ÉTUDE DU DATASET BANK MARKETING (UCI)

EL AMRAOUI ABIR

24010353

SOMMAIRE

1. Problématique, métier et objectif de la mission

1.1 Problématique du métier

Dans le secteur bancaire, les campagnes de télé-marketing constituent un levier stratégique pour promouvoir les dépôts à terme, un produit d'épargne essentiel pour assurer la liquidité et la stabilité financière des banques.

Cependant, ces campagnes présentent plusieurs limites opérationnelles et financières :

- Coûts élevés liés aux centres d'appels.
- Taux de conversion faible, car la majorité des clients contactés ne souscrivent pas au produit.
- Ciblage inefficace, entraînant une surcharge des agents et une allocation non optimale du budget marketing.
- Perte d'opportunités commerciales, lorsque des clients potentiellement intéressés ne sont pas identifiés à temps.

Ainsi, la banque se trouve confrontée à un défi majeur :

Comment identifier, avant même l'appel, les clients ayant la plus forte probabilité de souscrire à un dépôt à terme, afin d'optimiser les ressources et maximiser le rendement des campagnes ?

1.2 L'Enjeu Décisionnel : Une Matrice de Coûts d'Erreur Asymétrique

Contrairement à des problématiques classiques de classification, ici toutes les erreurs n'ont pas le même impact :

- Faux Positif (FP) : un client est prédit comme intéressé alors qu'il ne souscrira pas
  ➤ Coût opérationnel inutile, perte de temps agent, saturation des lignes d'appels
- Faux Négatif (FN) : un client réellement intéressé est classé comme non pertinent
  ➤ Manque à gagner commercial direct, perte potentielle de valeur client à long terme

Dans ce contexte, la priorité stratégique n'est pas seulement la précision, mais surtout :

Maximiser le rappel (Recall) de la classe positive, afin de réduire au maximum les faux négatifs et ne pas laisser passer les clients intéressés.

1.3 Objectifs de la mission IA

La mission consiste donc à concevoir un Assistant IA d'aide au ciblage client, basé sur le machine learning, permettant de :

1. Analyser le comportement et le profil client
2. Prédire la probabilité de souscription avant le contact téléphonique
3. Optimiser le ciblage des campagnes marketing
4. Réduire les coûts opérationnels liés aux appels non pertinents
5. Augmenter le taux de conversion et le rendement global des campagnes
6. Fournir un outil scalable pouvant être intégré dans un système CRM bancaire

1.4 Les Données Utilisées (L'Input du Modèle)

Pour répondre à cette mission, nous exploitons le dataset Bank Marketing UCI, qui contient des données historiques issues de campagnes réelles :

- Profil socio-démographique : âge, profession, niveau d'éducation, situation familiale…
- Indicateurs financiers : solde bancaire (balance)
- Historique marketing :
- nombre de contacts lors de la campagne (campaign), résultat des campagnes précédentes (poutcome), nombre de contacts antérieurs (previous), délai depuis le dernier contact (pdays)…
- Variable cible (y) : souscription au dépôt à terme → yes ou no
- Taille et format :
  - 45 211 observations
  - 17 variables
  - Données tabulaires, mélangeant variables numériques et catégorielles

Ces données permettent au modèle d'apprendre à partir de tendances comportementales réelles, tout en tenant compte des contraintes économiques du métier.

	age	job	marital	education	default	balance	housing	loan	\
0	58	management	married	tertiary	no	2143	yes	no	
1	44	technician	single	secondary	no	29	yes	no	
2	33	entrepreneur	married	secondary	no	2	yes	yes	
3	47	blue-collar	married	NaN	no	1506	yes	no	
4	33	NaN	single	NaN	no	1	no	no	

	contact	day_of_week	month	duration	campaign	pdays	previous	poutcome	y
0	NaN	5	may	261	1	-1	0	NaN	no
1	NaN	5	may	151	1	-1	0	NaN	no
2	NaN	5	may	76	1	-1	0	NaN	no
3	NaN	5	may	92	1	-1	0	NaN	no
4	NaN	5	may	198	1	-1	0	NaN	no

