Rapport Détaillé sur le Projet : Analyse Exploratoire de Données pour une Campagne Marketing Bancaire

1. Introduction au Projet Global

Titre et Contexte

Marketing Bancaire (avec contexte social/économique)

Le projet s'inscrit dans un contexte bancaire concurrentiel, où les campagnes marketing doivent être optimisées pour maximiser la rentabilité et l'expérience client. Qu'est-ce que c'est ? Une banque portugaise a mené des campagnes de marketing par téléphone (télémarketing) dans l'objectif de convaincre les clients de souscrire à un dépôt à terme (produit bancaire où on place de l'argent pour une durée fixe avec intérêts) à la période de : Mai 2008 à Novembre 2010 sur une population d'environ 10 millions d'habitants (population du Portugal à l'époque). Ma mission c'est donc de Prédire si un client va dire "OUI" ou "NON" à l'offre de dépôt à terme.

L'objectif principal est d'analyser des données d'une campagne de marketing direct (appels téléphoniques) d'une banque portugaise pour prédire si un client souscrira à un dépôt à terme (variable cible : y = "yes" ou "no").

Objectifs spécifiques :

- Explorer et analyser les données pour extraire des insights stratégiques.
- Appliquer des méthodes statistiques avancées (ex. : tests ANOVA, Khi 2, MANOVA) pour valider les résultats.
- Fournir des recommandations stratégiques basées sur les insights.
- Prédire la souscription (y) avec un modèle simple (focus sur l'analyse, pas sur l'optimisation ML).
- Question métier : "Comment identifier les clients qui ont le plus de chances de souscrire à un dépôt à terme, afin d'optimiser les campagnes de télémarketing ?" soit :
    - *1. Analyser quels types de clients souscrivent le plus.*
    - *2. Identifier les meilleurs moments pour appeler (mois, jour).*
    - *3. Déterminer l'influence du contexte économique.*
    - *4. Construire un modèle prédictif pour cibler les bons clients.*
    - *5. Faire des recommandations stratégiques à la banque.*

Le projet suit la méthodologie **CRISP-DM** (Cross-Industry Standard Process for Data Mining), avec 7 étapes :

1. Compréhension du contexte et des données.
2. Préparation des données (nettoyage, outliers, etc.).

3. Analyses statistiques (univariée, bivariée/multivariée).
4. Modélisation (modèle simple comme KNN).
5. Évaluation des modèles (métriques basiques : accuracy, précision).
6. Présentation des résultats et recommandations.
7. Mise en production (bonus : dashboard Streamlit).

Données Fournies et Taille du Dataset

- **Dataset principal** : bank-additional-full.csv , avec ***41 188*** *clients (lignes) et* ***21*** *variables (colonnes) : 20 variables explicatives et 1 variable cible (y).*
- **Variables** : 16 features d'entrée + 1 cible (y).
    - Numériques : age, balance, day, duration, campaign, pdays, previous, etc.
    - Catégorielles : job, marital, education, default, housing, loan, contact, month, poutcome, etc.
    - Variables socio-économiques : emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed.
- Le PDF insiste sur l'utilisation de la base la plus complète (bank-additional.zip).

Description détaillée des variables :

A) **INFORMATIONS DEMOGRAPHIQUES DU CLIENT**

- age : Numérique - Âge du client (ex: 35 ans).
- job : Catégorielle - Type d'emploi : admin, ouvrier, entrepreneur, femme au foyer, management, retraité, indépendant, services, étudiant, technicien, chômeur, inconnu.
- marital : Catégorielle - Statut marital : marié, célibataire, divorcé (inclut veufs), inconnu.
- education : Catégorielle - Niveau d'éducation : basic.4y (primaire 4 ans), basic.6y, basic.9y, lycée, illettré, formation professionnelle, diplôme universitaire, inconnu.

B) **INFORMATIONS FINANCIERES**

- default : Catégorielle - A un crédit en défaut de paiement ? oui/non/inconnu.
- housing : Catégorielle - A un prêt immobilier ? oui/non/inconnu.
- loan : Catégorielle - A un prêt personnel ? oui/non/inconnu.

C) **INFORMATIONS SUR LE DERNIER CONTACT**

- contact : Catégorielle - Type de communication : cellulaire ou téléphone fixe.
- month : Catégorielle - Mois du dernier contact.
- day_of_week : Catégorielle - Jour de la semaine.
- duration : Numérique - Durée du dernier appel en secondes . *Et Cette variable ne peut pas être utilisée pour prédire car on ne connaît la durée qu'APRÈS l'appel .*

D) **INFORMATIONS SUR LA CAMPAGNE**

- campaign : Numérique - Nombre de contacts effectués durant CETTE campagne (inclut le dernier contact).
- pdays : Numérique - Nombre de jours depuis le dernier contact d'une campagne PRÉCÉDENTE (999 = jamais contacté avant).
- previous : Numérique - Nombre de contacts effectués AVANT cette campagne.
- poutcome : Catégorielle - Résultat de la campagne précédente : succès, échec, inexistant (jamais contacté).

