Olivier Lebert

Rapport d'Analyse du Churn Télécom

1. Contexte & objectif

Ce projet vise à modéliser le churn (désabonnement) d'un opérateur télécom afin d'identifier les clients à risque et d'activer des actions de rétention ciblées (ex. offres promotionnelles, avantages fidélité, etc …). L'ambition est double : d'une part détecter le maximum de churners, d'autre part optimiser les coûts liés aux offres envoyées à tort aux clients fidèles.

2. Données & préparation

Jeu de données : 3333 clients et 20 variables issus de kaggle et provenant d'un opérateur telecom aux USA.
Variables clés : minutes jour/soir/nuit, appels, plan international, boîte vocale, nombre d'appels au service client, État, etc.
La cible Churn est booléenne (14,5% de churn = 1).

Préparation effectuée :

• Encodage binaire : International plan et Voice mail plan.

• Suppression de variables redondantes (charges dérivées des minutes, etc.).

• Split 80/20 stratifié (train/test).

• Standardisation des variables numériques (utile pour la LogReg, non nécessaire pour RF).

• Gestion du déséquilibre via class_weight='balanced'.

3. Analyse exploratoire – principaux constats

Comme souligné précédemment, le taux de churn ≈ 14,5% est mis en perspective sur un histogramme (***Figure_1***). Parmi les variables le plus fortement associées au churn (Random Forest), on retrouve principalement le 'total day minutes' (très forte), 'Customer service calls', 'International plan', 'Total eve minutes', ce qui semble assez cohérent vis-à-vis de l'utilisation et des problèmes fréquemment rencontrés dans la télécommunication (surfacturation, plans avantageux chez d'autres opérateurs, service client défectueux, …). Enfin, un dernier histogramme et une visualisation via Plotly nous permet de distinguer les états à risque plus élevé - CA, NJ, TX, MD, SC, MI, MS, WA, NV, ME (***Figure_8*** & ***Figure_10***).

4. Modélisation & performances globales

Deux modèles ont été entraînés : Régression Logistique et Forêt Aléatoire (Random Forest).
Performances sur le jeu de test (probabilités et ROC évaluées) :

La régression logistique présente une AUC = 0.84, ce qui en fait déjà un bon séparateur global mais Random Forest, AUC = 0.94, semble encore proposer de meilleures performances en vue de classer les churners vs non-churners. Néanmoins, quel que soit le modèle employé, le seuil de décision reste une décision subtile afin de déterminer la précision et le rappel, et donc les coûts de campagne envisagés (***Figure_2***).

5. Métriques – définitions et formules

True Positive (TP) = churner correctement prédit (prédit comme 1 correctement)
False Positive (FP) = fidèle prédit churner (prédit comme 1 mais en réalité 0)
False Negative (FN) = churner manqué (prédit comme 0 mais en réalité 1)
True Negative (TN) = fidèle prédit comme tel (prédit comme 0 correctement)

La Précision, vrais positifs parmi ceux prédits comme tels
= TP / (TP + FP), représente la fiabilité des prédictions positives, ce qui permet d'éviter de gaspiller des offres.

Le Rappel, Recall, ou True Positive Rate (TPR), les vrais positifs prédits parmi tous les positifs
= TP / (TP + FN), représente la proportion de churners découvert, ce qui permet d'éviter d'en rater.

F1-score, meilleur compromis entre précision et rappel = 2 × (Précision × Rappel) / (Précision + Rappel) compromis unique quand les deux importent.

False Positive Rate (FPR), les faux positif prédits parmi tous les négatifs
= FP / (FP + TN), représente la proportion de fidèles que l'on contacterait à tort.

La Spécificité,ou True Negative Rate (TNR), les vrai négatifs prédits parmi tous les négatifs
= TN / (TN + FP) = 1 – FPR

Accuracy, l'exactitude des prédictions
= (TP + TN) / (TP + FP + TN + FN)

ROC-AUC = aire sous la courbe *Receiver Operating Characteristic* (TPR vs FPR pour tous les seuils), représente la qualité de séparation globale par rapport à la diagonale qui correspond au hasard.
Mesure globale de séparation des classes, indépendante du seuil.
Plus on est en haut à gauche, mieux c'est (haut TPR et faible FPR).

