Compte Rendu d'Analyse Machine Learning sur le Dataset Bank Marketing

NADA EL IMANI

04/12/2025

Introduction

Ce travail porte sur l'analyse et la modélisation prédictive du dataset Bank Marketing issu du UCI Machine Learning Repository. Le contexte consiste en des campagnes de marketing direct d'une banque portugaise, visant à prédire si un client souscrira un produit bancaire spécifique (dépôt à terme). L'objectif principal est de développer un modèle capable de prédire efficacement cette souscription à partir des données clients et de la campagne téléphonique.

Méthodologie

Le traitement des données a inclus une exploration initiale, le nettoyage (suppression des doublons), la gestion des valeurs manquantes en combinant imputation KNN pour les variables numériques et traitement des modalités inconnues comme valeurs manquantes pour les catégorielles. L'encodage des variables catégorielles a été effectué en distinguant variables ordinales (avec un encodage ordinal) et nominales (encodage One-Hot ou label encoding pour binaires).

La normalisation a été appliquée via une standardisation StandardScaler sur les variables numériques, en excluant la variable *duration* qui représente la durée des appels et qui aurait pu entraîner un biais (data leakage).

Pour la modélisation, trois algorithmes ont été choisis pour leur complémentarité :

— Régression Logistique, pour sa simplicité et interprétabilité.
— Random Forest, pour sa robustesse face à la non-linéarité et la capacité à gérer l'importance des variables.

— Gradient Boosting (XGBoost), méthode avancée performante avec régularisation et possibilité d'optimisation poussée des hyperparamètres.

L'important déséquilibre de classes (majorité de clients ne souscrivant pas) a été corrigé par une technique d'oversampling SMOTE. La validation a été assurée par un train-test split (80/20) couplé avec une validation croisée à 5 folds. L'optimisation des hyperparamètres a été réalisée par RandomizedSearchCV, privilégiant un bon compromis entre exhaustivité et temps de calcul.

Résultats & Discussion

Les performances des modèles ont été évaluées grâce à plusieurs métriques : **Accuracy**, **F1-Score**, **ROC-AUC**, ainsi que l'analyse de la matrice de confusion.

— **Régression Logistique :** ROC-AUC environ 0.82, F1-Score aux alentours de 0.59. Le modèle est rapide mais limité par l'hypothèse de linéarité.

— **Random Forest :** Amélioration notable avec ROC-AUC autour de 0.85, F1-Score proche de 0.64. Meilleure gestion de variables complexes.

— **XGBoost :** Meilleure performance avec ROC-AUC après optimisation à 0.87 et F1-Score à 0.66. L'optimisation des hyperparamètres améliore significativement les résultats.

L'analyse de la matrice de confusion révèle un bon équilibre entre taux de vrais positifs et faibles taux de faux positifs, essentiel dans un contexte business pour limiter les coûts liés à un ciblage erroné.

Cependant, le dataset présente un déséquilibre marqué, bien atténué par SMOTE, mais qui reste une limite. Certaines variables comme *duration* ont été exclues pour éviter la fuite d'information.

Conclusion

Le modèle XGBoost optimisé constitue la meilleure approche pour cette problématique, offrant un bon compromis entre précision et robustesse. Néanmoins, des limites subsistent :

— Déséquilibre initial des classes et biais potentiels malgré SMOTE.

— Domaine d'application limité à la banque portugaise et aux variables disponibles.

— Complexité et temps d'entraînement plus élevés des modèles avancés.

Des pistes d'amélioration incluent l'expérimentation de modèles de deep learning, l'intégration de nouvelles sources de données externes (comportement client digital), ainsi qu'un suivi continu des performances en production avec des outils d'interprétabilité (ex : SHAP values).