E) **INDICATEURS SOCIO-ÉCONOMIQUES** *(variables très importantes car elles reflètent le contexte économique du Portugal)*

- emp.var.rate : Numérique - Taux de variation de l'emploi (indicateur trimestriel) - Si positif = création d'emplois, si négatif = perte d'emplois.
- cons.price.idx : Numérique - Indice des prix à la consommation (mensuel) - Mesure l'inflation.
- cons.conf.idx : Numérique - Indice de confiance des consommateurs (mensuel) - Si élevé = les gens sont optimistes.
- euribor3m : Numérique - Taux Euribor à 3 mois (quotidien) - Taux d'intérêt de référence en Europe.
- nr.employed : Numérique - Nombre d'employés (trimestriel) - Indicateur de l'emploi national.

F) **VARIABLE CIBLE**

- y : Binaire 6 Le client a-t-il souscrit au dépôt à terme ? "yes" ou "no".

Points Importants : A) **Valeurs Manquantes** : On n'a pas de valeurs NULL dans le dataset, seulement certaines variables catégorielles contiennent "unknown" (inconnu) et ces "unknown" représentent des valeurs manquantes déguisées. Je vais donc décider quoi en faire : supprimer, imputer, ou garder comme catégorie. B) **Variable "duration"** : La durée de l'appel prédit parfaitement le résultat : Si duration = 0 alors le client a dit "non" immédiatement ; Si duration est élevé alors le client était intéressé. Mais je ne connais pas la durée AVANT de faire l'appel. Cette variable je vais donc l'exclure. C) **Déséquilibre des Classes** : Le dataset est très déséquilibré : Environ 11% de "yes" (clients qui souscrivent) ; Environ 89% de "no" (clients qui refusent). Ce déséquilibre peut compliquer la modélisation.

Concepts Clés et Questions Professionnelles

- **Prise de décision basée sur les données** : Justification objective, réduction des biais, prévision des résultats, adaptabilité.
- **Questions orientées décision** : Segments à cibler, priorisation budgétaire, équilibre coût/opportunités, KPI (ex. : taux de conversion), intégration de scénarios économiques.

Livrables Attendus

1. Rapport d'analyse (description données, nettoyage, résultats stats, recommandations).

2. Code Python documenté (notebooks Jupyter).
3. Tableau de bord interactif (OPTIONNEL).

- **Outils** : Python, Pandas, NumPy, Scikit-Learn, Matplotlib, Seaborn, Streamlit (optionnel).

2. Phase 1 : Setup et Compréhension des Données

Cette phase correspond à l'étape 1 de CRISP-DM : Compréhension du contexte et des données.

Objectifs

- Vérifier l'installation des bibliothèques.
- Charger le dataset.
- Explorer superficiellement les données.
- Comprendre le contexte.

Étapes Réalisées

1. **Import et Configuration des Bibliothèques** :
    - Bibliothèques importées : pandas (v2.3.3), numpy (v2.4.2), matplotlib.pyplot, seaborn, scipy.stats.
    - Configurations : Affichage max colonnes, style de plot (seaborn-v0_8-darkgrid), suppression des warnings.
    - Justification logique : Assurer un environnement prêt pour l'analyse et la visualisation.
2. **Chargement des Données** :
    - Dataset : bank-additional-full.csv .
    - Taille : 41 188 lignes × 21 colonnes.
    - Aperçu initial : Affichage des 5 premières lignes (head), montrant des variables comme age, job , marital , etc., jusqu'à y.
3. **Exploration Superficielle** :
    - Focus sur la variable age (comme exemple d'analyse univariée initiale).
        - **Visualisation** : Figure avec 2 sous-plots (histogramme et boxplot).
            - Histogramme : Distribution de l'âge, avec lignes pour moyenne (**40.0**) et médiane (**38.0**).
            - Boxplot : Montre les quartiles et outliers potentiels.
            - Sauvegarde : results/figures/02_distribution_age.png .
        - **Statistiques descriptives** :
            - Moyenne : 40.0 ans.
            - Médiane : 38.0 ans.
            - Min : **17** ans.
            - Max : **98** ans.
            - Écart-type : 10.4 ans.

- Justification : Cela illustre une analyse univariée basique. L'âge est une variable clé démographique, et cette exploration montre une distribution asymétrique vers les âges plus élevés (*queue droite*), ce qui est logique pour une clientèle bancaire. ***L'âge est corrélé avec capacité d'épargne, et fait partie des caractéristiques démographiques importantes pour identifier les segments de clients.***

3. Phase 2 : Préparation et Nettoyage des Données

Cette phase correspond à l'étape 2 de CRISP-DM : Préparation des données.

Objectifs

- Traiter les valeurs "unknown" dans les catégorielles.
- Gérer les outliers potentiels dans les numériques.
- Analyser les corrélations.

