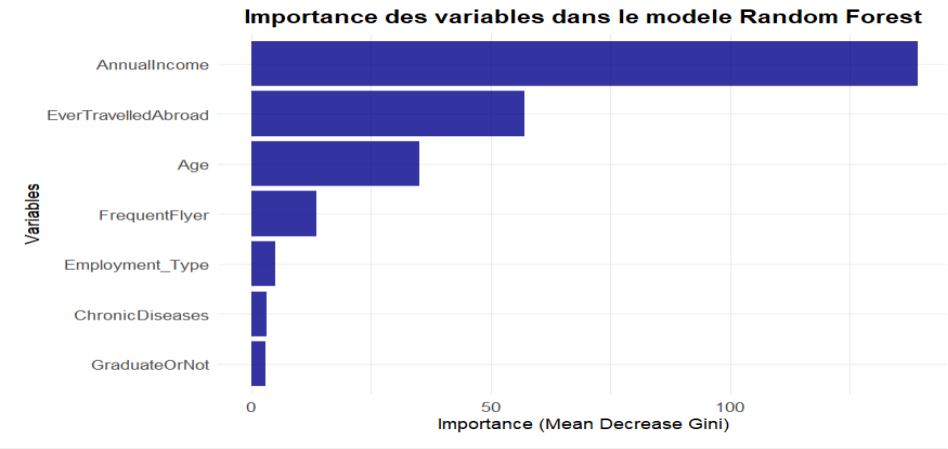


Figure – Importance des variables

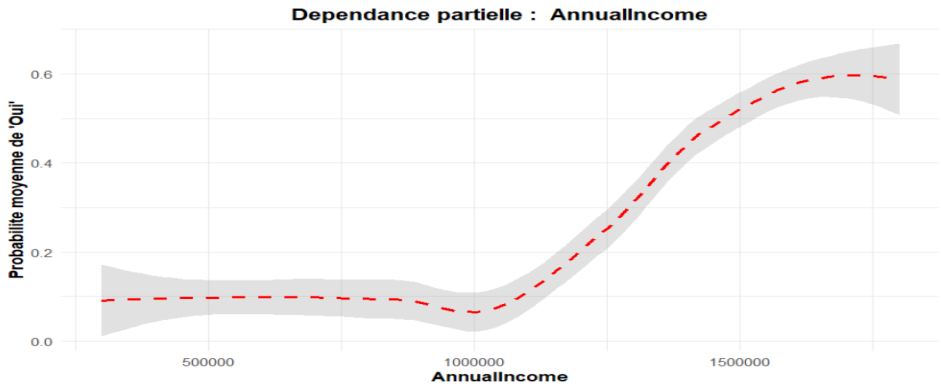


Figure – PDP - Revenu annuel

Analyse des dépendances partielles

- **Revenu :** Relation positive forte (seuil 1.2M)
- **Age :** Probabilité faible 25-28 ans, augmentation après 30 ans
- **Cohérence :** Avec analyses précédentes

