Projet - Predictive Models (TP5)

Samuel METIN (ENSIIE et UEVE (M2DS)
Daigo TELLIER-TERAWAKI (ENSIIE)

Décembre 2025

Introduction

Le but de ce projet est de prédire au mieux la résiliation (Churn) des forfaits d'un opérateur mobile. D'abord, on va proposer le/les meilleur(s) modèle(s) testés pour prédire la variable Churn. Ensuite, on étudiera les critères d'Independance, de Separation et de Sufficiency de ce(s) modèle(s) pour la variable sensible Gender.

Visualisation du dataset

Le dataset celldata.csv contient 8000 lignes et 11 colonnes dont la variable cible Churn.

	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	Salary	Churn
0	632	Germany	Female	50	5	107959.39	1	1	1	6985	1
1	649	France	Female	42	7	0.00	2	0	1	22974	0
2	595	France	Male	29	6	150685.79	1	1	0	87771	0
3	653	Spain	Male	35	6	116662.96	2	1	1	23864	0
4	559	Spain	Female	40	7	144470.77	1	1	1	18918	0

FIGURE 1 – *5 premières lignes de celldata.csv*

Seules les variables Geography et Gender sont catégorielles, et on les a encodées pour la suite. Ainsi en abandonnant la valeur France on a remplacé Geography par Geography_Germany et Geography_Spain, et en abandonnant la valeur Female on a remplacé Gender par Gender_Male.
Enfin, Churn est distribué selon 6391 valeurs 0 et 1609 valeurs 1.

Question 1 : Choix du/des modèle(s)

On a réalisé un benchmark sur 10 modèles de scikit-learn dont on a évalué les performances :

— `Bayes` : Naive Bayesian
— `LDA` : Linear Discriminant Analysis
— `QDA` : Quadratic Discriminant Analysis
— `LogReg` : Logistic Regression
— `KNN` : K-Nearest Neighbors (n_neighbors = 10)
— `Tree` : Decision Tree
— `RF` : Random Forest (max_depth=8)
— `ExTrees` : Extra Trees (max_depth=6)
— `AB` : Ada Boost (n_estimators=100)
— `GB` : Gradient Boosting (max_depth=5)

Les paramètres principaux de ces modèles ont été trouvés par grid search via l'optimisation de leur accuracy et de leur AUC sur le dataset de test. Voici les résultats des scores à travers 5-folds :

