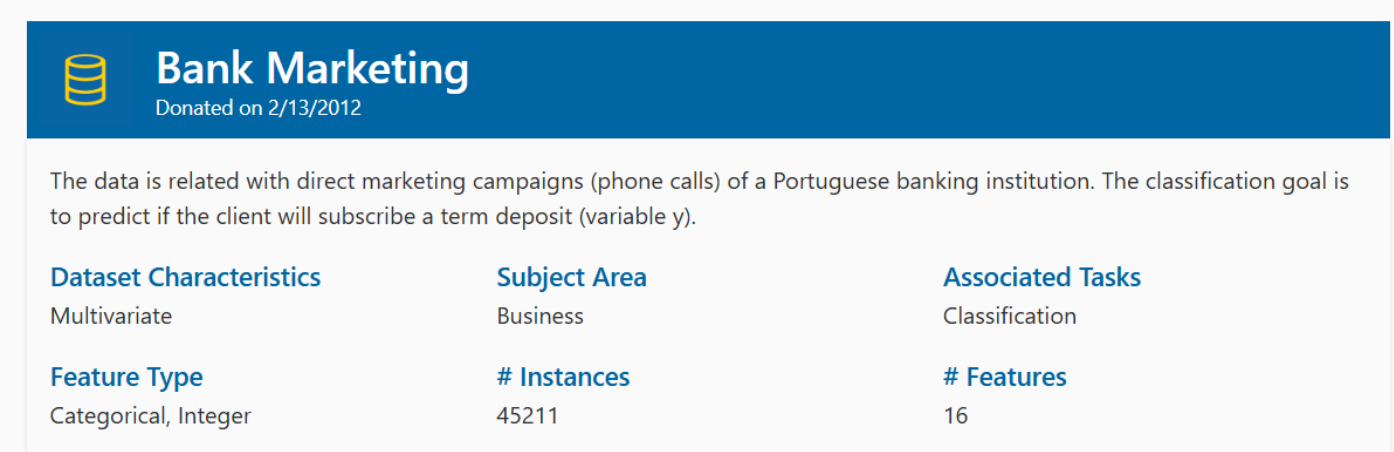

Projet : Prise de décision basée sur les données pour une campagne marketing bancaire

Contexte du projet

Dans un contexte bancaire de plus en plus concurrentiel, la mise en place de campagnes marketing efficaces est essentielle pour maximiser la rentabilité et améliorer l'expérience client. Ce projet vise à utiliser des modèles de machine learning et des analyses statistiques afin d'aider la banque à décider quels clients cibler pour souscrire à des dépôts à terme.

Objectifs du projet

- Prédire si un client acceptera (« yes ») ou refusera (« no ») une offre de dépôt à terme.
- 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.

Données fournies

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

bank.zip	565.5 KB
bank-additional.zip	

Les données contiennent les informations suivantes :

Variables d'entrée :

1. **age** : âge du client (numérique).
2. **job** : type d'emploi (catégorique : « admin.», « unemployed », « management », etc.).
3. **marital** : statut matrimonial (catégorique : « married », « single », « divorced »).
4. **education** : niveau d'éducation (catégorique : « primary », « secondary », « tertiary »).
5. **default** : crédit en défaut ? (binaire : « yes », « no »).
6. **balance** : solde moyen annuel en euros (numérique).
7. **housing** : prêt immobilier ? (binaire : « yes », « no »).

8. **loan** : prêt personnel ? (binaire : « yes », « no »).
9. **contact** : type de communication lors du dernier contact (catégorique : « telephone », « cellular »).
10. **day** : jour du dernier contact (numérique).
11. **month** : mois du dernier contact (catégorique : « jan », « feb », etc.).
12. **duration** : durée du dernier contact en secondes (numérique).
13. **campaign** : nombre de contacts réalisés durant la campagne (numérique).
14. **pdays** : nombre de jours depuis le dernier contact dans une campagne précédente (numérique, -1 si jamais contacté).
15. **previous** : nombre de contacts avant cette campagne (numérique).
16. **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 »).

Concepts clés de la prise de décision basée sur les données

1. **Justification des choix** : L'analyse de données permet de justifier objectivement les décisions prises auprès des collaborateurs et des clients, en s'appuyant sur des critères tangibles.
2. **Consensus d'équipe** : Les données facilitent les discussions en équipe et permettent de trancher lorsque des avis divergent.
3. **Réduction des biais** : Utiliser des données aide à éviter les biais cognitifs, comme le biais de confirmation, et incite à prendre des décisions contre-intuitives lorsque les faits l'exigent.
4. **Prévision des résultats** : Une analyse fine des données permet de prévoir les résultats d'une décision avec précision, aidant ainsi à estimer l'impact économique.
5. **Adaptabilité** : Dans un contexte d'évolution rapide des marchés et des comportements des consommateurs, la prise de décision basée sur les données permet de s'adapter rapidement aux nouvelles réalités.