Étapes Réalisées

1. **Import et Configuration**
2. **Partie 1 : Analyse des Valeurs "Unknown"**
   - Identification : Dans 6 variables catégorielles (default, education, housing, loan, job, marital).
   - Quantification :
     - Tableau : Counts (ex. : default = 8597, **20.87%**).
     - Total unknowns : **12 718**.
     - Classification : Élevé (> 5% : default), Moyen (**1-5%** : education, housing, loan), Faible (< **1%** : job, marital).
   - Méthodologie d'analyse :

     - Étape 1 : Quantifier (pourcentage par variable).
     - Étape 2 : Analyser l'impact (taux souscription avec/sans unknown).
     - Étape 3 : Décider selon le pourcentage (< 5% supprimer lignes ; 5-20% imputer ou garder ; > 20% garder comme catégorie).
     - Options : Supprimer lignes (pas de biais mais perte données), Imputer mode (simple mais biais), Garder comme catégorie (aucune perte, "unknown" a du sens métier).

   - Impact sur y : Tableau récapitulatif avec taux "yes" avec/sans unknown, différence, décision.
     - Ex. : Default : **5.15% yes** avec unknown vs **12.88% sans** entraine une différence de -7.72% c'est donc à garder (impact significatif).

        - Décisions : Garder pour la plupart (unknown est une info en soi), supprimer lignes pour job/marital (faible %).
    - Décisions finales :

        - *GARDÉES (5) :* default (20.67%, impact -7.72, clients "unknown" souscrivent 2.5 x moins), education (3.92%, impact +3.38), marital (0.17%, impact +3.74), housing (2.41%, impact -0.47), loan (2.41%, impact -0.47)
        - *SUPPRIMÉES (1) :* job (impact -0.05, quasi nul*).*

    - Action : Suppression de **330 lignes (0.80%)**, équilibre de **y** préservé .
    - Justification : Approche conservatrice : les unknowns ne sont pas des erreurs, mais des infos (ex : refus de répondre).
3. **Partie 2 : Détection et Traitement des Outliers** :
    - Méthode : **IQR (Q1 - 1.5*IQR à Q3 + 1.5*IQR).**
    - Résultats : Tableau avec outliers par variable (ex : previous = 5588, 13.68% ; duration = 2922, 7.15%).
    - Décision : Garder TOUS (**13 288 outliers**) ; Valeurs réelles (ex : appels longs = engagements positifs).
    - Justification : Outliers ne veut pas dire erreurs ils portent de l'information.
4. **Partie 3 : Analyse des Corrélations** :
    - Méthode : Coefficient **Pearson sur numériques**.
    - Résultats : Matrice (45 paires), corr moyenne **0.221** ; 4 paires supérieurs à **0.7** (ex : emp.var.rate & euribor3m = 0.972).
    - Focus socio-éco : 6 paires fortes (supérieur à **0.5**), normales pour indicateurs macro.
    - Décision : Garder toutes.
    - Justification : Corrélations attendues, ajoutent nuance (ex. : contexte économique impacte souscriptions).

Sélection des Variables pour la Modélisation

A) **VARIABLES A INCLURE (et le pourquoi)**

1. *VARIABLES SOCIO-ECONOMIQUES* (les plus importantes). Elles reflètent le contexte économique global ; Quand l'économie est forte, les gens épargnent plus ; Ces variables changent dans le temps donc ils sont bon pour la prédiction.
2. *HISTORIQUE CLIENT* (aussi important) : Le comportement passé prédit le comportement futur ; Si on a un "success" avant il y'a une forte probabilité de re-souscription.
3. *VARIABLES TEMPORELLES* (Importantes) : Certains mois/jours sont plus propices.
4. *CARACTERISTIQUES DEMOGRAPHIQUES* (Importantes) : elles sont corrélées avec la capacité d'épargne, revenus, culture financière.
5. *SITUATION FINANCIERE* (Modérément importante) elle indique la capacité financière restante.
6. *INFORMATIONS CAMPAGNE* (faiblement important) : Trop de contacts peuvent agacer et le type contact impacte.

B) **VARIABLES A EXCLURE (et pourquoi)** : *DURATION* : On ne connaît la durée qu'après l'appel c'est une conséquence et pas une cause.

4. Phase 3 : Analyses Statistiques Approfondies

Cette phase correspond à l'étape 3 de la méthodologie CRISP-DM : **Analyses statistiques**. Elle se concentre exclusivement sur l'exploration des données nettoyées (issues du Notebook 2) pour extraire des insights métier, sans aucune modélisation prédictive, feature engineering ou rééquilibrage de classes. Le déséquilibre naturel de la cible y (11.27 % « yes » / 88.73 % « no ») est préservé pour analyser les patterns réels et authentiques du dataset.

Objectifs de la phase

- Réaliser des analyses univariées approfondies pour comprendre la distribution des variables clés.
- Effectuer des analyses bivariées et multivariées pour identifier les relations significatives avec la cible y (souscription au dépôt à terme).
- Appliquer des tests statistiques appropriés pour valider les hypothèses (ex : relations entre variables catégorielles/numériques et y).
- Répondre directement aux 5 questions métier du cahier des charges (PDF du projet).
- Formuler des recommandations stratégiques concrètes, basées sur les insights, avec un profil client idéal et des KPI de suivi.

