RNCP35288 CDSD

Conversion rate challenge

Supervised Machine Learning

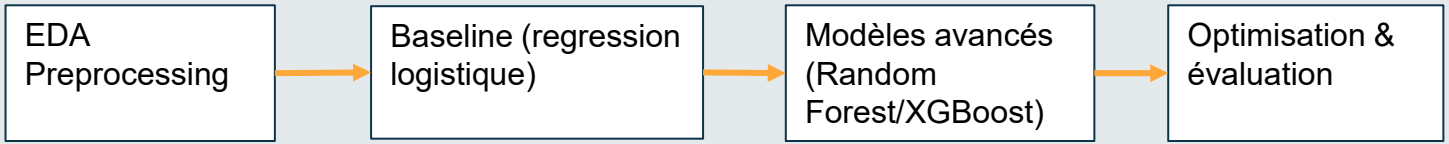
Rédacteur : Jean-Yves Vuillequez

Contexte

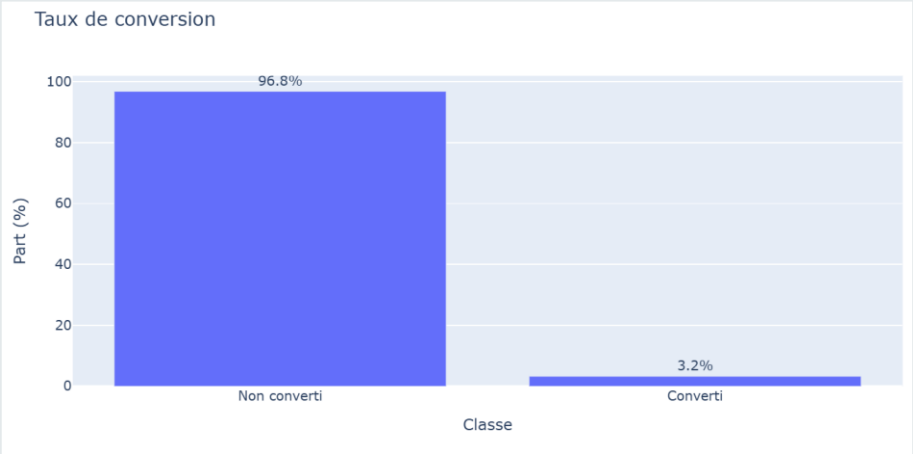
Objectif : prédire le comportement de conversion des visiteurs à partir de leurs données de navigation.

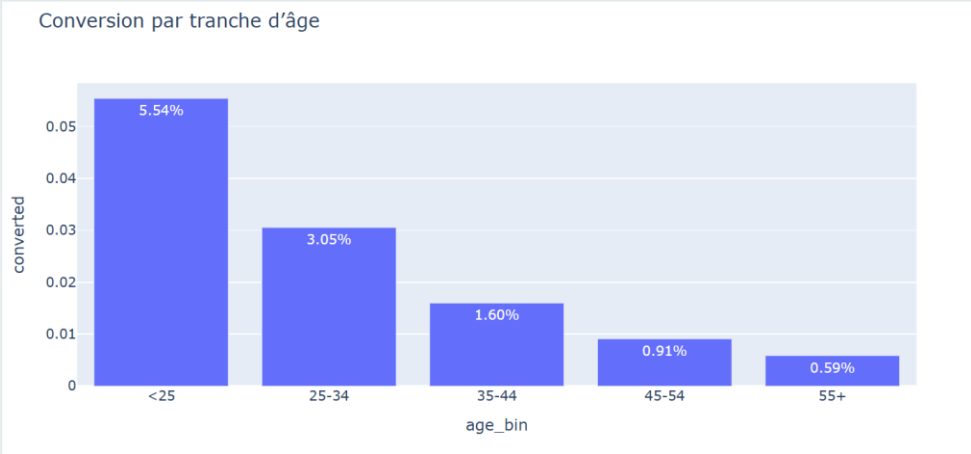
Données à disposition : Caractéristiques utilisateur (âge, nouveau client / existant), comportement de navigation (nombre de pages visitées), acquisition (source de trafic), variable cible

Approche :