2. Code python simplifié

```
# INSTALLATION & IMPORTS
# pip install ucimirep:seaborn matplotlib pandas numpy
from ucimlrepo import fetch_ucirepo
import pandas as pd
import numpy as np
import matplotbib.pyplot as plt
import seaborn as sns

# 1) CHARGEMENT DES DONNÉES
bank_marketing = fetch_ucirepo(id=222)
X = bank_marketing.data.features
y = bank_marketing.data.targets
df = pd.concat([X, y], axis=1)

# 2) ANALYSE DES VARIABLES NUMERQUES
num_cols = df.select_dtypes(include=['int64', 'float64']..columns
# HISTOGRAMMES
for col in nn.cols:
   sns.histplot( df[col]), kde=True)
   plt.title(f'Distribution de 'col')
plt.show()
# BOXPLOTS
for col in nn cols:
   sns.boxplot(x=df[col])
   plt.title(''Boxplot de l
plt.show()
# HEATMAP
plt.figure(figzeine=(10,6)
corr = dfinum_cols,corr()
sns.heatmap:corr, annot=True:, cmap="coolwarm"')
plt.title("Matrice de correlation")
  :   |    }
plt.show()

# 4) MODELISATION : LOGISTIC REGRESSION + RANDOM FOREST
from skleern.model_selection +train_test_split
from skleern.processing : OneHotEncoder
from skleern.composse ı ColumnTransformer
from skleern.pipeline ı Pipeline
from skleern.linear_model LogisticRegression
from skleern.ensemble . RandomForestClassifier
from skleern.metrics ı classification_report, conupe.ımattric

df' y' = df' y').map('yes':1, 'no ! 0"))

X = df.drop:columns['y'])
y = df'y'

num_cols = X select_dtypes(include=[`int64`,`float64`].columms
cat_cols =  X select_dtypes(include=['0bject','category`).columns
preprocess = Column trandexmner
    transformers=[: {cot, OnecOroder.maadunxrowr`'ignore', cat_cols)
    remainder= pasthrougn')
)),
```

Interprétation du code:

Le code présenté ci-dessus illustre l'intégralité du pipeline de traitement et de modélisation appliqué au dataset Bank Marketing (UCI). Il suit une approche méthodologique rigoureuse, structurée autour de quatre axes principaux :

1. **Chargement et consolidation des données**, permettant d'unifier les variables explicatives (X) et la variable cible (y) pour faciliter l'analyse et le prétraitement.
2. **Analyse statistique des variables numériques** (EDA), réalisée à l'aide d'histogrammes, de boxplots et d'une matrice de corrélation. Cette étape permet d'examiner :
    - la forme des distributions,
    - la dispersion des données,
    - la présence de valeurs extrêmes pouvant influencer les performances des modèles prédictifs.
3. **Feature Engineering**, visant à enrichir le jeu de données par la création de nouvelles variables plus interprétables et pertinentes pour le métier bancaire, notamment :
    - age_group pour la segmentation client,
    - duration_min pour convertir la durée d'appel en minutes,
    - et total_contacts pour mesurer l'intensité du contact marketing.
4. **Modélisation supervisée**, basée sur deux algorithmes complémentaires :
    - La Régression Logistique, utilisée comme modèle linéaire de référence (baseline), adaptée à une première estimation de la probabilité de souscription.
    - Le Random Forest, modèle non-linéaire puissant, capable de capturer des interactions complexes entre les variables et de réduire le risque de surapprentissage grâce à l'agrégation d'arbres multiples.

Ce pipeline technique constitue une base solide pour la prédiction client, tout en tenant compte de l'enjeu prioritaire du projet : Minimiser les faux négatifs en maximisant le rappel de la classe positive, dans un contexte où le coût des erreurs est fortement asymétrique.

3. Résultat de l'analyse exploratoire

3.1 Statistique descriptive

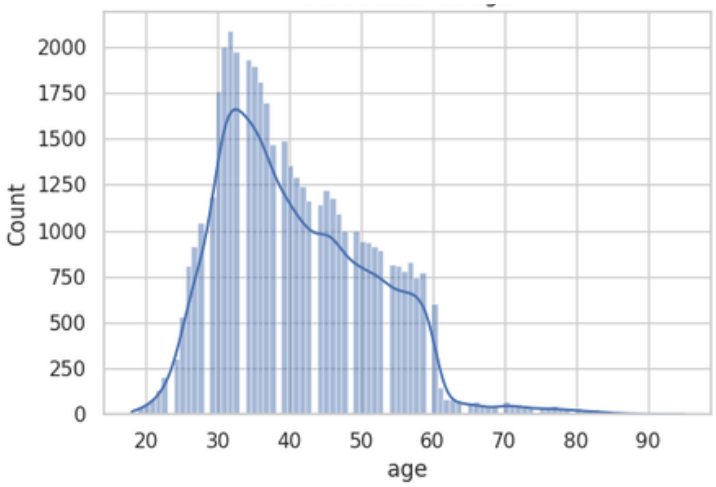
	age	balance	day_of_week	duration	campaign
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841
std	10.618762	3044.765829	8.322476	257.527812	3.098021
min	18.000000	-8019.000000	1.000000	0.000000	1.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000

	pdays	previous
count	45211.000000	45211.000000
mean	40.197828	0.580323
std	100.128746	2.303441
min	-1.000000	0.000000
25%	-1.000000	0.000000
50%	-1.000000	0.000000
75%	-1.000000	0.000000
max	871.000000	275.000000

La différence observée entre la moyenne et la médiane pour plusieurs variables indique des distributions asymétriques, souvent dues à la présence de valeurs extrêmes.

