PROJET PGE3 CODING FOR AI AND DATA SCIENCE

A rendre le 01 décembre (groupe de 2 maximum)

Ce projet se concentre non seulement sur les aspects techniques, mais aussi sur l'application métier et le choix stratégique du meilleur modèle de machine learning selon des métriques pertinentes.

Contexte du Projet : Optimisation d'une Campagne Marketing Bancaire

Dans un contexte de plus en plus concurrentiel, les institutions financières cherchent à maximiser l'efficacité de leurs campagnes de marketing tout en optimisant l'expérience client. La campagne de marketing direct de notre projet cible la promotion de dépôts à terme auprès des clients existants. Pour réduire les coûts et améliorer le taux de conversion, l'institution souhaite identifier les clients les plus susceptibles de souscrire à cette offre.

Objectif du Projet :

Le but est de prédire si un client acceptera ("yes") ou refusera ("no") l'offre de dépôt à terme, en s'appuyant sur les données issues des interactions précédentes et les caractéristiques socioéconomiques. En tant que data scientists, vous devrez explorer, préparer et modéliser les données afin de fournir un outil de prédiction précis et interprétable. Automatisez certaines parties avec vos propres fonctions (orienté objet si possible).

Étapes Détaillées du Projet :

1. Exploration des Données et Contexte Métier

Description des données :

Les données comprennent les informations démographiques des clients, leurs interactions passées avec la banque, et les résultats de campagnes marketing antérieures. Vous les trouverez sur le site https://archive.ics.uci.edu/dataset/222/bank+marketing

Files

bank.zip	565.5 KB
bank-additional.zip	

Problématique métier :

La souscription à un dépôt à terme est une décision importante pour un client. La banque souhaite minimiser les coûts liés à des campagnes marketing inefficaces et maximiser le retour sur investissement (ROI).

Métrique métier clé :

En plus des métriques standards comme l'accuracy ou le F1score, la précision et le recall sont essentielles. En effet, un faux positif (prédire que le client acceptera alors qu'il refuse) entraîne des coûts de marketing inutiles, tandis qu'un faux négatif (manquer un client prêt à souscrire) représente une opportunité perdue.

2. Préparation des Données

Nettoyage et transformation :

Traitement des valeurs manquantes, normalisation des données numériques.

Encodage des variables catégorielles (par exemple : OneHot Encoding).

Analyse exploratoire :

Visualisation des distributions de variables clés (âge, profession, éducation, etc.) et exploration de leur influence sur la souscription (variable cible `y`).

Stratégies de mise à l'échelle :

Normalisation des valeurs numériques pour une meilleure convergence des modèles.

3. Création et Entraînement de Modèles

Modèles proposés :

Un simple Knn

Régression Logistique : Modèle de base interprétable pour les prédictions binaires.

Arbres de Décision : Modèle intuitif, utile pour l'interprétabilité.

Random Forest : Ensemble d'arbres de décision offrant souvent de meilleures performances globales.

Gradient Boosting (XGBoost) : Modèle puissant pour les prédictions, avec une capacité élevée à gérer les données déséquilibrées.

Méthodes avancées (optionnel)

4. Évaluation des Modèles :

Métriques de performance clés :

Accuracy : Pourcentage de prédictions correctes. Utile mais peut être trompeur si les classes sont déséquilibrées.

Précision (Precision) : Proportion de prédictions positives correctes par rapport à l'ensemble des prédictions positives. Réduit les faux positifs (évite les coûts inutiles).

Rappel (Recall ou Sensibilité) : Proportion de vraies positives détectées parmi toutes les vraies positives possibles. Réduit les faux négatifs (maximise les opportunités).

F1score : Moyenne harmonique de la précision et du rappel. Idéal pour évaluer l'équilibre entre ces deux métriques.

Courbe ROC AUC : Indicateur de la performance globale du modèle.

Choix du modèle le plus adapté :

Si la banque souhaite maximiser les souscriptions positives (minimiser les faux négatifs), le rappel sera privilégié.

Si la banque cherche à réduire les coûts liés à des campagnes inutiles (minimiser les faux positifs), la précision sera le critère principal.

Un F1score élevé peut représenter un bon compromis lorsque les enjeux de précision et de rappel sont équilibrés.

5. Optimisation du Modèle :

Tuning des hyperparamètres à l'aide de la recherche en grille ou en random search pour obtenir la meilleure performance.

Analyse des variables importantes pour comprendre les facteurs influençant la souscription.

6. Interprétation et Application Métier

Présentation des résultats au travers de visualisations claires (ex : importance des variables, courbe ROC, matrice de confusion).

Cas métier : Par exemple, identifier des segments de clients à cibler en priorité lors des prochaines campagnes.

7. Mise en Production

Intégration du modèle avec un tableau de bord (par exemple, via Streamlit) pour permettre aux utilisateurs métiers de tester différents scénarios et visualiser les prévisions.

Outils Utilisés :

Langage de programmation : Python

Bibliothèques : Pandas, NumPy, Scikitlearn, XGBoost, Matplotlib, Seaborn etc

Visualisation et tableau de bord : Streamlit (ou flask ou autres)

Résumé du Projet :

Vous travailleraient non seulement sur la préparation, l'entraînement, et l'optimisation d'un modèle de machine learning, mais aussi considérer le métrique métier les plus pertinentes pour prendre des décisions éclairées, tout en gardant à l'esprit les implications économiques de leurs prédictions. Ce projet permet de mettre en pratique l'optimisation des campagnes marketing, une compétence clé pour tout data scientist dans le secteur financier ou du marketing. Vous pouvez aussi rajouter du webscraping (optionnel) pour enrichir vos donnees.