6. Choix de seuil – résultats observés

6.1 Régression Logistique

Matrice de confusion (***Figure_5***) au seuil ≈ 0.55 à F1-max observé ((***Figure_3***) : TN=466, FP=104, FN=25, TP=72
Rappel = 72 / (72 + 25) ≈ 0.74
Précision = 72 / (72 + 104) ≈ 0.41
F1 = 2 × (0.74 × 0.41) / (0.74 + 0.41) ≈ 0.53
Spécificité = 466 / 570 = 0.82 → FPR = 1 – Spécificité = 0.18
Accuracy = 538 / 667 ≈ 0.81
% de prédits positifs = 176/ 667 ≈ 26.4%
FP parmi les prédits positifs = 104 / 176 = 59%
TP parmi les prédits positifs (précision) = 41%

→ Le rappel est bon (74%) mais la précision est faible (41%), il y a donc beaucoup d'offre envoyées à tort. On a une assez bonne couverture des churners mais comme 59% des personnes contactées ne churnent pas, le coût marketing assez élevé et l'utilisation du budget n'est pas très efficace.

Si l'objectif est de maximiser le rappel, abaisser le seuil (≈ 0.40 par exemple) augmenterait encore la détection des churners, mais au prix d'une précision encore plus faible. La LogReg resterait donc surtout utile pour l'explicabilité des drivers dans ce contexte (permettant d'observer les valeurs des coefficients), mais est inférieure à Random Forest en terme de séparation globale (AUC plus petite).

6.2 Random Forest

Matrice de confusion (***Figure_6***) au seuil ≈ 0.26 à F1-max observé (***Figure_4***) : TN=548, FP=22, FN=15, TP=82
Précision = 82 /(82 + 22) ≈ 0.79
Rappel = 82 /(82 + 15) ≈ 0.85
F1 = 2 × (0.79 × 0.85) / (0.79 + 0.85) ≈ 0.82

Spécificité = 548/570 = 0.96 → FPR = 1 – Spécificité = 0.04
Accuracy = 630/667 ≈ 0.94
% prédits positifs = 104/667 = 15,6%
FP parmi les prédits positifs = 22/104 = 21%
TP parmi les prédits positifs (précision) = 79%

→ Très bon compromis : peu de faux positifs pour une couverture élevée.

Variante rappel prioritaire avec contrainte de précision ≥ 60%
Seuil ≈ 0.17, matrice de confusion (***Figure_7***) : TN=519, FP=51, FN=11, TP=86
Précision = 86 / (86 + 51) ≈ 0.63
Rappel = 86 / (86 + 11) ≈ 0.89
F1 = 2 × (0.63 × 0.89) / (0.63 + 0.89) ≈ 0.74
Accuracy = 605 / 667 ≈ 0.91
Spécificité = 519/570 = 0.91 → FPR = 0.09
% prédits positifs = 137/667 = 20.5%
FP parmi les prédits positifs = 51/137 = 37%
TP parmi les prédits positifs (précision) = 63%

Augmenter le rappel marginalement (on passe de 0.85 à 0.89) fait baisser substantiellement la précision (on passe de 0.79 à 0.63). En d'autres termes, on ne sauve que 4 clients en plus (TP= 86 vs 82 au seuil 0.26) alors qu'on a près de 30 fausses alertes en plus (FP = 51 vs 22 au préalable). Chaque client en plus détecté coute donc presque 8 fausses alertes en plus, ce qui montre bien que le seuil F1 permet effectivement d'éviter pas mal de gaspillage.

Approche Top-k : si le budget ne permet de cibler que 10% de la base client, on sélectionne chaque semaine les 10% des clients ayant la proba la plus élevée (seuil = quantile 90% des probas), c'est souvent plus simple et plus réaliste d'un point de vue opérationnel.