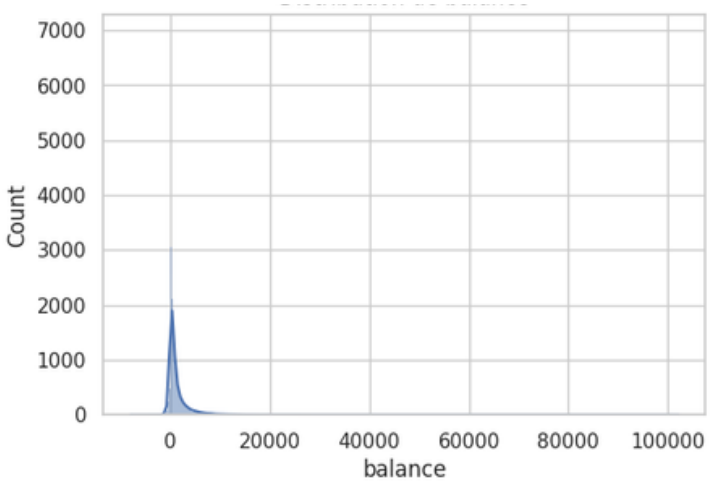
3.2 Histogrammes _ distribution

Distribution âge:



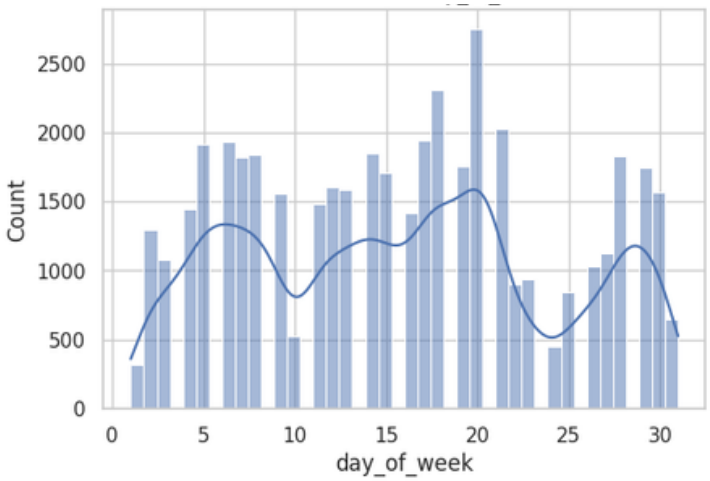
Une concentration importante des clients dans les classes d'âge intermédiaires entre 30 et 60 ans. Cette distribution asymétrique suggère que la clientèle bancaire est majoritairement composée d'adultes actifs, ce qui est cohérent avec la cible des produits d'épargne à terme. La présence de clients très jeunes ou très âgés reste marginale, indiquant des segments spécifiques pouvant nécessiter des stratégies marketing différenciées.

Distribution balance:



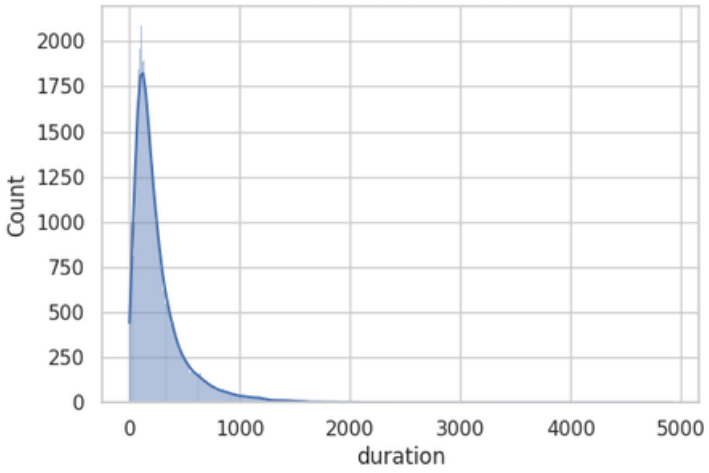
La distribution du solde bancaire est fortement asymétrique à droite, avec une majorité de clients présentant des soldes faibles ou proches de zéro, et quelques valeurs très élevées. Cette forte dispersion met en évidence la présence d'outliers, typique des données financières.