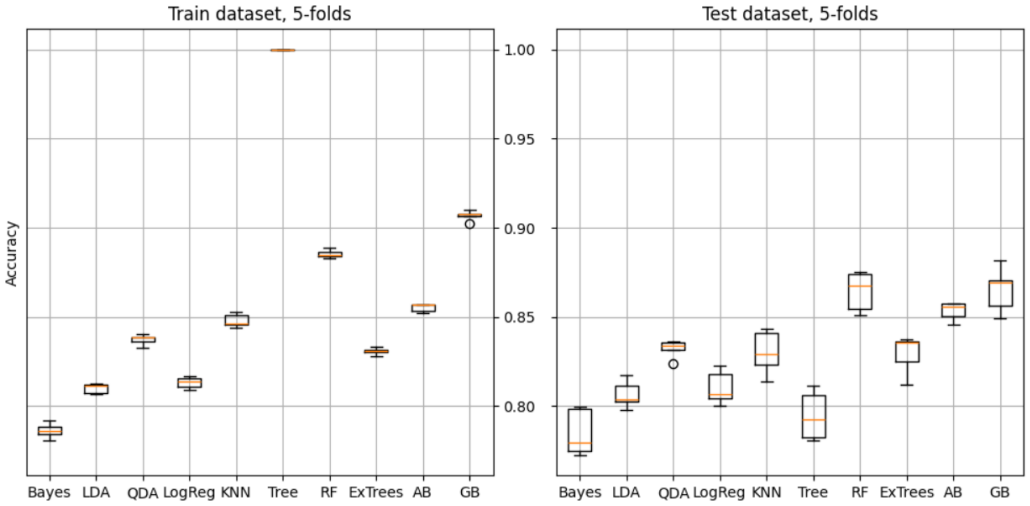


FIGURE 2 – *Accuracy des modèles*

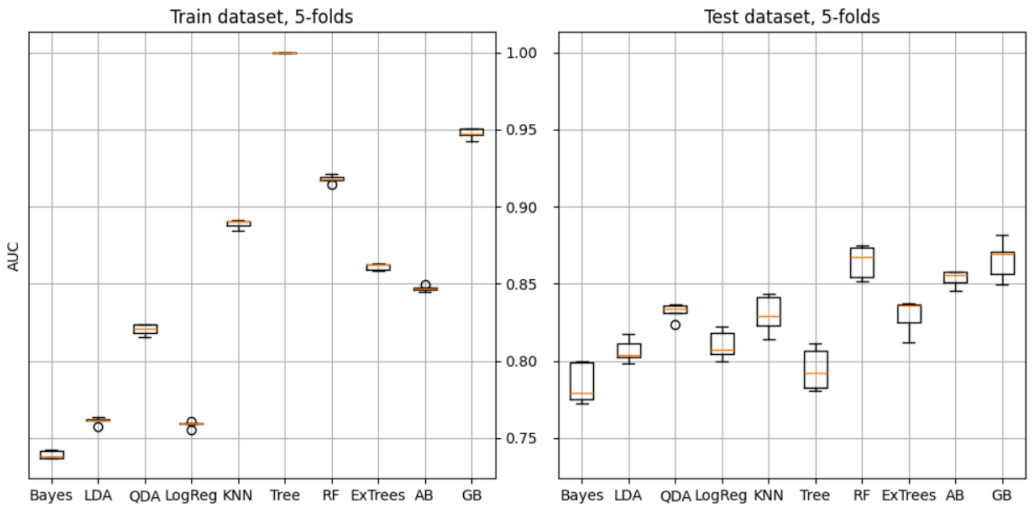


FIGURE 3 – *AUC des modèles*

Ainsi, les 2 modèles qui sont les plus perfomants sur le benchmark pour la prédiction de "Churn" sont le Random Forest (RF) et le Gradient Boosting (GB). En effet, en ajustant notamment la profondeur maximale des arbres, on arrive à des valeurs d'accuracy et d'AUC moyennes supérieures à 0.85 avec 5 folds sur le test. Le modèle Adaboost est aussi une variante pertinente, ce qui fait penser que la structure sous-jacente d'arbres est la plus compétitive dans ce contexte.

Ensuite on a évalué le nombre d'arbres nécéssaires et suffisants pour retrouver ces score de prédiction compétitifs. En effet, ce type de modèles de classification ne gagne pas en indéfiniment en précision (accuracy) notamment avec l'ajout d'arbres. De plus, un nombre trop important d'arbres dans les modèles Random Forest et Gradient Boosting augmente le temps d'entrainement sans gain de performances, au contraire.