7. Recommandations d'exploitation

Vu les performances plus élevées, on préfèrera Random Forest (AUC 0.94) comme modèle de production. On distingue 3 modes de pilotage en fonction de l'objectif :

- Compromis (recommandé) : seuil ≈ 0.26 (F1-max)
  → très bonne précision (≈ 0.79) et rappel élevé (≈ 0.85).
- Rappel prioritaire : seuil ≈ 0.17 avec précision ≥ 60%
  → rappel ≈0.89 pour rater très peu de churners, au prix de plus de FP.
- Budget fixe (Top-k) : 10% signifie que le seuil = quantile 90% des probas.
  → simple et aligné à l'enveloppe marketing.

Intégration CRM : score de probabilité quotidien/hebdomadaire des clients actifs selon l'un des 3 modes, mise en relation en fonction des politiques internes (canal de communication prioritaire, exclusion, fréquence maximale), personnaliser l'offre en fonction de l'importance des variables (***Figure_9***), des motifs spécifiques et du profil de consommation, enregistrer chaque contact pour pouvoir observer l'efficacité des interventions.

On pourrait dès lors imaginer de réaliser des tests d'hypothèses (A/B ou test & control), de créer des KPIs (coût par client retenu, ROI, taux de rétention, …) et ajuster le seuil (ou la proportion de client visés) progressivement en fonction des résultats.

8. Implications, limites & bonnes pratiques

Pour rappel, un seuil plus bas augmente le rappel (moins de churners ratés) mais aussi le FPR (plus d'offres inutiles). Par ailleurs, en plus de ré-entraîner le modèle régulièrement, il faut également veiller à tuner les hyperparamètres (max_depth, min_samples_split, max_features…) pour optimiser le Random Forest. Notons aussi que le churn peut varier par segment (offres, usage) : envisager des seuils différenciés ou des modèles par segment pourrait également améliorer la performance des prédictions. Enfin, pour un bon pilotage, il serait préférable d'aller au-delà du scoring en observant qui réagit à une offre ou non ainsi que les circonstances précises de l'occurrence des churns afin de mieux gérer le timing des interventions.

9. Conclusion

La Random Forest surclasse la Régression Logistique en séparation globale (AUC 0.94 vs 0.84) et offre d'excellents compromis selon le seuil. Deux points d'opération recommandés : 0.26 (F1-max) pour un équilibre robuste, et environ 0.17 quand la priorité est de sauver un maximum de clients en gardant un minimum de précision. Le choix final dépendra du budget marketing et du coût relatif d'un churn raté (FN) par rapport à celui d'une offre inutile (FP).

Olivier Lebert

Figure_1 :

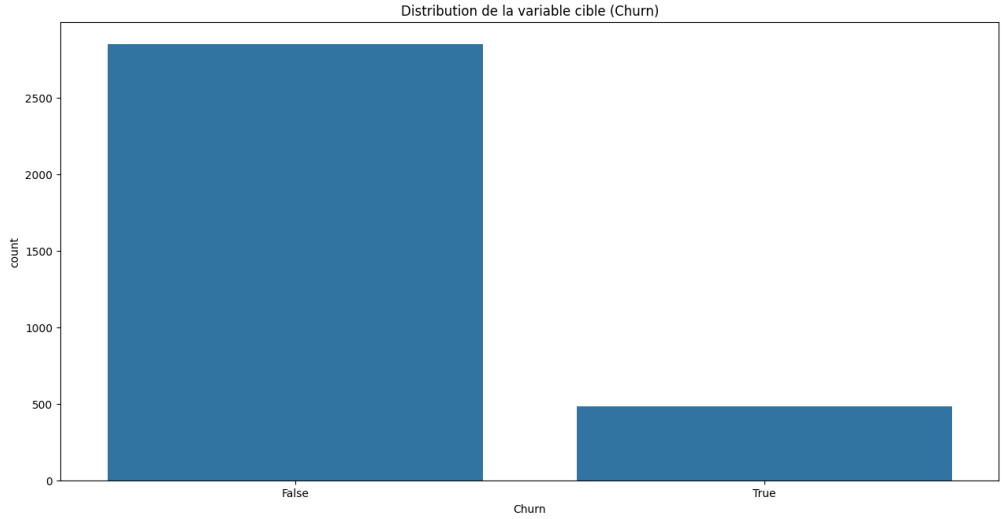

Figure_2 :