Distribution day of week:



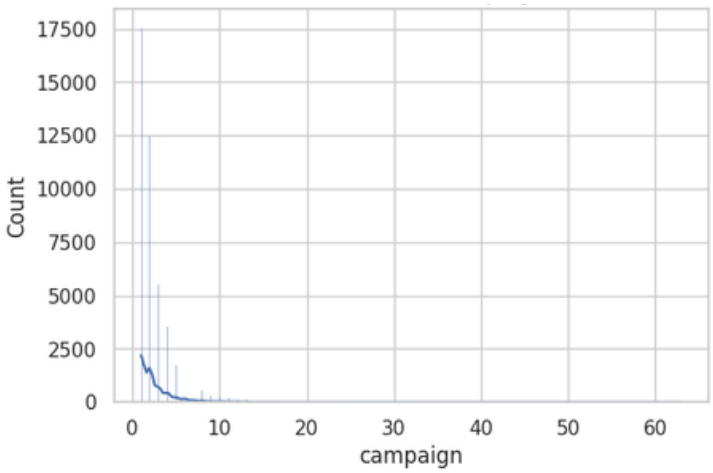
L'histogramme du jour de la semaine montre une répartition relativement homogène des appels marketing sur les différents jours ouvrables. Cela suggère une organisation équilibrée des campagnes de contact, sans concentration excessive sur un jour spécifique, ce qui limite les biais temporels dans les données.

Distribution duration:



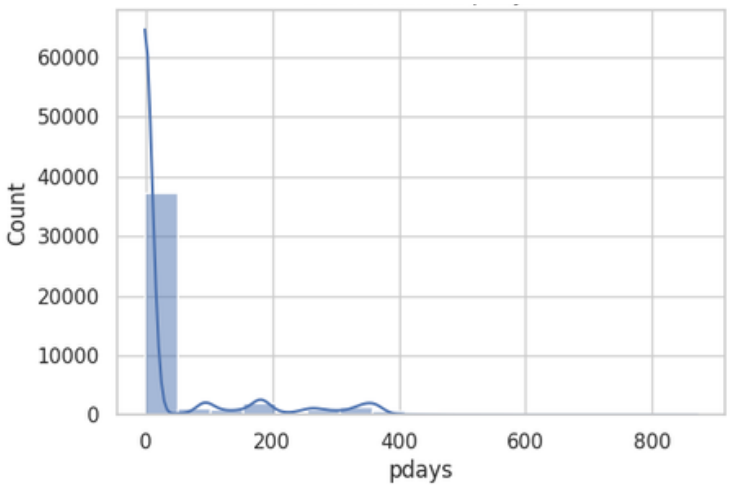
La durée des appels présente une distribution très asymétrique, avec une majorité d'appels courts et un faible nombre d'appels très longs. Cette variable est particulièrement discriminante, car les appels plus longs sont souvent associés à un intérêt plus marqué du client. Toutefois, son utilisation doit être prudente, car elle peut introduire une fuite de données si elle est observée après la décision de souscription.

Distribution nombre du contact:



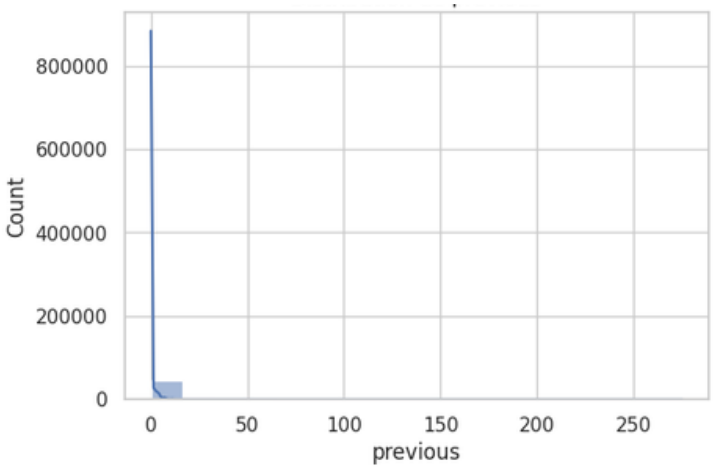
La majorité des clients est contactée un nombre limité de fois durant une campagne, tandis qu'un petit nombre de clients fait l'objet de sollicitations répétées. Cette asymétrie reflète une stratégie de relance ciblée, mais peut également indiquer un risque de saturation ou de rejet de la part des clients fortement sollicités.

