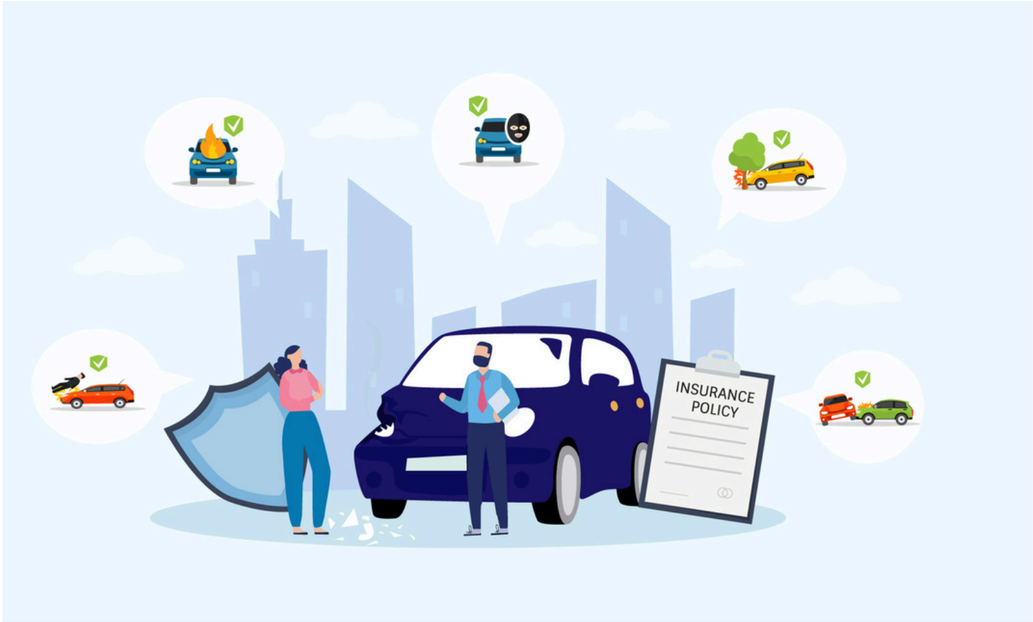
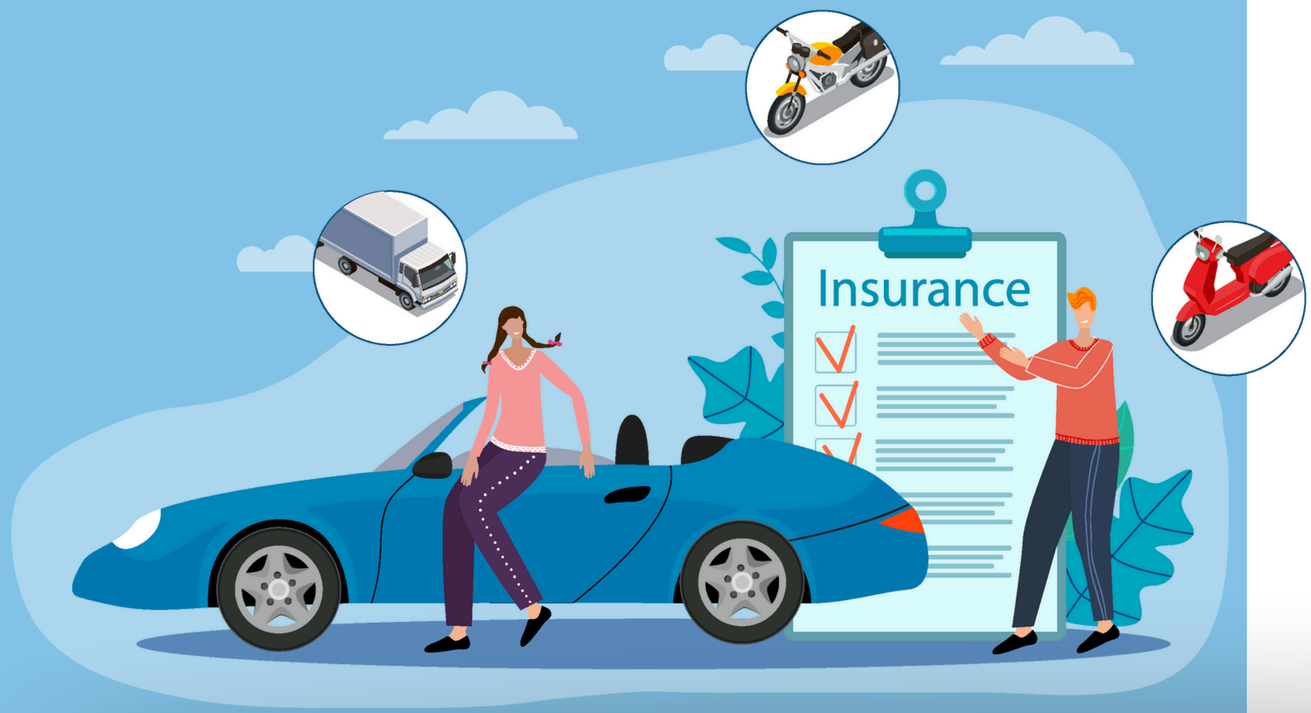
UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Prédiction du risque de résiliation des contrats auto