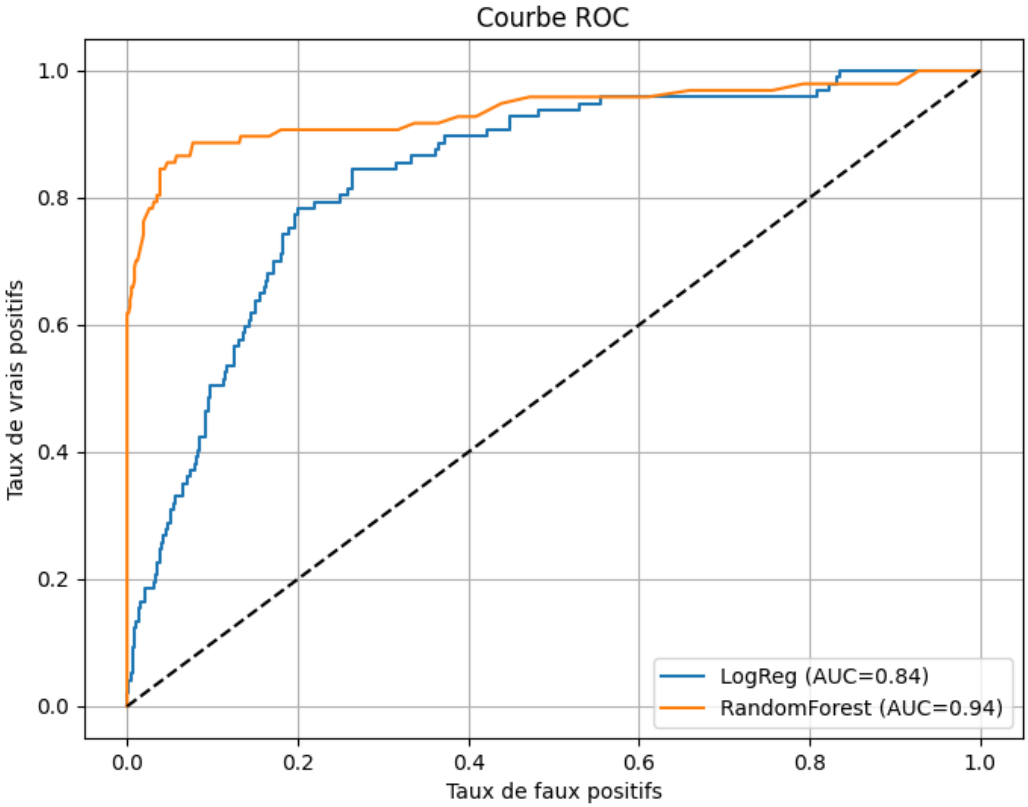

Figure_3 :