Distribution pdays:



La distribution de pdays met en évidence une forte concentration sur certaines valeurs spécifiques, traduisant des périodes standardisées de relance commerciale. Les valeurs élevées indiquent des clients peu ou jamais recontactés, ce qui peut influencer leur probabilité de souscription.

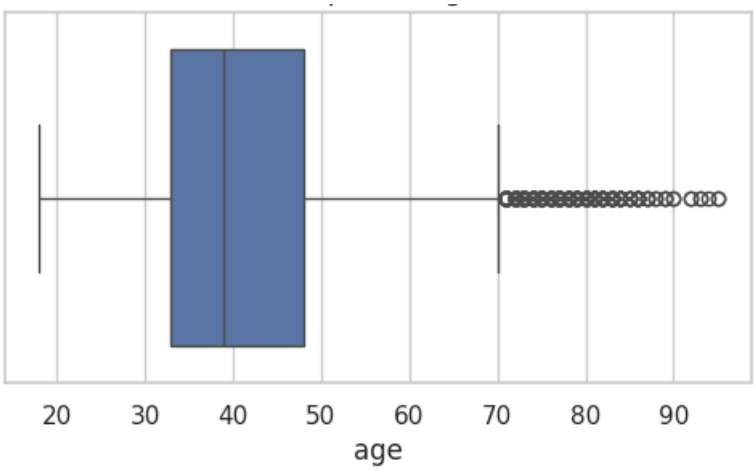
Distribution previous:



La variable previous montre que la majorité des clients n'a eu que peu ou pas de contacts lors de campagnes précédentes. Cela suggère que la base clients est régulièrement renouvelée et que les campagnes antérieures ont un impact limité pour une grande partie de la population.

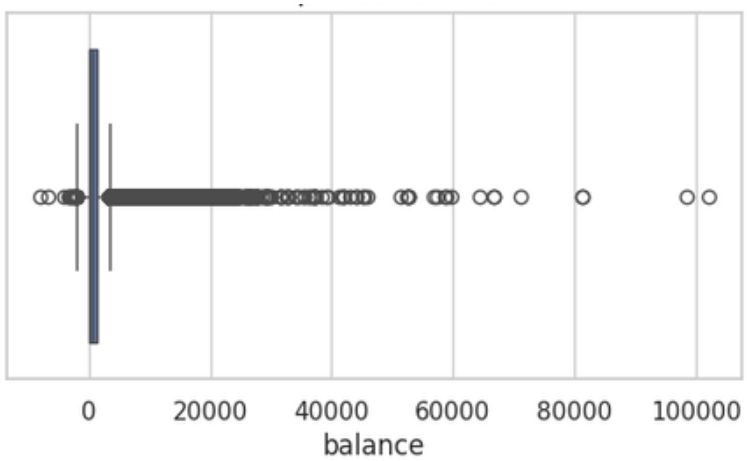
3.3 Boxplot

Boxplot âge:



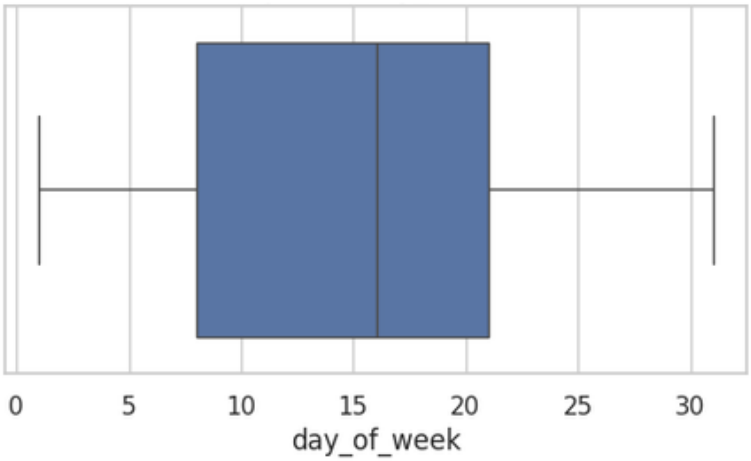
Le boxplot de l'âge révèle une dispersion modérée autour de la médiane, avec quelques valeurs aberrantes correspondant à des clients très âgés. Ces outliers restent cohérents dans un contexte bancaire et ne nécessitent pas forcément d'exclusion, mais ils doivent être pris en compte dans l'interprétation des résultats.