Rappel sur les tests statistiques utilisés :

- **Test du Chi-deux** : Utilisé pour vérifier si deux variables catégorielles sont liées (ex : métier et souscription). Principe : Il compare les fréquences observées (dans les données) aux fréquences attendues si les variables étaient indépendantes. Si la p-value (probabilité que le résultat soit dû au hasard) est < 0.05, la relation est significative (non due au hasard).
- **Test t de Student** : Pour comparer les moyennes de deux groupes (ex. : âge moyen pour « yes » vs « no »). Principe : Il mesure si la différence observée entre moyennes est réelle ou due à la variabilité aléatoire des données. P-value < 0.05 entraine une différence significative.
- **Test ANOVA** : Extension du t-test pour plus de deux groupes (ex. : âge moyen par niveau d'éducation). Principe : Il évalue si au moins une moyenne diffère des autres. P-value < 0.05 entraine au moins une différence significative.
- **MANOVA** : Pour analyser plusieurs variables numériques dépendantes simultanément (ex. : indicateurs socio-économiques vs y). Principe : Comme l'ANOVA, mais multivarié ; il teste si les groupes diffèrent globalement sur un ensemble de variables.

Étapes réalisées

Le Notebook 3 charge le dataset nettoyé (bank-additional-full-cleaned.csv : **40 858 lignes × 21 colonnes**) et suit une structure en 6 parties, avec visualisations sauvegardées.

1. **Partie 1 : Analyses univariées approfondies** Focus sur 9 variables clés (démographiques, temporelles, campagne, financières). Pour les numériques : statistiques descriptives (moyenne, médiane, écart-type, min/max, quartiles) et skewness (asymétrie) et kurtosis (aplatissement pour détecter outliers/extremes). Pour les catégorielles : fréquences relatives (%) et barplots.

Résultats principaux (numériques) :

- age : Moyenne 40.02 ans, médiane 38, min 17, max 98 ; skewness 0.79 (queue droite : plus de seniors) ; kurtosis 0.79 (légèrement pointue). Interprétation : Clientèle mature, outliers âgés (retraités).
- campaign : Moyenne 2.57 contacts, médiane 2 ; skewness 4.76 (majorité 1-2 contacts, queue longue pour les sur-sollicités).
- previous : Moyenne 0.17, médiane 0 ; skewness 3.83 (majorité sans contacts précédents).
- Variables socio-économiques : emp.var.rate moyenne 0.08 (neutre), skewness -0.72 ; cons.conf.idx moyenne -40.5 (basse confiance, crise 2008-2010).

Résultats principaux (catégorielles) :

- job : Admin 25.4 %, blue-collar 22.5 %, étudiants 2.1 %.
- month : Mai 33.4 %, mars 1.3 % (campagnes concentrées printemps/été).
- poutcome : Nonexistent 86.2 % (majorité sans campagne précédente).
- y : Confirmé 11.27 % yes (déséquilibre réel).

Justification : Ces analyses révèlent la structure du dataset (ex : asymétries indiquant segments rares mais potentiellement clés comme les seniors).

2. **Partie 2 : Analyses bivariées (variables vs y)**

A. Variables catégorielles vs y : Crosstabs (% « yes » par catégorie) et barplots et test Chi2.

- job : Étudiants 31.3 % yes, retraités 25.2 %, blue-collar 6.9 % ; Chi2 733.96, $p < 0.001$ (significatif).
- education : Illétrés 22.2 %, university.degree 13.7 %, basic.9y 7.8 % ; Chi2 220.66, $p < 0.001$.
- marital : Célibataires 14.0 %, mariés 10.1 % ; Chi2 122.31, $p < 0.001$.
- month : Décembre 51.1 %, mars 45.1 %, mai 6.4 % Chi2 1745.37, $p < 0.001$.
- day_of_week : Légère variation (mercredi/jeudi ~11.8 %) ; Chi2 11.47, $p = 0.022$ (significatif mais faible).
- poutcome : Success 65.1 %, nonexistent 8.8 % ; Chi2 2187.96, $p < 0.001$.
- default : No 12.8 %, unknown 5.2 % ; Chi2 391.13, $p < 0.001$.

B. Variables numériques vs y : Boxplots et t-test (Welch pour variances inégales).

- age : Moyenne yes 40.91 vs no 39.91 (diff +1) ; $t=7.94$, $p < 0.001$.
- campaign : Moyenne yes 2.14 vs no 2.63 (moins de contacts pour yes) ; $t=-9.21$, $p < 0.001$.
- previous : Moyenne yes 0.49 vs no 0.13 ; $t=32.41$, $p < 0.001$.
- Socio-économiques : Ex. emp.var.rate moyenne yes -1.23 vs no 0.24 (croissance emploi favorise yes) ; $t=-53.28$, $p < 0.001$.

Justification : Ces tests confirment les relations (p-values très basses : non dues au hasard), priorisant les variables discriminantes comme poutcome et month.

3. **Partie 3 : Analyses multivariées** Crosstabs multi-variables et MANOVA pour socio-économiques.
    - month et poutcome vs y : Mars et success 81.5 % yes ; mai et nonexistent 5.6 %. Interprétation : Succès précédent booste en périodes optimistes.
    - age_group et education vs y : >60 et university.degree 28.5 % ; <30 et basic.4y 6.2 %. Interprétation : Seniors éduqués haut taux.
    - MANOVA socio-éco vs y : Wilks' lambda 0.77, F=246.3, $p < 0.001$ (groupe socio-éco discriminant globalement).