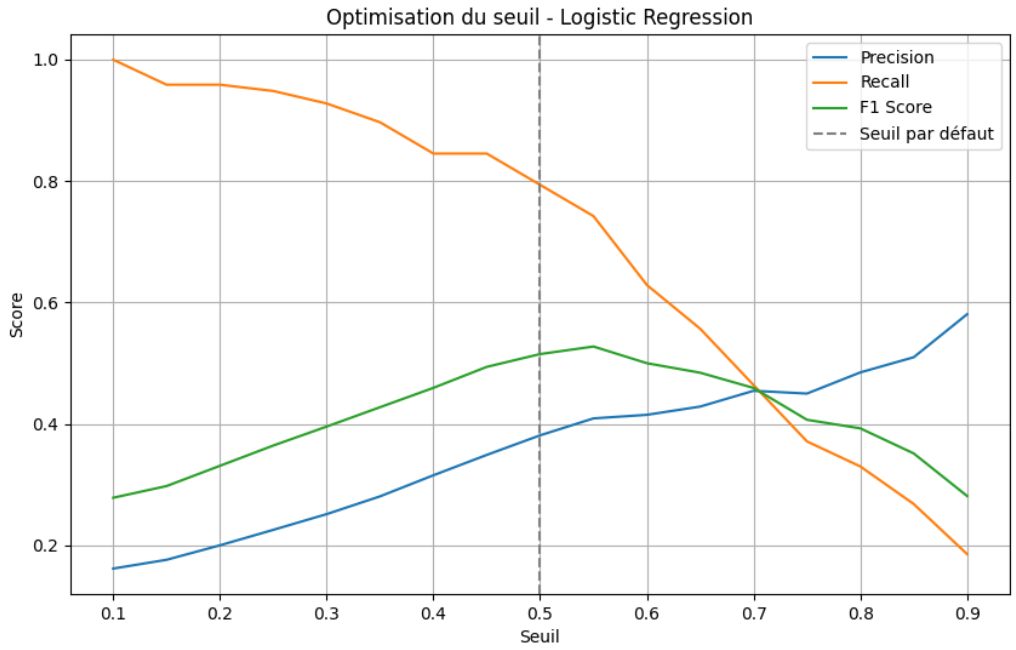

Figure_4 :

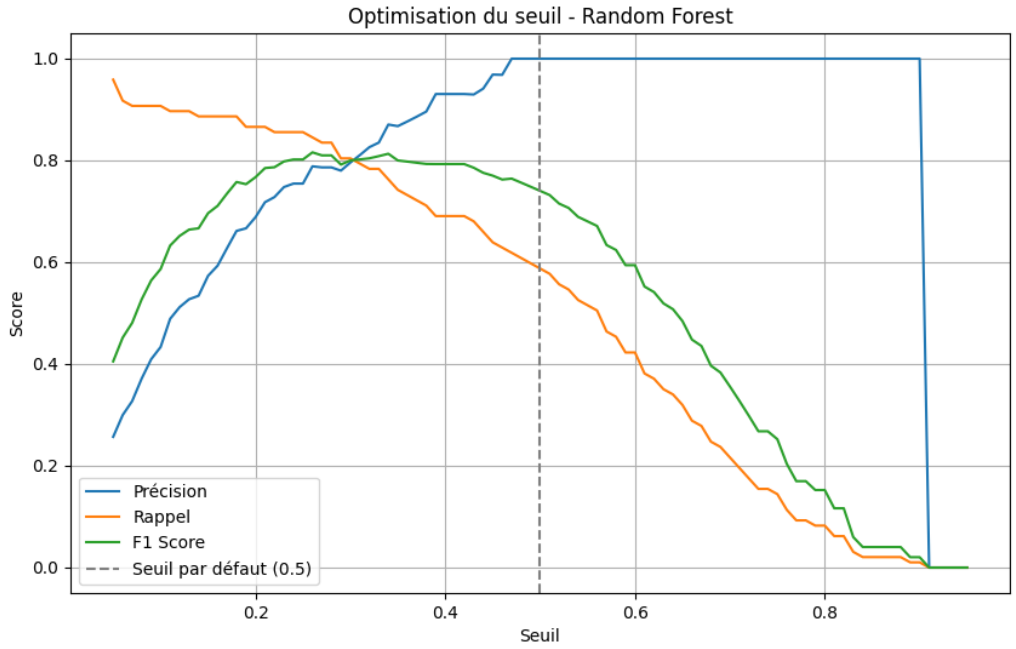

Figure_5 :