ISSA Karen
ADRIEN Davidson
TAVSIEV Askhab

Master 2 TIDE - Juillet 2025

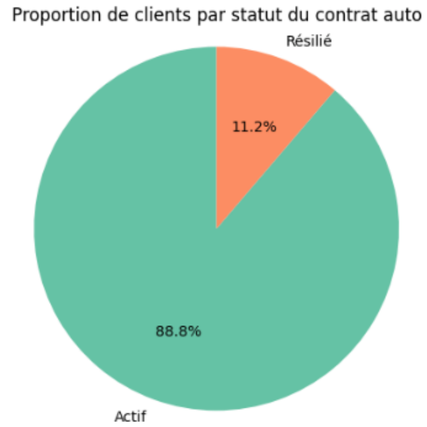


Sommaire

01. Objectif & Données

Objectif principal :

→ Prédire la résiliation des contrats auto à partir des données clients et des informations contractuelles.



Description des données :

→ 90 247 clients

→ 58 variables (données client, historique des contrats, sinistres, etc.)

→ Variable cible : **CONTRAT** (1 = résilié, 0 = actif)

Approche de la problématique métier :

→ Fidélisation des clients : identifier les facteurs conduisant à la résiliation des contrats, afin de l'anticiper et de mettre en place des actions ciblées.

Indicateurs créés :

→ Indice de responsabilité : proportion de sinistres responsables sur le total pour résumer le comportement (responsable vs non-responsable).

→ Écart de prime : ratio MTPAATTC / MTPAAREF, indicateur d'éventuelle remise ou surcoût appliqué.

→ Score résiliation pondéré : pondération des contrats résiliés selon leur importance (auto > habitation...)

Variables sans définition :

→ CRM (bonus-malus) → COEFCOMM (coefficient commercial),
→ COEFPFLT (coefficient inconnu) → CLIACTIF (client actif) :
à conserver.
→ U → NOCLIGES → MOTIFRSL → CDPRGES → AUTO4R → ETAT :
à exclure (trop de manquants, peu informatifs ou redondants).

02. Exploration & Préparation

Nature des variables :

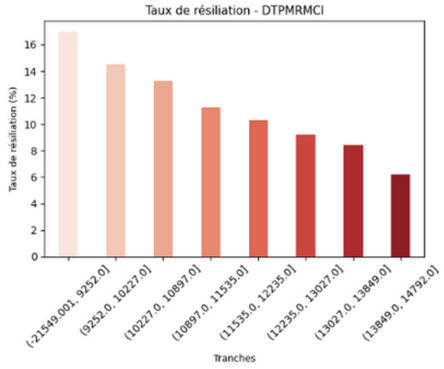
→ **Profil du client :** ancienneté du client (ANCCLI), sexe (CD_SEX), date de naissance (DT_NAI), etc.
→ **Informations sur le véhicule :** marque de la voiture (CDMARVEH), puissance fiscale (PUI_TRE), etc.
→ **Dates :** début du contrat (DTDBUCON), date du dernier mouvement (DTEFTMVT), date d'échéance (MMJECHPP), etc.
→ **Sinistralité :** nombre de sinistres responsables vs non-responsables sur différentes périodes (S_0_N, S_1_O, etc.).
→ **Score & tarification :** niveau de bonus-malus (NIVBM), score CRM, tarif de référence (NOTAREFF).
→ **Autres :** agent, région, etc.

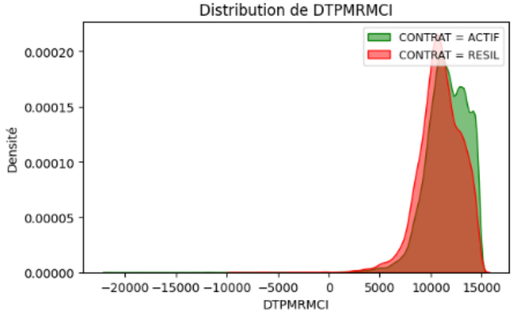
Attention particulière aux variables dont la signification n'est pas clairement définie ou absente.