    Justification : Révèle des interactions complexes (ex : effet combiné mois et historique), invisibles en bivarié.

4. **Partie 4 : Réponses aux questions métier**

    Les analyses statistiques univariées, bivariées et multivariées réalisées dans cette phase permettent de répondre de manière concrète et data-driven aux cinq questions professionnelles posées dans le cahier des charges du projet.

    Question 1 : Quels segments de clients montrent la plus grande propension à souscrire à l'offre, et quels types d'actions marketing recommanderiez-vous pour les cibler efficacement ?

    Les segments présentant la plus forte propension à souscrire un dépôt à terme sont les suivants :

- Les clients ayant déjà souscrit lors d'une campagne précédente (**poutcome = success**) affichent un taux de conversion de **65,6 %**, soit un écart de **+56,7 points** par rapport à la moyenne globale.
- Les seniors de **65 ans et plus** présentent un taux de **46,6 %**, soit **+38,1 points** par rapport à la tranche 35-45 ans.
- Les étudiants atteignent **31,4 %** de souscription, contre seulement **6,9 %** pour les ouvriers (blue-collar).
- Les clients célibataires (**marital = single**) affichent **14,0 %** de conversion, contre **10,1 %** pour les mariés.
- Les diplômés universitaires (**university.degree**) souscrivent à **13,7 %**, contre **7,8 %** pour les niveaux basic.9y.

    Sur le plan du contexte économique, les périodes de **taux d'intérêt bas** (euribor3m faible), de **croissance de l'emploi** (emp.var.rate négatif ou faible) et de **confiance des consommateurs relativement élevée** favorisent fortement la souscription.

    Recommandations d'actions marketing : Il est fortement recommandé de cibler en priorité les anciens clients ayant déjà souscrit (success), les seniors de 65 ans et plus, ainsi que les étudiants et les célibataires diplômés. Les messages doivent être adaptés : sécurité financière et transmission patrimoniale pour les seniors ; construction d'un avenir financier pour les étudiants ; fidélisation et opportunité renouvelée pour les anciens success. Le canal cellulaire est à privilégier.

    Question 2 : En cas de contraintes budgétaires, quels critères utiliseriez-vous pour prioriser les segments à cibler ?

En situation de budget limité, la priorisation doit reposer sur deux critères principaux :

- Le **taux de conversion observé** (efficacité intrinsèque du segment),
- Le **volume potentiel** (effectif du segment dans la base).

Les segments à cibler en priorité absolue sont ceux dont le taux dépasse **30 %** :

- Anciens clients success (**65,6 %**) : priorité maximale, même en faible volume.
- Seniors 65+ (**46,6 %**) : très fort potentiel si la base contient suffisamment de retraités.
- Étudiants (**31,4 %**) : segment à fort rendement malgré un effectif plus réduit.

Les segments intermédiaires (taux 10-30 %) peuvent être inclus en deuxième vague si le budget le permet (retraités, diplômés universitaires, célibataires sans défaut). Les segments inférieurs à **8 %** (blue-collar 35-45 ans, unknown default, >6 contacts) doivent être systématiquement exclus pour éviter un gaspillage budgétaire.

Question 3 : Comment équilibrer le coût d'une campagne marketing avec les opportunités manquées liées aux faux négatifs ?

L'analyse met en évidence un trade-off clair entre intensité de sollicitation et efficacité :

- Au-delà de **2-3 contacts** par client, le taux de conversion chute significativement (**-7 points** par palier supplémentaire), traduisant une fatigue client confirmée.
- Les segments à très fort taux (>30 %) génèrent un retour sur investissement élevé même avec un nombre limité de contacts (1-2 suffisent).

Pour minimiser les faux négatifs (opportunités manquées), il est essentiel de :

- concentrer les efforts sur les segments à haut rendement (>30 %) plutôt que de disperser les contacts sur une large base peu réceptive,
- limiter strictement le nombre de relances à **2 maximum** par client, même pour les segments prioritaires, afin d'éviter la saturation.

Hypothèse économique : avec un coût moyen de **7 €** par contact et un revenu estimé de **300 €** par souscription, le seuil de rentabilité est de **2,33 %**. Tous les segments >5 % sont rentables, mais seuls ceux >30 % offrent un ROI exceptionnel (x15 à x20).

Question 4 : Quels indicateurs clés de performance (KPI) recommanderiez-vous de suivre pour évaluer le succès de la campagne ?

Les KPI suivants sont recommandés pour un suivi efficace et orienté résultats :

- **Taux de conversion global** : objectif >15 % (vs 11,27 % historique).
- **Taux de conversion par segment prioritaire** : >40 % pour mars, >60 % pour anciens success, >40 % pour seniors 65+.
- **Retour sur investissement (ROI)** : objectif >300 % (formule : (Revenus – Coûts) / Coûts × 100).
- **Coût par acquisition (CPA)** : objectif <25 € par souscription.
- **Nombre moyen de contacts par client** : objectif <2,5 (éviter sur-sollicitation).
- **Taux de satisfaction client** (post-campagne) : objectif >80 %.