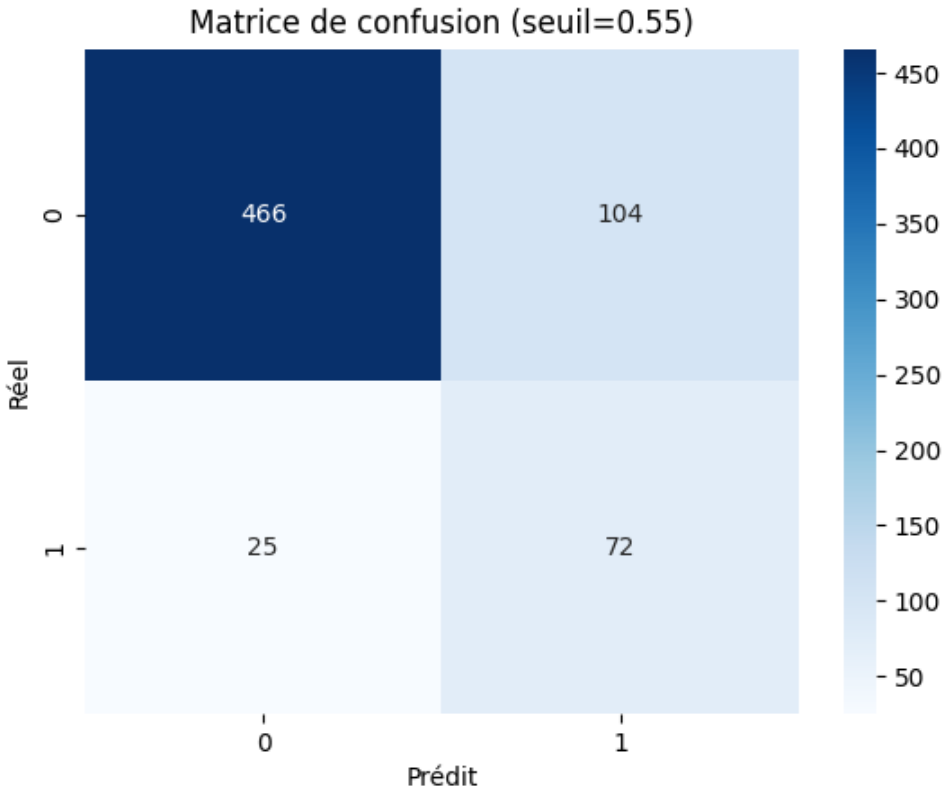
Matrice de confusion (seuil=0.55)
466
104
25
72
0
1
Réel
0
1
Prédit
450
400
350
300
250
200
150
100
50

Figure_6 :