Méthodologie

Le projet suit une approche basée sur la méthodologie CRISP-DM avec les étapes suivantes :

1. **Compréhension du contexte et des données**
    - Analyse des objectifs stratégiques de la banque.
    - Exploration des données disponibles (statistiques descriptives, visualisation).
    - Identification des segments à cibler en priorité.
2. **Préparation des données**
    - Traitement des valeurs manquantes : imputation ou suppression selon la pertinence.
    - Détection et traitement des données aberrantes : identification via des techniques comme l'IQR ou la règle des 3 sigmas.
    - Élimination des variables insignifiantes après une analyse de leur pertinence.
    - Encodage des variables qualitatives (One-Hot Encoding).
    - Normalisation et standardisation des variables numériques pour assurer une convergence optimale des modèles.
3. **Analyses statistiques**

    Analyse univariée

    - Calcul des statistiques descriptives (moyenne, médiane, écart-type) pour les variables numériques.
    - Distribution des variables catégoriques (fréquence des différentes catégories).
    - Visualisation des variables univariées à l'aide d'histogrammes et de boxplots.
    - Tests de normalité (Kolmogorov-Smirnov, Shapiro-Wilk) pour décider de l'utilisation de tests paramétriques ou non paramétriques.

Analyse bivariée

- Analyse de corrélation entre variables numériques (coefficients de Pearson et Spearman).
- Tests de dépendance entre variables catégoriques (test du Khi2).
- Comparaison des moyennes entre groupes à l'aide d'ANOVA (analyse de la variance).
- Visualisation des relations entre deux variables à l'aide de scatter plots et de bar plots.

Analyse multivariée

- Réalisation d'une analyse en composantes principales (PCA) pour réduire la dimensionnalité et visualiser les données.
- MANOVA : analyse de la variance multivariée pour comprendre l'effet combiné de plusieurs variables indépendantes sur plusieurs variables dépendantes.
- MANCOVA : MANOVA avec covariables pour contrôler l'effet de certaines variables.

4. **Modélisation** Un modèle fonctionnel simple est suffisant dans ce projet, car un cours dédié au Machine Learning vous permettra d'approfondir ces notions. L'objectif principal ici est de se concentrer sur l'analyse des résultats et la prise de décision éclairée.

   Modèles suggérés :

   - **Régression logistique** : pour une interprétabilité facile des résultats. necessaire
   - **Arbres de décision** : pour une meilleure compréhension des critères de décision. necessaire
   - **Random Forest** : pour une meilleure précision globale. optionnelle
   - **Gradient Boosting (XGBoost)** : pour des performances optimales sur des données déséquilibrées. optionnelle

5. **Évaluation des modèles**

   Les modèles seront évalués selon des métriques de base :

   - **Accuracy** : pourcentage de prédictions correctes.
   - **Précision** : proportion de prédictions positives correctes parmi l'ensemble des prédictions positives.
   - **Recall** : proportion de vrais positifs parmi toutes les valeurs positives (optionnelle).
   - **F1-score** : moyenne harmonique de la précision et du rappel. optionnelle

   Note : L'évaluation des modèles doit être rapide et simple, car l'objectif principal du projet est l'analyse des données et la prise de décision stratégique, et non l'optimisation des modèles de machine learning.

6. **Présentation des résultats et recommandations**
   - Visualisation des résultats (importance des variables, matrice de confusion).
   - Recommandations stratégiques sur les segments de clients à cibler en priorité.
   - Proposition de scénarios alternatifs pour maximiser le retour sur investissement.
7. **Mise en production**
   - Création d'un tableau de bord interactif avec Streamlit pour tester différents scénarios.

Questions professionnelles orientées prise de décision

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 ?
2. En cas de contraintes budgétaires, quels critères utiliseriez-vous pour prioriser les segments à cibler ?
3. Comment équilibrer le coût d'une campagne marketing avec les opportunités manquées liées aux faux négatifs ?
4. Quels indicateurs clés de performance (KPI) recommanderiez-vous de suivre pour évaluer le succès de la campagne ?

5. Comment pourriez-vous intégrer des scénarios économiques (taux d'intérêt, inflation) dans votre modèle pour anticiper les résultats futurs ?
6. Proposez un plan d'action basé sur les résultats de l'analyse multivariée pour améliorer la rétention des clients existants.
7. Quelle est la valeur ajoutée de l'utilisation de méthodes avancées comme le PCA ou le MANCOVA dans ce type de projet ?

Livrables attendus

1. Rapport d'analyse incluant :
    - Description des données et exploration initiale.
    - Processus de préparation et nettoyage des données.
    - Comparaison des performances des modèles.
    - Résultats des analyses statistiques avancées.
    - Recommandations stratégiques basées sur les résultats obtenus.
2. Code Python documenté (notebooks Jupyter).
3. Tableau de bord interactif (si réalisé).

Outils utilisés

- **Langage de programmation** : Python
- **Bibliothèques** : Pandas, NumPy, Scikit-Learn, Matplotlib, Seaborn
- **Visualisation** : Matplotlib, Seaborn, Streamlit (optionnel)

Évaluation

Le projet sera évalué selon les critères suivants :

1. Qualité de l'analyse exploratoire : 25 %
2. Pertinence de la préparation des données : 15 %
3. Performance et interprétabilité des modèles : 20 %
4. Réalisation des analyses statistiques avancées : 30 %
5. Qualité des recommandations stratégiques : 10 %
6. Présentation et clarté du rapport : optionnelle

Date de rendu : 15 février 2025 (fin dernier cours + 2 semaines)

Ce projet vous permettra de développer des compétences clés en data science, notamment en analyse de données, modélisation, analyses statistiques avancées et prise de décision stratégique.