***Rapport d’analyse :***

**Prise de décision avec un modèle d’IA pour une campagne marketing bancaire**

Dans un contexte bancaire marqué par une intensification de la concurrence et une exigence croissante en matière de rentabilité, les campagnes marketing doivent être de plus en plus ciblées, efficaces et mesurables. La réussite d'une offre telle qu’un dépôt à terme dépend fortement de la capacité de la banque à identifier les bons profils de clients, à anticiper leur comportement et à personnaliser l’approche.

Ce projet s’inscrit dans cette démarche stratégique : exploiter l’analyse de données et machine Learning pour améliorer les performances marketings d’une banque.

Les objectifs de ce projet sont :

* Explorer et analyser les données pour extraire des insights stratégiques.
* Appliquer des méthodes statistiques avancées pour valider les résultats.
* Fournir des recommandations stratégiques basées sur les résultats obtenus.
* Prédire si un client acceptera (« yes ») ou refusera (« no ») une offre de dépôt à terme

1. **Description des données et exploration initiale**

Le dataset utilisé provient d’une campagne marketing bancaire. Elle contient des informations socio-démographiques, des données de contact ainsi que les résultats de campagnes précédentes.

La variable cible y indique si le client a souscrit (yes) ou non (no) à une offre de dépôt à terme.

Les variables sont les suivantes :

* **age** : âge du client (numérique).
* **job** : type d'emploi (catégorique : « admin.», « unemployed », « management », etc.).
* **marital** : statut matrimonial (catégorique : « married », « single », « divorced »).
* **education** : niveau d'éducation (catégorique : « primary », « secondary », « tertiary »).
* **default** : crédit en défaut ? (binaire : « yes », « no »).
* **balance** : solde moyen annuel en euros (numérique).
* **housing** : prêt immobilier ? (binaire : « yes », « no »).
* **loan** : prêt personnel ? (binaire : « yes », « no »).
* **contact** : type de communication lors du dernier contact (catégorique : « telephone », « cellular »).
* **day** : jour du dernier contact (numérique).
* **month** : mois du dernier contact (catégorique : « jan », « feb », etc.).
* **duration** : durée du dernier contact en secondes (numérique).
* **campaign** : nombre de contacts réalisés durant la campagne (numérique).
* **pdays** : nombre de jours depuis le dernier contact dans une campagne précédente (numérique, -1 si jamais contacté).
* **previous** : nombre de contacts avant cette campagne (numérique).
* **poutcome** : résultat de la campagne précédente (catégorique : « success », « failure », « unknown »).
* **Variable cible : y** : le client a-t-il souscrit à un dépôt à terme ? (binaire : « yes », « no »).

L’exploration initiale a commencé par une vérification des dimensions du fichier et un aperçu des premières lignes. Une table de statistiques descriptives a ensuite permis d’identifier des écarts importants dans certaines variables comme le solde ou la durée des appels. Une attention particulière a été portée à la distribution des classes de la variable cible, qui est déséquilibrée en faveur du 'non'.

1. **Processus de préparation et nettoyage de données**

Avant toute modélisation, il était indispensable de procéder à une préparation rigoureuse des données afin de garantir la fiabilité des résultats. La première étape a été de vérifier s’il existait des valeurs manquantes. Le dataset ne contenait aucune valeur nulle, ce qui a permis d’éviter l’étape d’imputation. Ensuite, les doublons ont été supprimés, bien que très peu soient présents.

L’analyse des valeurs aberrantes a été faite sur les variables numériques à l’aide de la méthode de Tukey. Cela a permis d’identifier certains écarts importants (notamment sur les soldes très élevés ou le nombre de contacts), tout en décidant de conserver ces valeurs dans certains cas puisqu’elles représentaient une réalité client exploitable (ex : client fortuné ou très sollicité).

Les variables catégorielles ont ensuite été encodées avec ***get\_dummies***, afin de rendre le dataset compatible avec les modèles. La variable duration, bien que très corrélée avec la cible, a été exclue pour éviter tout biais, car elle ne peut être connue qu’après le contact avec le client.

Enfin, les données ont été standardisées à l’aide de ***StandardScaler***, notamment pour KNN et le clustering, deux méthodes très sensibles à l’échelle des variables.

1. **Comparaison des performances des modèles**

Dans un premier temps, le modèle principal utilisé a été le K-Nearest Neighbors (KNN). Simple à comprendre et à mettre en œuvre, il a été un bon point de départ. Le choix du paramètre k a été fait par validation croisée, et la meilleure performance a été obtenue avec k = 13.

Cependant, même si l’accuracy globale était plutôt bonne (environ 89 %), les performances sur la classe yes étaient décevantes : le modèle détectait peu de clients réellement intéressés (rappel faible). Cela montre la limite du KNN dans un contexte de classes déséquilibrées.

En complément, un Random Forest a été testé. Ce modèle a permis de mieux capturer les clients potentiellement intéressés tout en étant moins sensible aux déséquilibres. Il fournit aussi une importance des variables, ce qui est utile pour l’interprétation métier. Enfin, XGBoost et Naive Bayes ont été évoqués comme modèles alternatifs. XGBoost serait adapté à une future mise en production (robuste, optimisé), tandis que Naive Bayes est surtout utile pour des cas simples.

1. **Résultats des analyses statistiques avancées**
   1. **Clustering**

Pour mieux comprendre les profils clients, un clustering KMeans a été réalisé. La méthode du coude et le score de silhouette ont permis de fixer k = 3. Les clients ont été segmentés comme suit :

* Cluster 1 : clients réactifs, ayant été contactés auparavant, et présentant un taux de souscription élevé.
* Cluster 0 : clients moyens, avec des caractéristiques modérées (solde, âge, etc.)
* Cluster 2 : jeunes, peu sollicités, peu enclins à souscrire.

Les clusters ont été visualisés via une PCA à deux dimensions, et la silhouette par cluster a confirmé que le cluster 1 était le plus cohérent et le plus exploitable

* 1. **Visualisations des taux de souscriptions**

Des graphiques ont été générés pour afficher les taux de souscription par métier, niveau d’éducation, mois, tranche d’âge, etc. Ces visuels ont permis d’identifier des profils plus réceptifs aux campagnes marketing (ex : les étudiants, les retraités, les cadres ou les personnes contactées en mars et octobre).

1. **Recommandations stratégiques basées sur les résultats obtenus**

Sur la base de l’analyse, plusieurs recommandations peuvent être faites :

* Cibler prioritairement les clients du cluster 1, ayant déjà été sollicités avec succès, et présentant un solde positif.
* Éviter les sur-sollicitations : les clients trop contactés (campaign > 10) répondent rarement favorablement.

En cas de budget marketing limité, concentrer les efforts sur les clients avec :

* Un bon historique (previous > 0)
* un solde supérieur à la moyenne
* des contacts passés positifs (poutcome = success)

Du point de vue métier, il serait préférable de baisser légèrement le seuil de prédiction pour augmenter le rappel, même si cela implique quelques faux positifs. Cela permettrait de réduire les opportunités manquées (faux négatifs).

Enfin, l’intégration future de variables économiques (ex : taux d’intérêt, inflation) permettrait de simuler plusieurs scénarios et d’adapter les campagnes aux contextes économiques.