Boxplot balance:



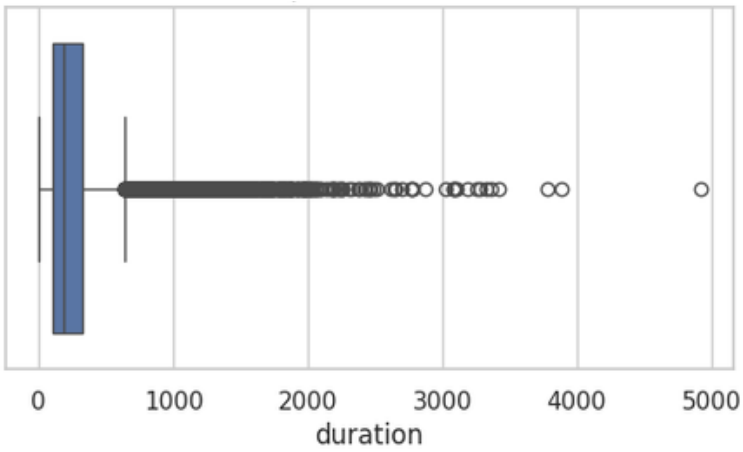
Le boxplot du solde bancaire met en évidence de nombreux outliers positifs, confirmant une forte hétérogénéité financière entre les clients. Cette variabilité est typique des données bancaires et justifie l'utilisation de modèles robustes aux valeurs extrêmes, comme les méthodes ensemblistes.

Boxplot day of week:



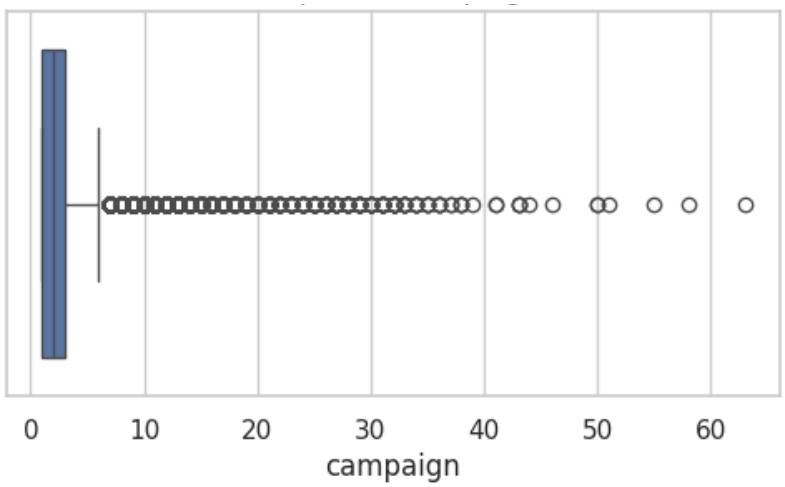
Le boxplot du jour de la semaine montre une faible dispersion autour de la médiane et peu de valeurs extrêmes. Cela indique une répartition relativement homogène des contacts sur les différents jours ouvrables, sans concentration excessive susceptible d'introduire un biais temporel dans les campagnes marketing.

Boxplot duration:



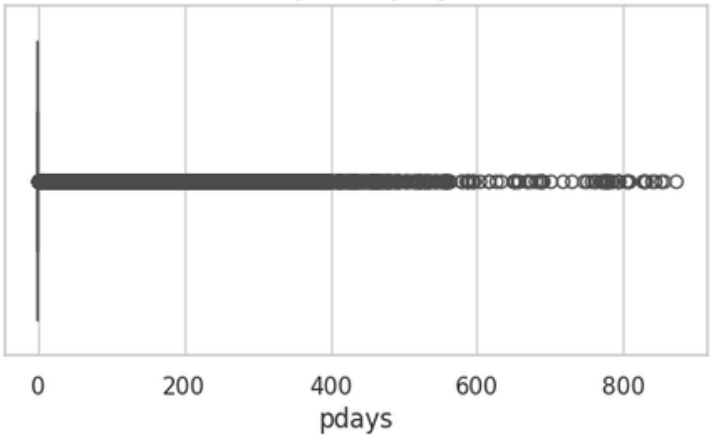
La présence de nombreux points extrêmes dans la durée des appels confirme une forte asymétrie de la variable. Ces valeurs élevées correspondent à des interactions longues, potentiellement associées à une forte probabilité de souscription, mais elles peuvent aussi biaiser certains algorithmes si elles ne sont pas correctement traitées.