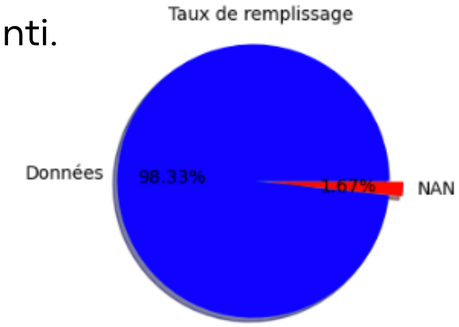
Tests réalisés :

→ **Significativité par rapport à la cible :** test du Chi² pour les var. quali., test de Student pour les variables quanti.
→ **Corrélation :** V de Cramér pour les var. quali., corrélation de Pearson pour les var. quanti.
→ **Analyse des relations avec la cible à l'aide de graphiques d'interactions avec la variable CONTRAT.**

Exemple : → DTPMRMCI (date de mise en circulation du véhicule)







Outliers & NaN :
→ Split : 80% train / 20% test
→ Données manquantes : < 2%
→ Détection des outliers (var. quanti.) : méthode de Tukey, KDE, stats. desc. (quantiles extrêmes)

Méthodes d'imputation :
→ Var. quali. : par **mode** (.transform à X_test)
→ Var. quanti. : par **médiane** (.transform à X_test)

03. Modélisation - Version 1

Étapes suivies :

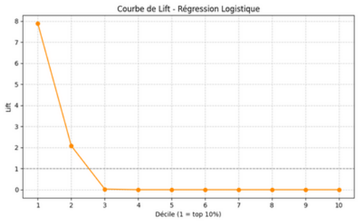
→ One-Hot Encoding
→ Target Encoding
→ RobustScaler

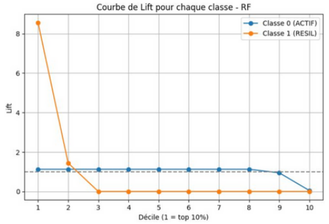
→ Réalisation d'un benchmark avec LazyPredict (~30 modèles testés)
→ Meilleures performances : Random Forest & XGBoost
→ Régression logistique comme modèle de base

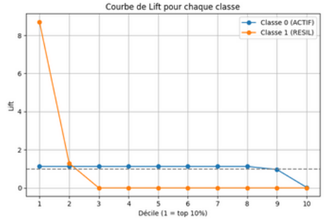
→ Sélection des variables via Stepwise, LASSO et Feature Importance
→ Contrôle de la multicolinéarité* pour éviter les variables redondantes ou corrélées

→ Métriques utilisées : AUC, Gini, F1-score global, F1-score de la classe 1
→ Optimisation des hyperparamètres via BayesSearchCV