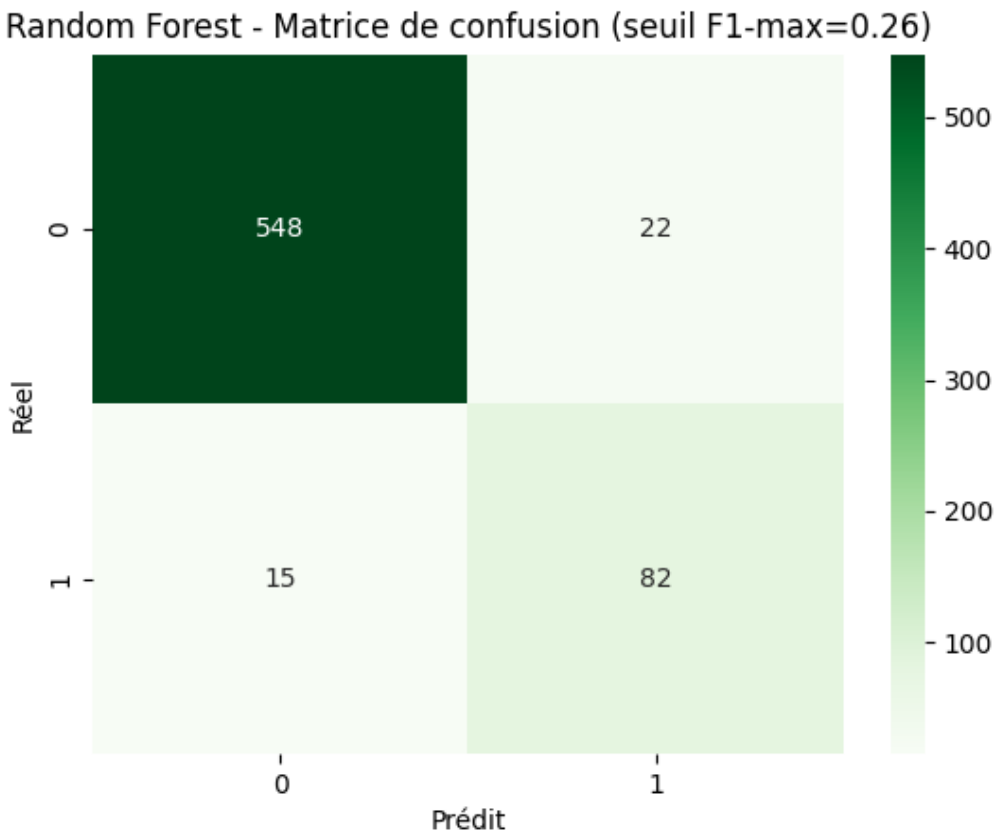

Figure_7 :

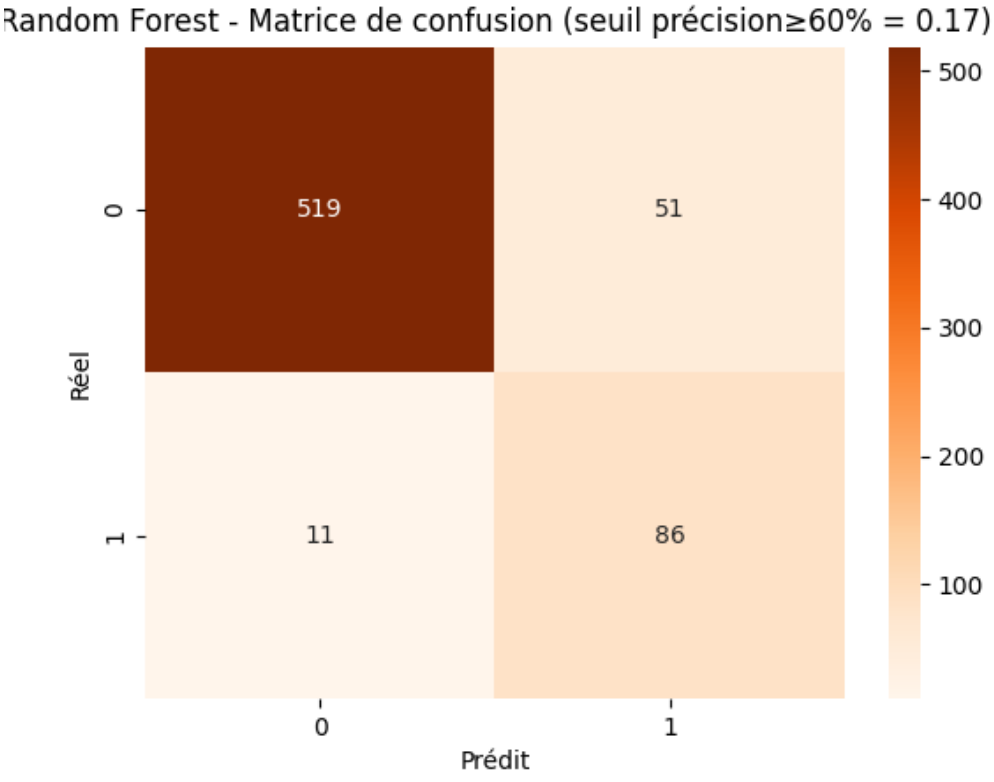

Figure_8 :