Boxplot compaign:



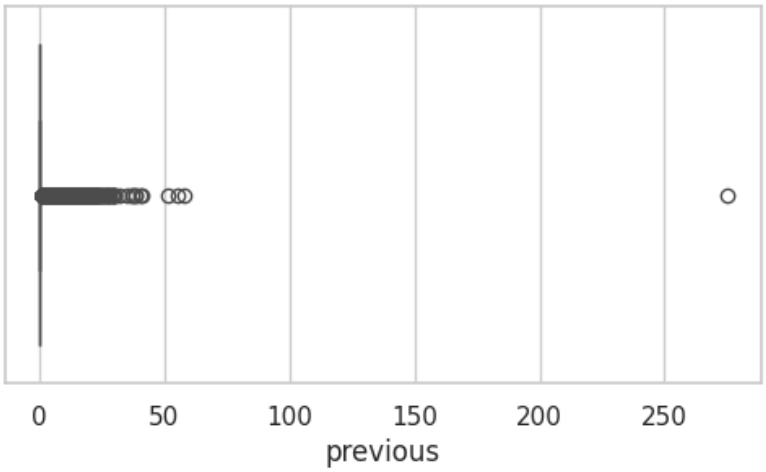
Le boxplot de la variable campaign met en évidence une distribution asymétrique avec plusieurs valeurs aberrantes. La majorité des clients est contactée un nombre limité de fois, tandis qu'un petit groupe fait l'objet de relances répétées, traduisant une stratégie commerciale ciblée mais potentiellement intrusive.

Boxplot pdays:



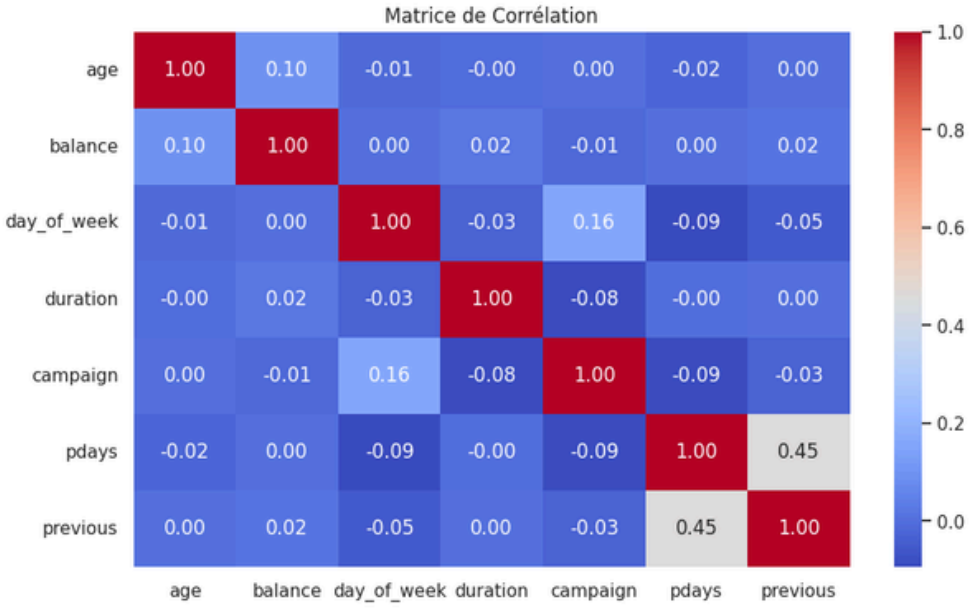
Le boxplot de pdays révèle une distribution très asymétrique, marquée par des valeurs extrêmes élevées. Ces observations correspondent à des clients n'ayant pas été recontactés depuis longtemps, voire jamais, ce qui peut influencer significativement leur probabilité de souscription.

Boxplot previous:



La variable previous présente une dispersion faible pour la majorité des observations, avec quelques valeurs extrêmes. Cela suggère que la plupart des clients ont eu peu de contacts lors des campagnes précédentes, alors qu'une minorité a été sollicitée de manière plus intensive.

3.4 Matrice de corrélation



La matrice de corrélation montre globalement de faibles corrélations entre les variables numériques, indiquant une faible redondance de l'information. La variable duration se distingue par une corrélation plus élevée avec la variable cible, ce qui confirme son pouvoir explicatif. L'absence de multi-colinéarité forte est favorable à l'utilisation de modèles comme la régression logistique.