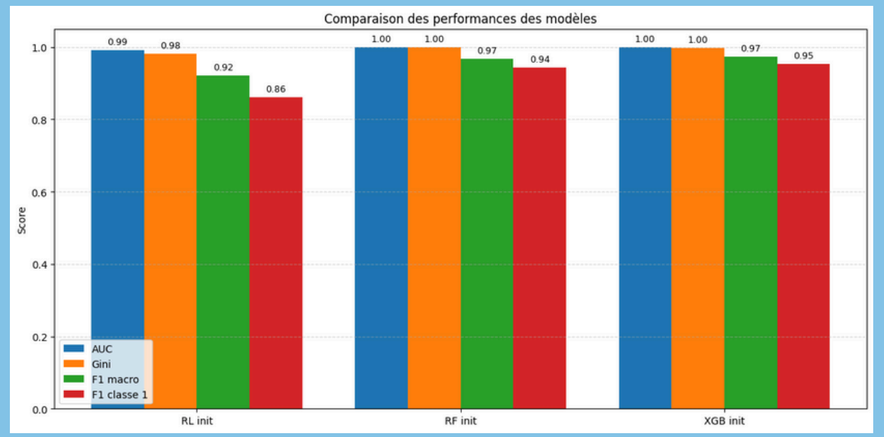
LIFT



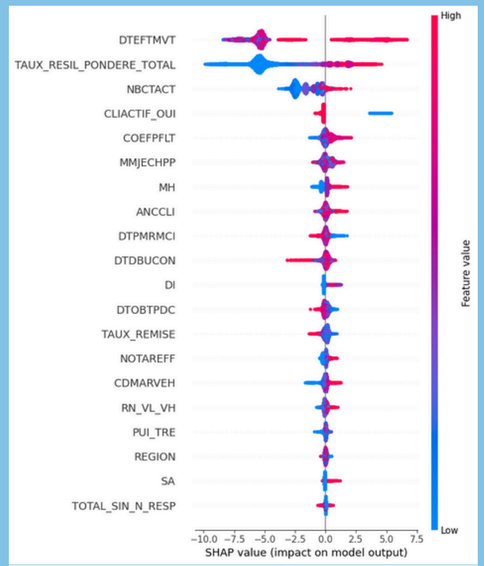


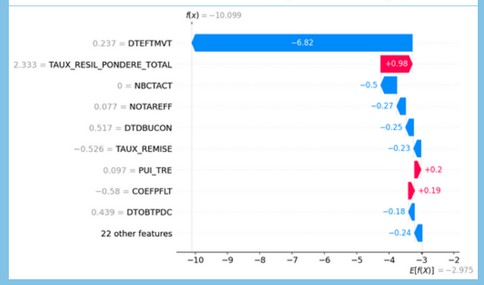


La variable **DTEFTMVT** semble causer une fuite de données en étant mise à jour juste avant la résiliation ; elle a donc été exclue des modèles pour éviter un biais et on fait une autre version du modèle plus fiable.

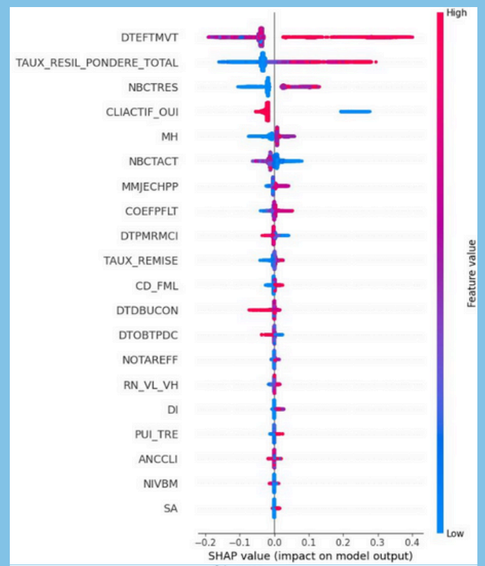


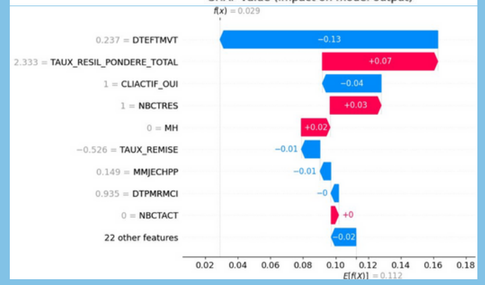
Boosting:





Bagging:





*pour la régression logistique

04. Modélisation - Version 2

Étapes suivies :

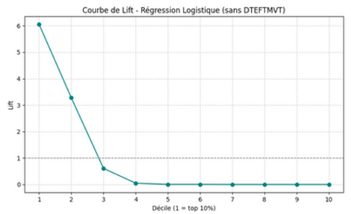
→ One-Hot Encoding
→ Target Encoding
→ RobustScaler

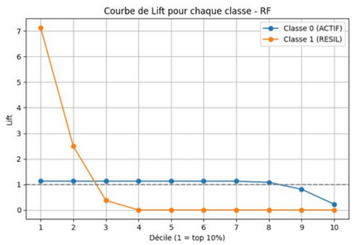
→ Réalisation d'un benchmark avec LazyPredict (~30 modèles testés)
→ Meilleures performances : Random Forest & XGBoost
→ Régression logistique comme modèle de base

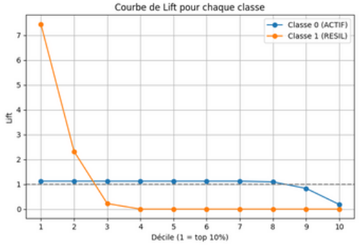
→ Sélection des variables via Stepwise, LASSO et Feature Importance
→ Contrôle de la multicolinéarité* pour éviter les variables redondantes ou corrélées