Fréquence de suivi suggérée :

- Quotidien : taux de joignabilité.
- Hebdomadaire : conversion, CPA, nombre de contacts/client.
- Mensuel : ROI, conversion par segment.
- Fin de campagne : satisfaction et réclamations.

Question 5 : Comment pourriez-vous intégrer des scénarios économiques (taux d'intérêt, inflation) dans votre modèle pour anticiper les résultats futurs ?

Bien que cette phase ne traite pas de modélisation prédictive, les analyses montrent que les variables socio-économiques (**euribor3m**, **emp.var.rate**, **cons.conf.idx**) expliquent une part significative de la variance de la souscription.

Pour anticiper les résultats futurs, je pourrai :

- conserver ces cinq variables dans le modèle final (elles restent discriminantes même en présence de multicolinéarité),
- simuler différents scénarios macroéconomiques (ex : hausse de l'euribor3m à 4,5 %, chute de la confiance à -45) en appliquant le modèle entraîné sur des valeurs modifiées de ces variables
- analyser l'impact marginal de chaque indicateur via une analyse de sensibilité post-modélisation.

Ces approches permettront d'estimer l'impact prévisionnel d'une récession ou d'une reprise sur les taux de conversion attendus.

5. **Partie 5 : Recommandations stratégiques**
    - Segments prioritaires : Anciens success en mars (>60 % yes).
    - Timing : Mars prioritaire, éviter mai/juillet/novembre.
    - Budget : Allouer 60 % à success passés ; limiter 2 contacts.

5. Phase 4 : Modélisation et Évaluation

Cette phase correspond à l'étape 4 de la méthodologie CRISP-DM : **Modélisation**. L'objectif est de construire des modèles prédictifs simples pour identifier les clients les plus susceptibles de souscrire à un dépôt à terme, tout en respectant les consignes du projet : un modèle KNN obligatoire, une évaluation rapide et une forte interprétation métier (et non une optimisation complexe).

Préparation finale des données pour la modélisation

À partir du dataset nettoyé du Notebook 2 (bank-additional-full-cleaned.csv - 40 858 lignes × 21 colonnes) :

- Exclusion de la variable **duration** (data leakage : cette information n'est connue qu'après l'appel, elle ne peut pas être utilisée pour prédire).
- Encodage de la cible y : « no » en 0 et « yes » en 1.

- Séparation entre features (X : 19 variables) et cible (y).
- Identification des colonnes : 10 variables catégorielles et 9 variables numériques.
- Split stratifié : 80 % d'entraînement (32 686 lignes) / 20 % de test (8 172 lignes), afin de préserver le déséquilibre (11.27 % « yes »).
- Création d'un **ColumnTransformer** : OneHotEncoder (avec drop='first') pour les variables catégorielles et StandardScaler pour les variables numériques.

Le déséquilibre de classe étant très marqué, deux approches ont été testées : une baseline sans rééquilibrage et une approche avec **SMOTE** (uniquement sur le jeu d'entraînement).

Modèle Baseline : KNN (k=5) sans rééquilibrage

Ce premier modèle sert de point de référence pour mesurer l'impact du déséquilibre.

Performances sur le jeu de test :

- Accuracy : **0.8906** (89.06 %)
- Précision : **0.5277**
- Recall : **0.2790**
- F1-score : **0.3651**

Interprétation : L'accuracy est artificiellement élevée car le modèle prédit majoritairement la classe « no » (la plus fréquente). Cependant, le recall très faible (27.9 %) montre que le modèle ne détecte qu'un quart des vrais clients qui auraient souscrit. Cela confirme l'analyse faite a la pahse 3 : sans correction du déséquilibre, le modèle n'est pas opérationnel d'un point de vue métier.

Amélioration avec SMOTE et optimisation KNN

Application de **SMOTE** (création d'exemples synthétiques de la classe « yes » uniquement sur le train set) suivie d'une optimisation légère des hyperparamètres de KNN via GridSearchCV (n_neighbors, weights, metric).

Meilleur modèle KNN : SMOTE + k=3

- Accuracy : **0.7794**
- Précision : **0.2694**
- Recall : **0.5592**
- F1-score : **0.3636**

Analyse : Le rééquilibrage permet de presque doubler le recall (+28 points par rapport à la baseline). Le modèle détecte désormais plus de la moitié des vrais souscripteurs, ce qui est bien plus intéressant pour une campagne marketing. Le F1-score reste stable, montrant un meilleur équilibre entre précision et recall.