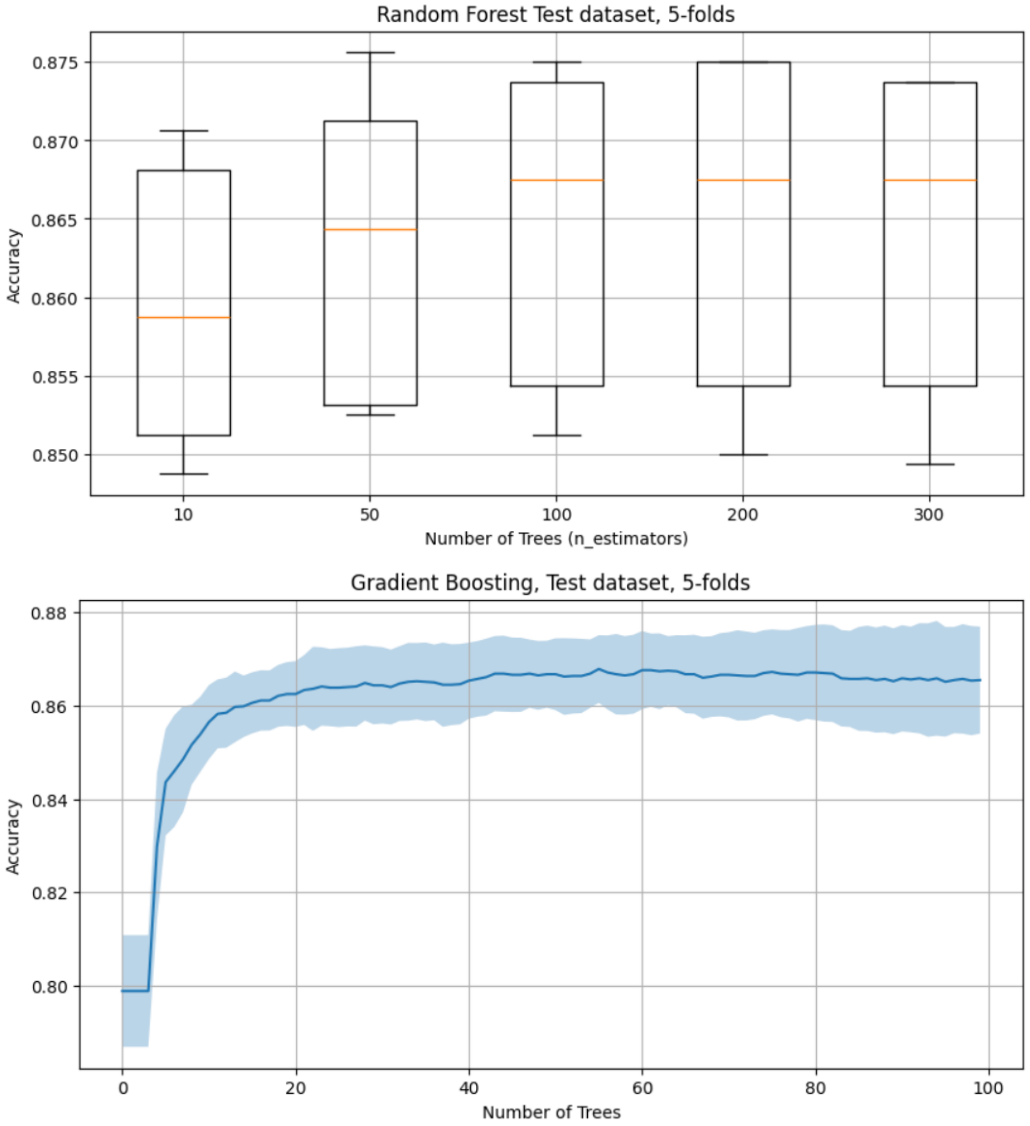


FIGURE 4 – *Evaluation du nombre d'arbres optimal pour chaque modèle*

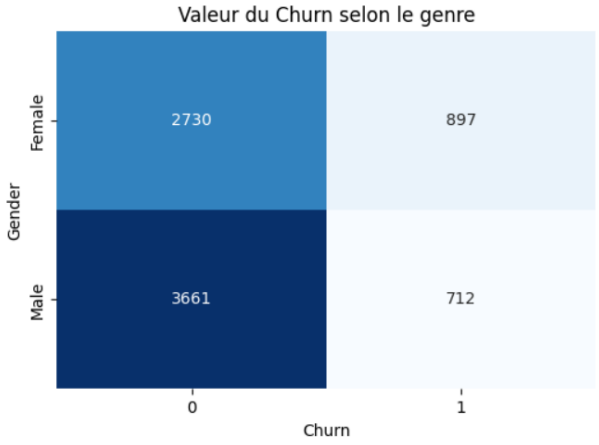
Ainsi pour le Random Forest un nombre d'arbres n_estimators=100 est suffisant, l'accuracy moyenne n'augmente pas en augmentant davantage ce paramètre, alors que ce nombre optimal est d'environ 25 pour le Gradient Boosting.

Conclusion

En nous basant sur les scores d'accuracy et d'AUC sur 5-folds on a comparé 10 modèles de classification pour prédire la variable "churn". On a trouvé que les modèles les plus performants sont : un Random Forest avec 100 arbres de profondeur maximale 8 et un Gradient Boosting de 25 arbres de profondeur maximale 5. Ces deux modèles atteignent leur accuracy moyenne maximale (plus de 86%) avec des paramètres raisonnables, et ils sont donc ceux que l'on recommande.

Question 2 : Independance, Separation et Sufficiency pour la variable sensible Gender

Tout d'abord nous avons choisi de regarder la distribution du Churn selon Gender afin de voir si il y a un biais statistique qui serait donc inévitable.



On remarque que les femmes sont fortement surreprésentées parmi les clients ayant un Churn=1 et on s'attend donc à ce que l'Independance ne soit vérifiée pour aucun des modèles. Voici les résultats en prenant les paramètres des modèles estimés précédemment :

Critère	Mesure	Female	Male
Independence	$P(\hat{y} = 1 \mid Gender)$	0.1295	0.0814
Separation	TPR $= P(\hat{y} = 1 \mid y = 1, Gender)$	0.4365	0.3994
	FPR $= P(\hat{y} = 1 \mid y = 0, Gender)$	0.0286	0.0199
Sufficiency	$P(y = 1 \mid \hat{y} = 1, Gender)$	0.8326	0.7986

TABLE 1 – Random Forest - métriques de fairness moyennes sur 5-folds

Critère	Mesure	Female	Male
Independence	$P(\hat{y} = 1 \mid Gender)$	0.1003	0.0550
Separation	TPR $= P(\hat{y} = 1 \mid y = 1, Gender)$	0.3496	0.2838
	FPR $= P(\hat{y} = 1 \mid y = 0, Gender)$	0.0183	0.0107
Sufficiency	$P(y = 1 \mid \hat{y} = 1, Gender)$	0.8612	0.8372

TABLE 2 – Gradient Boosting - métriques de fairness moyennes sur 5-folds

Conclusion

Les deux modèles ne respectent pas le critère d'Independance, car les prédictions positives du Churn pour les femmes sont significativement plus élevées que pour les hommes, ce qui était attendu vu la distribution du Churn selon le genre (biais statistique). La Separation elle n'est que partiellement respectée par Random Forest là où elle est totalement violée par le Gradient Boosting, ce qui montre que ce modèle accentue le biais sur le genre au niveau des erreurs de classification. Enfin la Sufficiency est respéctée par les deux modèles mais davantage par le Gradient Boosting, ce qui indique que les prédictions positives du churn sont plus fiables et moins biaisées entre les genres.