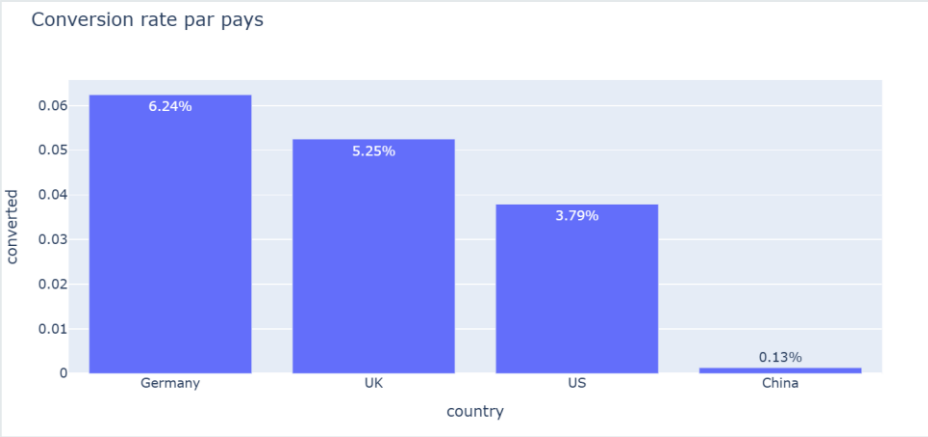
EDA 1/2

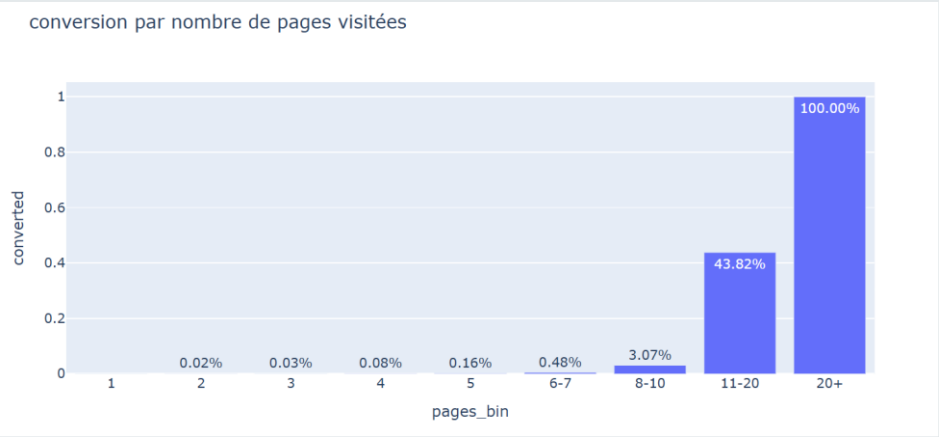




- **Le taux de conversion est très déséquilibré :** environ 3% de conversions
- **Un effet âge très marqué :** les plus jeunes convertissent significativement plus que les 45 et plus

EDA 2/2





- **Un effet pays très marqué : Les** comportements de conversion varient selon le marché avec une sous représentation des conversions en Chine
- **Un effet engagement** : le taux de conversion augmente fortement avec le nombre de pages visitées

Métriques et protocole

Critères d'évaluation du modèle de régression

Métriques choisies :

- **F1-score - Mesure l'équilibre entre faux positifs et faux négatifs :** Indicateur synthétique adapté aux classes déséquilibrées
- **Précision – « Parmi les utilisateurs prédits convertis, combien le sont réellement ? » :** Mesure la fiabilité des prédictions positives et permet de limiter le « gaspillage marketing » (
- **Recall (sensibilité) - « Parmi les vrais convertis, combien sont détectés ? » :** Mesure la capacité du modèle à identifier les opportunités de conversion sans en manquer
- **ROC AUC - « Le modèle sait-il classer correctement les utilisateurs ? »** : Évalue la capacité globale du modèle à distinguer convertis et non-convertis

Résultats

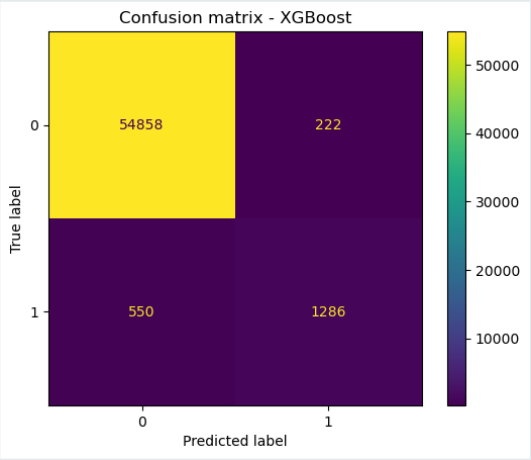
Baseline vs Random Forest / XGBoost

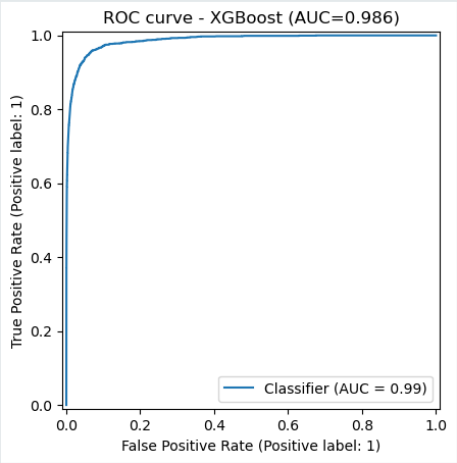
Pipeline : Preprocessing → Feature engineering → Modélisation → Optimisation hyperparamètres

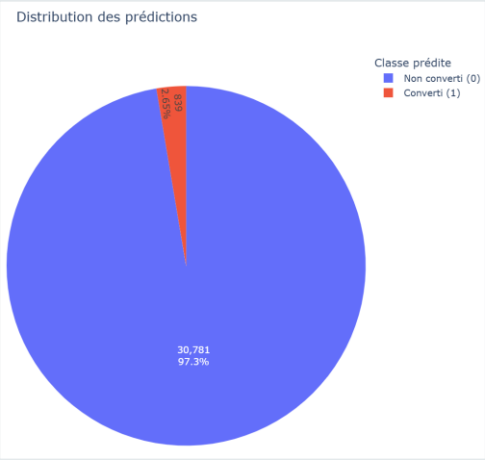
Modèle	F1 score	precision	recall	roc_auc	Diagnostic
LogReg	0.767	**0.865**	0.690	**0.987**	Baseline : bonne précision, peu de faux positifs
Random Forest	0.763	0.853	0.691	0.986	Modèle plus complexe sans gain significatif avec la baseline
XGBoost	**0.769**	0.853	**0.700**	0.986	Meilleur compromis global : F1 et recall légèrement supérieurs.

Résultats

Analyse XGBoost







- Le modèle détecte efficacement les conversions réelles (TP = 1 286) avec quelques faux positifs (FP = 222) mais avec des faux négatifs (FN = 550), donc des utilisateurs **convertis**, mais prédits comme non convertis ➔ **Coût métier potentiellement fort** : opportunités manquées
- La distribution des prédictions reste majoritairement « non converti **» en phase avec le faible taux de conversion observé dans les données d'entraînement**

Jedha

Merci pour votre attention
Des questions ?