Proposition de Projet Capstone
Prédiction du Churn des Clients Bancaires

Ce projet est réalisé en duo avec mon frère **Micka Louis.**

Compréhension du Business

Les banques investissent massivement pour acquérir de nouveaux clients, mais fidéliser les clients existants est bien plus rentable. L'objectif de ce projet est de construire un modèle capable de prédire si un client quittera la banque en se basant sur des informations démographiques et comportementales.
Nous avons choisi ce sujet car la prédiction du churn combine la réflexion économique avec la prise de décision basée sur les données. Dans le contexte bancaire réel, identifier les clients susceptibles de partir permet aux managers d'agir tôt et de mettre en place des stratégies pour les retenir. Le projet s'applique au secteur bancaire et financier et cible les analystes de données, les équipes marketing et les responsables de la relation client.
Si le modèle est performant, il pourrait aider les banques à réduire le taux d'attrition des clients, à renforcer les programmes de fidélité et à améliorer la rentabilité.

Compréhension des Données

Le jeu de données utilisé pour ce projet provient de Kaggle et contient des informations sur 10 000 clients bancaires. Il inclut des variables démographiques telles que la géographie, le sexe et l'âge, ainsi que des variables financières comme le score de crédit, le solde du compte, l'ancienneté, le nombre de produits détenus et le salaire estimé. La variable cible « Exited » indique si un client a quitté la banque.
Ce jeu de données est public et couramment utilisé pour étudier le comportement de churn des clients. Il constitue une base solide pour l'apprentissage supervisé et la modélisation de classification.

Préparation des Données

Les données sont stockées dans un fichier CSV. Nous commencerons par nettoyer le jeu de données en supprimant les colonnes non pertinentes telles que RowNumber, CustomerId et Surname. Les variables catégorielles comme Geography et Gender seront encodées sous forme numérique, et les variables numériques seront normalisées pour assurer une échelle cohérente. Les valeurs manquantes seront traitées avec soin afin de préserver l'intégrité des données. L'ingénierie des caractéristiques comprendra la création de nouveaux indicateurs, tels que le ratio entre le solde du compte et le salaire, ou un indicateur indiquant si le solde d'un client dépasse un certain seuil. Une fois les données prêtes, elles seront divisées en ensembles d'entraînement et de test pour la construction du modèle.

Modélisation

Le premier modèle sera une régression logistique utilisée comme référence. Des modèles plus avancés tels que Random Forest, XGBoost et LightGBM seront ensuite testés afin d'améliorer la performance prédictive. L'objectif principal est de trouver le modèle le plus précis et fiable capable d'identifier les clients à risque de churn tout en restant interprétable.
Une analyse de l'importance des caractéristiques à l'aide des valeurs SHAP sera réalisée pour comprendre les principaux facteurs influençant les décisions des clients.

Évaluation

Les métriques d'évaluation incluront l'Accuracy, la Précision, le Recall, le F1-Score et le ROC-AUC. Comme l'objectif principal est d'identifier le plus grand nombre possible de clients à risque de départ, le Recall sera priorisé. Le produit minimum viable consistera en un modèle de classification fonctionnel avec un pipeline de prétraitement complet et une visualisation claire des performances.
Les objectifs supplémentaires incluront l'optimisation du modèle, l'affinement de la sélection des caractéristiques et la visualisation via un tableau de bord.

Déploiement

Le modèle final sera déployé via une application web Streamlit. Les utilisateurs pourront saisir le profil d'un client, et l'application affichera la probabilité que ce client quitte la banque. Cet outil peut aider les managers et analystes à prendre des décisions basées sur les données afin d'améliorer les stratégies de fidélisation.

Outils et Méthodologies

Le projet sera implémenté en Python en utilisant des bibliothèques telles que Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, XGBoost et SHAP. Toutes les analyses seront effectuées localement sur nos machines. Le rapport final et l'application web résumeront les principales conclusions et prédictions dans un format clair et orienté business.