Modèles avancés : Random Forest et XGBoost

Pour aller plus loin, deux modèles plus puissants ont été testés avec la même pipeline (preprocessing et SMOTE) et une optimisation via GridSearchCV :

- **Random Forest** : Meilleur modèle global

    - Accuracy : 0.8521
    - Précision : 0.4286
    - Recall : **0.5700**
    - F1-score : **0.4920** (meilleur F1)
    - AUC-ROC : **0.7990**
- **XGBoost** :
    - F1-score : 0.4785
    - AUC-ROC : 0.7854

Comparaison finale des modèles

MODELE	ACCURACY	PRECISION	RECALL	F1-SCORE	AUC-ROC
BASELINE KNN (K=5)	0.8906	0.5277	0.2790	0.3651	-
KNN SMOTE (K=3)	0.7794	0.2694	0.5592	0.3636	-
RANDOM FOREST	**0.8521**	0.4286	**0.5700**	**0.4920**	**0.7990**
XGBOOST	0.8473	0.4152	0.5624	0.4785	0.7854

Modèle retenu : Random Forest Raisons du choix :

- Meilleur F1-score global (+34.8 % par rapport à la baseline).
- Excellent recall (57 %), essentiel pour ne pas rater les clients potentiels.
- Meilleure AUC-ROC, preuve d'une bonne capacité de discrimination.
- Feature Importance très interprétable et cohérente avec les analyses du Notebook 3 (poutcome, month, age, euribor3m, default ressortent parmi les variables les plus importantes).

Interprétation métier et recommandations stratégiques

Le modèle Random Forest final confirme et renforce les insights de la phase 3 :

- poutcome = success reste la variable la plus puissante (les anciens clients satisfaits ont une probabilité très élevée de resouscrire).
- Les variables temporelles (month) et démographiques (age) sont très discriminantes, comme observé statistiquement.
- Les indicateurs socio-économiques (euribor3m, emp.var.rate) jouent un rôle important : le modèle apprend que les périodes de taux bas et de confiance élevée favorisent la souscription.

7. Phase 5 : Résultats et Recommandations Finales

Cette phase correspond aux étapes 5 et 6 de la méthodologie CRISP-DM : **Évaluation** et **Présentation des résultats**. Elle synthétise l'ensemble du projet, met en perspective les découvertes des phases précédentes et formule des recommandations actionnables pour optimiser les campagnes de télémarketing. Basée sur les analyses statistiques et la modélisation, cette phase évalue l'impact business potentiel et propose un plan de mise en œuvre progressif.

Synthèse du projet

Le projet a analysé un dataset de 41 188 clients (réduit à 40 858 après nettoyage minimal) couvrant des campagnes de télémarketing entre mai 2008 et novembre 2010. L'objectif central était de prédire la souscription à un dépôt à terme (y = yes/no) tout en extrayant des insights stratégiques pour améliorer l'efficacité des campagnes.

Méthodologie appliquée (CRISP-DM) : Le projet a été structuré en 5 notebooks, couvrant toutes les étapes :

- **Phase 1 (Compréhension)** : Exploration initiale, identification du déséquilibre (11.27 % « yes ») et exclusion de duration (data leakage).
- **Phase 2 (Préparation)** : Traitement conservateur des « unknown » (gardés comme catégorie dans 5 variables), conservation des outliers et analyse des corrélations (fortes dans les socio-économiques, toutes gardées).
- **Phase 3 (Analyse)** : Univariées (9 variables), bivariées (8 vs y), multivariées (interactions) avec tests statistiques (Chi 2, ANOVA, t-test, MANOVA) - tous significatifs pour les variables clés.
- **Phase 4 (Modélisation)** : KNN baseline, amélioré avec SMOTE, comparaison avec Random Forest et XGBoost - modèle final : Random Forest (F1-score 0.4920).
- **Phase 5 (Évaluation)** : Synthèse globale, impact chiffré et recommandations.

Données finales utilisées : 40 858 clients × 20 variables (après exclusions). Déséquilibre préservé pour les analyses réelles, corrigé via SMOTE en modélisation.

Résultats clés par phase

Phase 1 : Compréhension des données

- Dataset de qualité : aucune valeur NULL, mais des « unknown » à traiter.
- Déséquilibre majeur : 11.27 % « yes », nécessitant une gestion spécifique en modélisation.
- duration identifiée comme non utilisable (prédit parfaitement mais post-hoc).

Phase 2 : Préparation et nettoyage

- Perte minimale : 0.80 % (330 lignes supprimées pour job=unknown).
- 13 288 outliers conservés (valeurs réelles informatives).
- Corrélations socio-économiques fortes (ex. : emp.var.rate et euribor3m = 0.972) – gardées pour nuance métier.

Phase 3 : Analyses statistiques

- Variables les plus discriminantes (top 3 par écart de taux) :
  - poutcome : Success 65.58 % vs nonexistent 8.84 % (+56.74 points) - Chi2 $p < 0.001$.
  - month : Mars 50.74 % vs mai 6.47 % (+44.27 points) – Chi2 $p < 0.001$.
  - age : 65+ ans 46.63 % vs 35-45 ans 8.52 % (+38.12 points) – Chi2 $p < 0.001$.
- Autres : job (étudiants 31.43 % vs blue-collar 6.89 %), education (+14.37 points), default (+7.69 points).
- Multivariées : Interactions fortes (ex. : mars + success = 81.5 % yes) - MANOVA socio-éco $p < 0.001$.