Cross-validation

Performance optimale avec mtry=2-3, minimisant la corrélation entre arbres

Gradient Boosting - Optimisation et convergence

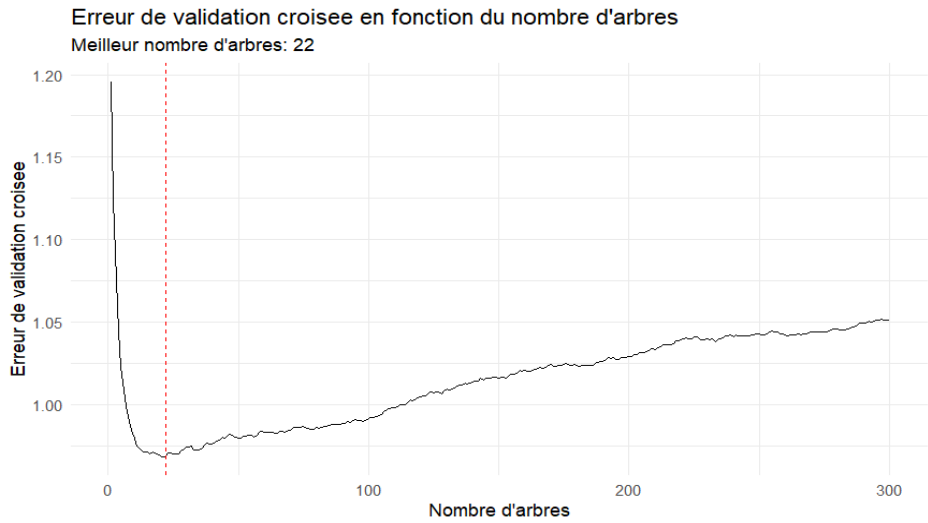


Figure – Évolution erreur vs nb arbres

Hyperparamètres optimaux

- **n_trees :** 22
- **interaction_depth :** 3
- **bag_fraction :** 0.9
- **shrinkage :** 0.2
- **best_iter :** 21

Performance

- **CV error :** 0.9623
- **Distribution :** Bernoulli
- **Fonction lien :** Logit

Cross-validation

Optimisation fine par grid search avec validation croisée extensive

Gradient Boosting - Importance et effets partiels

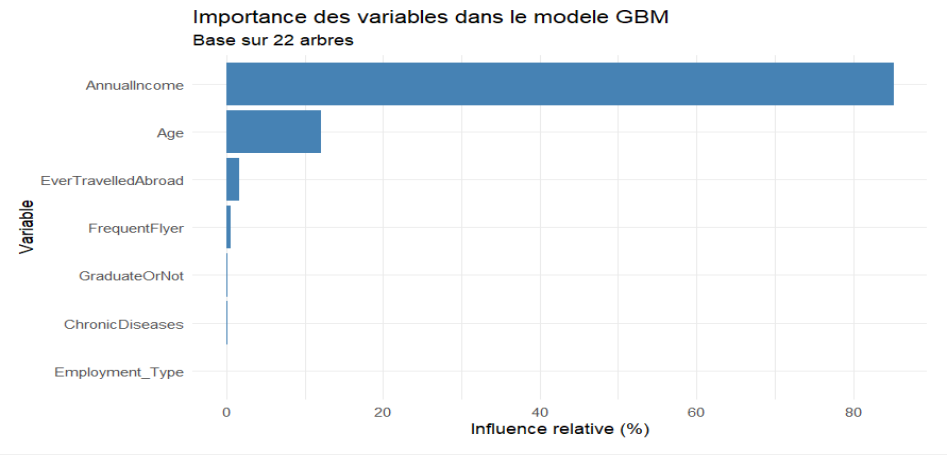


Figure – Importance des variables

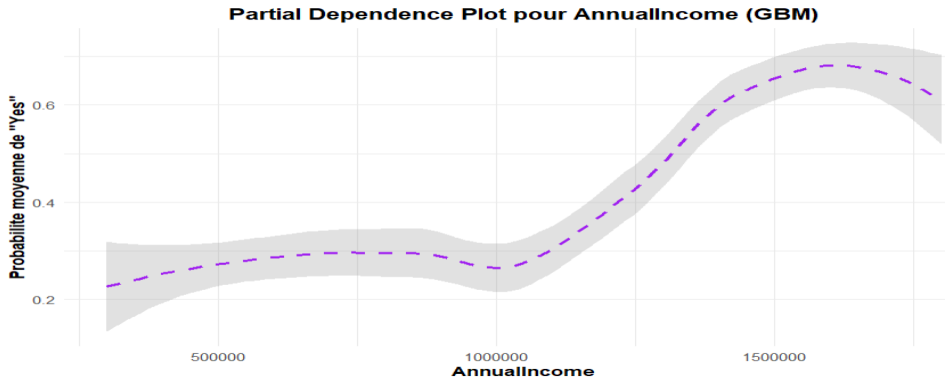


Figure – PDP - Effet du revenu

Analyse PDP confirmée

- **Revenu :** Augmentation nette à partir de 1M, stabilisation vers 1.6M
- **Age :** Probabilité faible vers 30 ans, croissance régulière au-delà
- **Cohérence :** Avec tous les modèles précédents