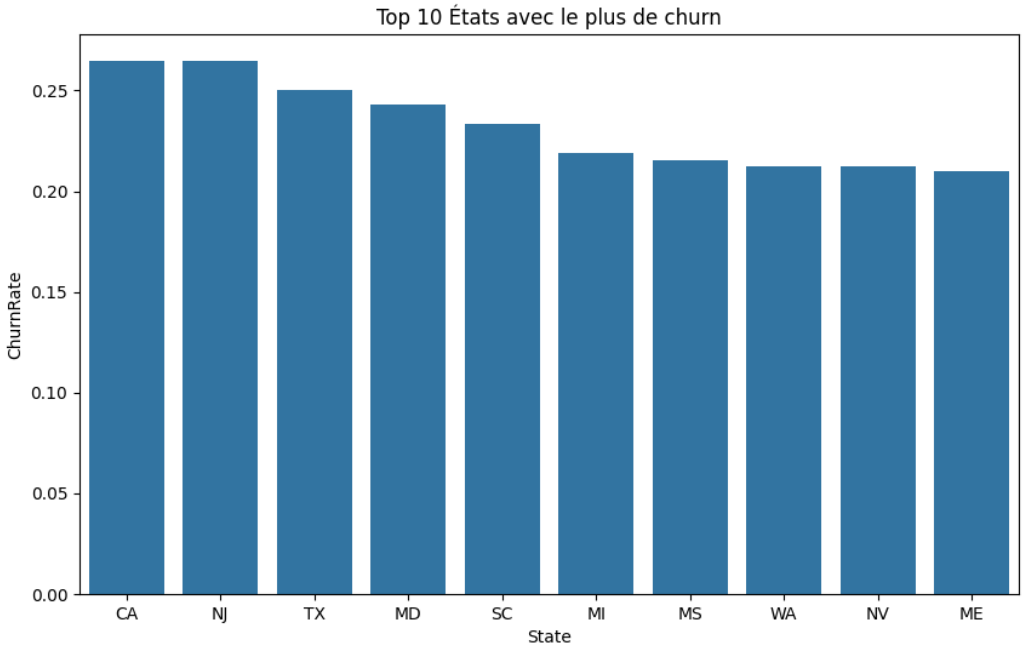

Figure_9 :

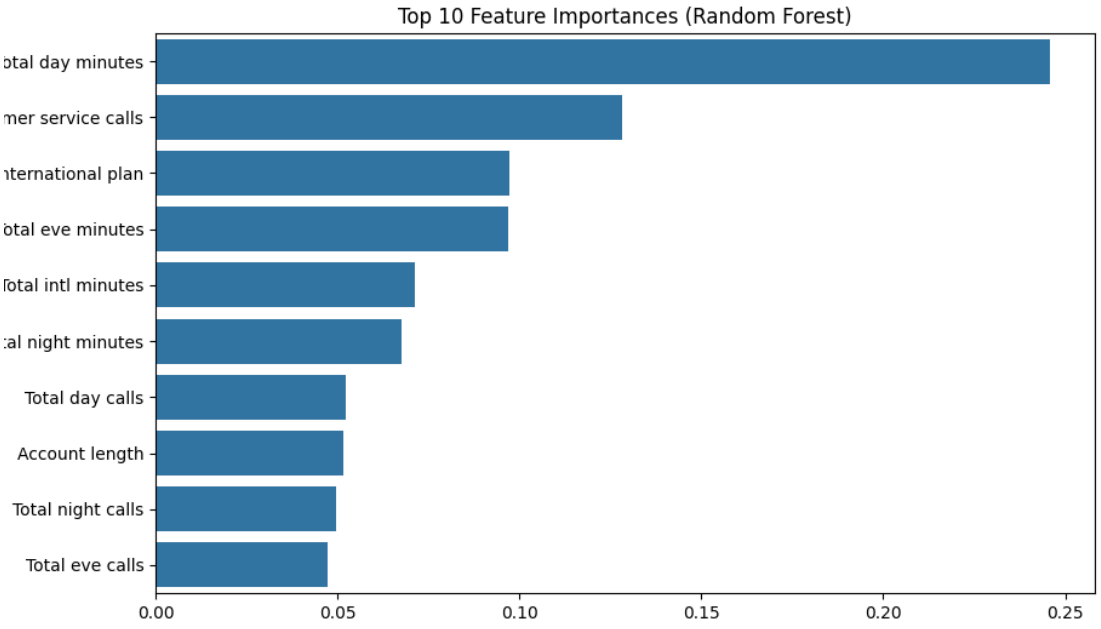

Figure_10 :