Phase 4 : Modélisation

- Modèle retenu : **Random Forest** avec SMOTE (amélioration +34.8 % F1 vs baseline).
- Performances : F1-score 0.4920, recall 0.5700 (57 % des opportunités détectées), AUC-ROC 0.7990.
- Feature Importance : poutcome, month, age, euribor3m, default confirme les insights statistiques.

Profil client idéal

Profil cible prioritaire (probabilité de souscription 60-70 %) :

- poutcome = success (ancien client ayant souscrit).
- Contacté en mars (ou septembre/décembre).
- Âge 60-70 ans (retraité).
- job = student ou retired.
- default = no (crédit évalué sans défaut).
- marital = single.
- Maximum 1-2 contacts.

Profil à éviter (probabilité < 5 %) :

- Contacté en mai, juillet ou novembre.
- Âge 35-45 ans.
- job = blue-collar.
- default = unknown (non évalué).
- Plus de 6 contacts déjà effectués.
- poutcome = failure.

Modèle prédictif final

Modèle retenu : **Random Forest** (optimisé avec GridSearchCV sur n_estimators, max_depth, etc.).

- **Performances détaillées** : Voir tableau ci-dessus (comparaison).
- **Interprétation** : Le modèle excelle en recall (57 %), essentiel pour capturer les opportunités. L'AUC-ROC de 0.7990 indique une bonne discrimination globale.
- **Utilisation** : Scoring automatisé des clients avant appel prioriser ceux avec probabilité > 0.35.

Recommandations stratégiques

Stratégie court terme (0-1 mois) :

- Planifier une campagne pilote en mars 2026 : cibler 1 354 anciens « success », 609 seniors 65+ et 875 étudiants (budget environ 30 000 €, souscriptions attendues : 1 400-1 500, taux 45-50 %, ROI x15-20).
- Limiter à 2 contacts max par client pour éviter la fatigue.
- Adapter les messages : sécurité pour seniors, avenir pour étudiants, fidélité pour success.

Stratégie moyen terme (1-3 mois) :

- Implémenter le scoring Random Forest dans le CRM pour un ciblage mensuel.

- Tester sur un échantillon pilote (1 000 clients : test vs contrôle) pour valider les hypothèses.
- Former les équipes commerciales sur les profils prioritaires et la gestion de la fatigue client.

Stratégie long terme (3-6 mois) :

- Enrichir la base de données avec des scores externes (comportementaux, crédit).
- Réentraîner le modèle sur de nouvelles données pour une amélioration continue.
- Industrialiser : dashboard KPI temps réel et monitoring automatique.

Règle de décision budgétaire : Utiliser un score = (Taux × 0.6) + (Effectif × 0.0001) - (Coût × 0.4) ; prioriser score > 30.

Impact business estimé

Scénario actuel (sans optimisation) :

- Taux de conversion : 11.27 %.
- Sur 10 000 contacts : environ 1 127 souscriptions.
- Coûts : 70 000 € (7 €/contact).
- Revenus : 338 100 € (300 €/souscription).
- Bénéfice : 268 100 €.

Scénario optimisé (avec ciblage et modèle) :

- Taux : 30-40 %.
- Souscriptions : 3 500-4 000.
- Revenus : 1 050 000 - 1 200 000 €.
- Bénéfice : 980 000 - 1 130 000 €.
- **Gain** : +700 000 € (+260 %).
- **ROI** : +337 % vs approche actuelle.

Gains qualitatifs : Réduction du gaspillage (moins de contacts inutiles), meilleure satisfaction client (moins de sur-sollicitation).

Limites et améliorations

Limites :

- Déséquilibre des classes : Impacte les métriques malgré SMOTE.
- Période des données (2008-2010) : Contexte économique daté (crise) - réentraînement nécessaire sur données récentes.
- Modèle simple : Performances modérées (F1 environ 0.49) donc pas d'optimisation avancée.
- Pas d'enrichissement externe : Limite la précision (ex. : données comportementales manquantes).

Améliorations proposées :

- Ajout de features (ex. : scores crédit externes).
- Tester des modèles avancés (ex. : stacking KNN + RF).

- Valider en production avec A/B testing.
- Intégrer un dashboard interactif (Streamlit) pour scoring temps réel.

5.8 Conclusion générale

Ce projet démontre le pouvoir de l'analyse de données et du machine learning pour transformer les campagnes marketing bancaires. En identifiant les segments prioritaires (anciens success, seniors, étudiants) et en optimisant le timing (mars/septembre/décembre), nous multiplions le ROI par plus de 3 tout en réduisant les coûts inutiles. Le modèle Random Forest retenu, avec un recall de 57 %, permet de capturer 268 opportunités supplémentaires sur 10 000 contacts.

Les recommandations sont immédiatement applicables : une campagne pilote en mars 2026 pourrait générer un gain net de +700 000 €. À long terme, cette approche data-driven rend les stratégies marketing proactives et adaptables aux évolutions du marché.

Le projet, réalisé avec rigueur suivant CRISP-DM, fournit tous les outils pour une implémentation réussie. Il reste à passer à l'action pour maximiser la rentabilité et l'expérience client.

Auteur : Adja Fatou SAGNA (KIM) **Date de finalisation** : 27 février 2026