Cross-validation

Validation croisée confirme la robustesse avec early stopping à 21 arbres

Comparaison des performances

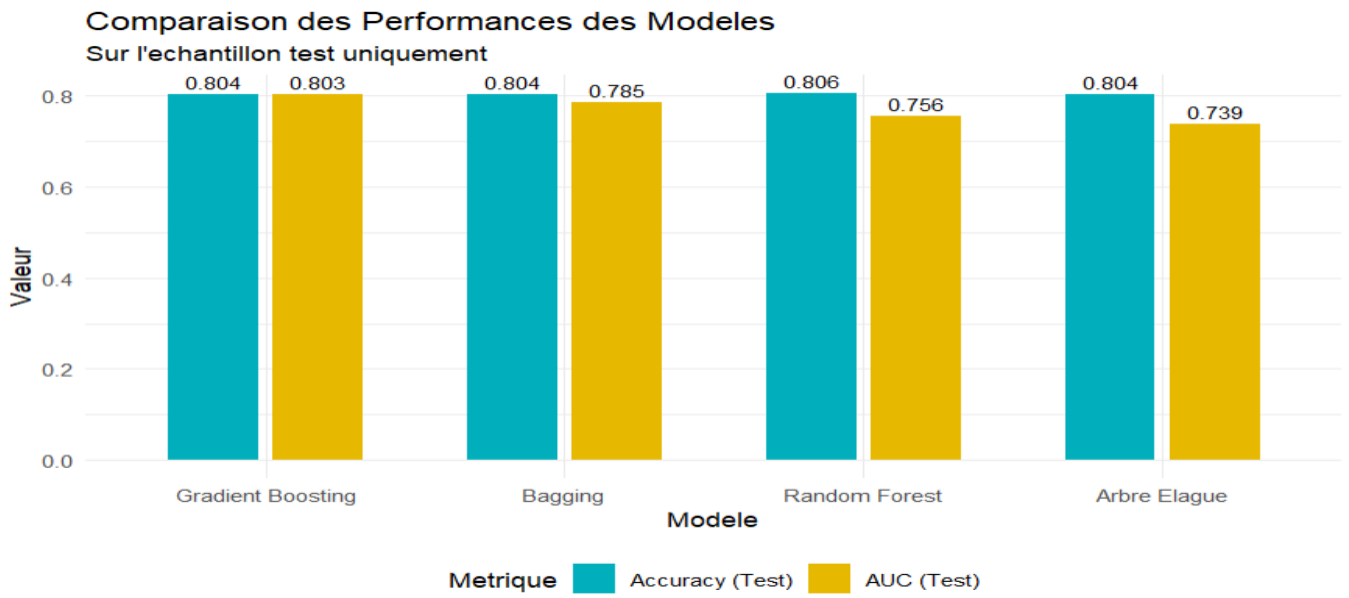


Figure – Performance des différents modèles (Accuracy et AUC)

Synthèse des résultats

- **Méthodes ensemblistes** surpassent l'arbre simple
- **Random Forest et Gradient Boosting :** Meilleures performances
- **Arbre élagué :** Moins performant mais plus interprétable
- **Consistency :** Toutes les méthodes identifient les mêmes variables clés

Conclusion et perspectives

Résultats clés

- **Variable principale :** AnnualIncome (effet seuil à 1M)
- **Variables secondaires :** EverTravelledAbroad, Age
- **Performance :** Méthodes ensemblistes supérieures
- **Robustesse :** Validation croisée confirme la stabilité

Implications actuarielles

- **Segmentation :** Ciblage clients haut revenu (>1M)
- **Stratégie :** Focus sur les voyageurs expérimentés
- **Communication :** Arbre simple pour équipes commerciales

Perspectives

- Enrichissement données comportementales
- Test méthodes avancées (XGBoost, Deep Learning)
- Optimisation seuils selon objectifs business

Merci pour votre attention