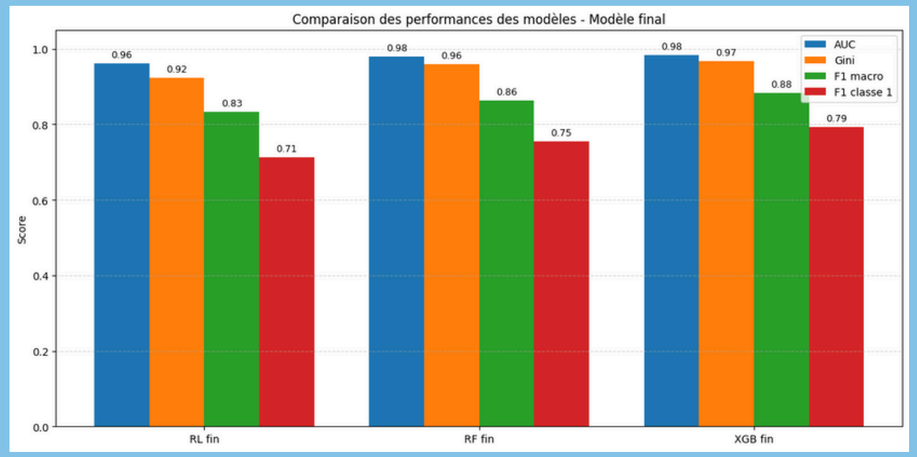
→ Métriques utilisées : AUC, Gini, F1-score global, F1-score de la classe 1
→ Optimisation des hyperparamètres via BayesSearchCV

LIFT

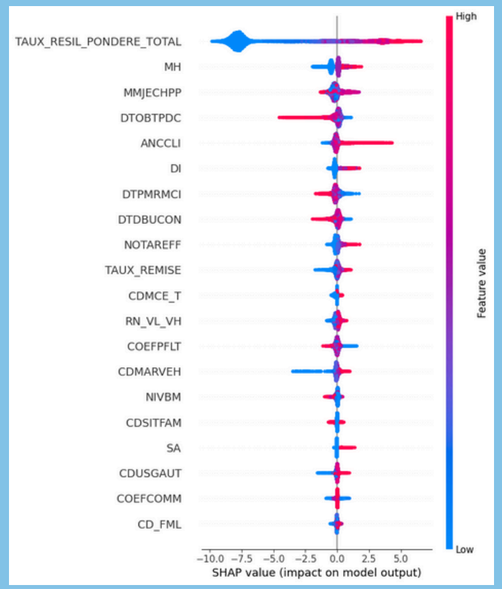


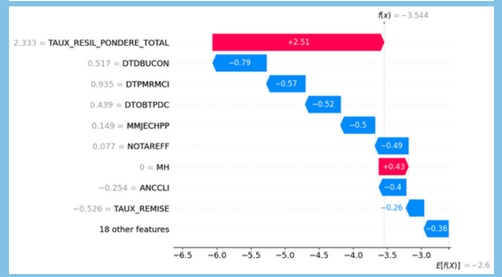




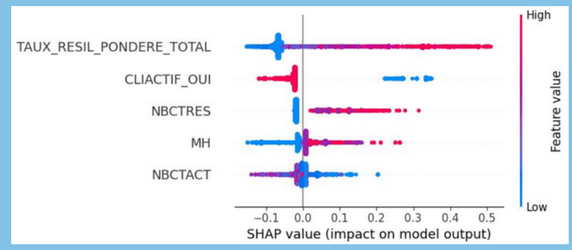


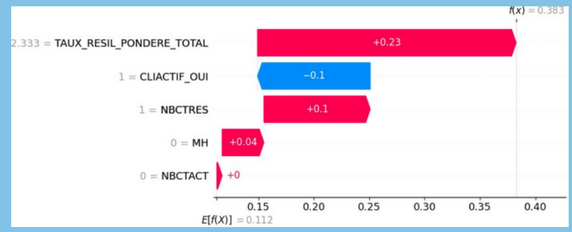
Boosting:





Bagging:





*pour la régression logistique

05. Conclusion

Deux versions comparées :
→ Initiale (avec la variable DTEFTMVT)
→ Finale (sans DTEFTMVT, suspectée de data leakage)

Constat sur la version initiale :
→ Scores très élevés (AUC et Gini proches de 1)
→ F1-score classe 1 = 0.94 (RF) et 0.95 (XGBoost)
→ Régression Logistique : AUC = 0.99, F1 classe 1 = 0.86
→ Performances surélevées → suspicion de fuite de données

Après suppression de DTEFTMVT (version finale) :
→ Scores plus crédibles
→ XGBoost : AUC = 0.98, F1 classe 1 = 0.79
→ Random Forest : AUC = 0.98, F1 classe 1 = 0.75
→ Régression Logistique : AUC = 0.96, F1 classe 1 = 0.71

Choix du modèle selon le besoin :
→ XGBoost si l'objectif est la performance maximale
→ Régression Logistique si l'on privilégie transparence, déploiement simple, interprétabilité