3.5 Feature Engineering

Le feature engineering vise à enrichir le jeu de données initial en créant de nouvelles variables plus interprétables et plus pertinentes pour la modélisation prédictive. Dans cette analyse, trois nouvelles variables ont été construites:

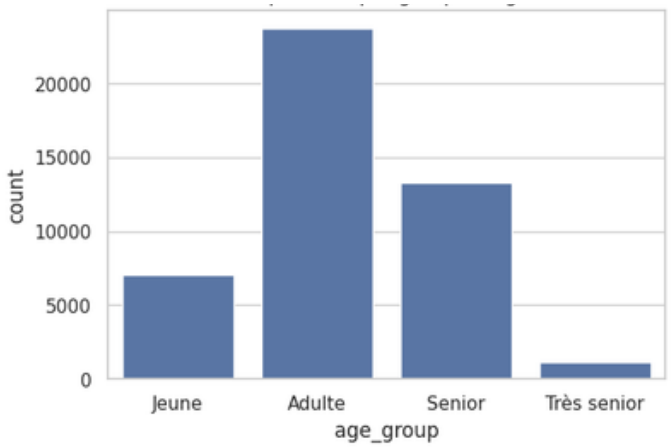
La variable **age_group** permet de segmenter les clients en classes d'âge homogènes, facilitant l'interprétation métier et l'identification des segments les plus susceptibles de souscrire à un dépôt à terme. Cette transformation réduit également la complexité liée à la variabilité individuelle de l'âge.

La variable **duration_min**, obtenue par conversion de la durée des appels en minutes, améliore la lisibilité et l'interprétation de l'intensité des interactions commerciales, tout en conservant l'information essentielle contenue dans la variable initiale.

Enfin, la variable **total_contacts**, qui agrège le nombre de contacts de la campagne en cours et des campagnes précédentes, fournit une mesure globale de la pression commerciale exercée sur le client. Cette information est particulièrement pertinente pour analyser l'impact de la fréquence des sollicitations sur la probabilité de souscription.

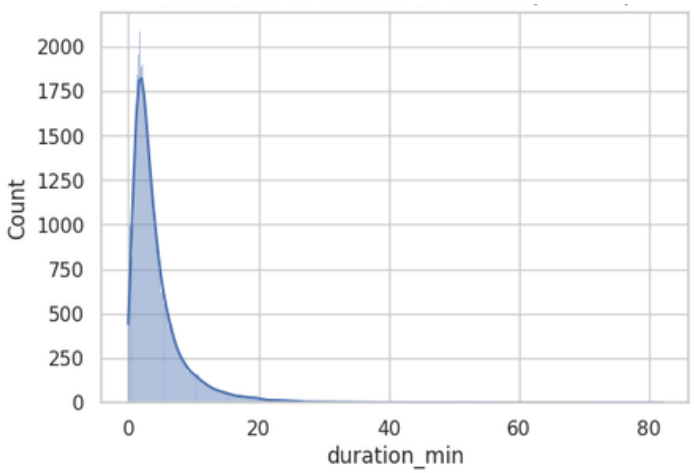
Graphes issus du Feature Engineering

Répartition par âge



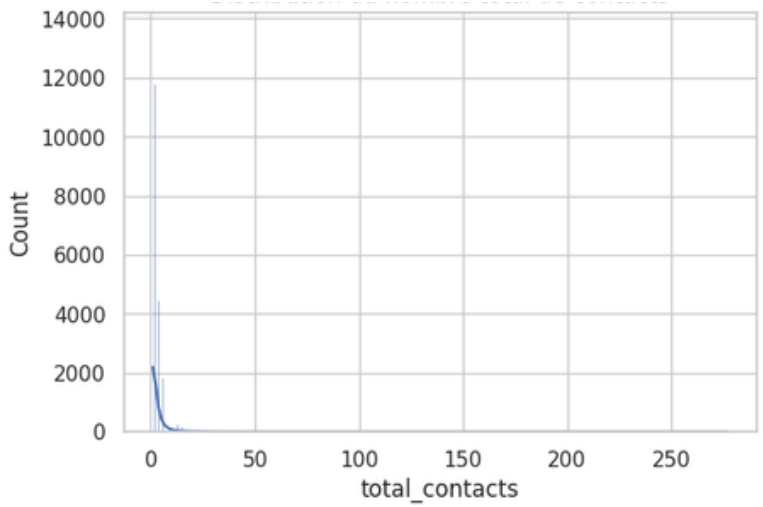
La segmentation en groupes d'âge montre une dominance claire des adultes, suivis des seniors. Cette représentation facilite l'interprétation métier et permet d'adapter les stratégies commerciales selon les segments démographiques les plus rentables.

Répartition durée contact :



La transformation de la durée en minutes améliore la lisibilité du phénomène observé. La dominance des appels courts confirme que l'efficacité commerciale repose sur un faible nombre d'interactions longues et qualitatives.

Répartition total de contact:



La majorité des clients est contactée peu de fois, ce qui traduit une stratégie de ciblage relativement sélective. Les clients très sollicités représentent une minorité, ce qui limite les coûts opérationnels tout en concentrant les efforts sur les profils jugés prometteurs.

4. Résultat de la régression logistique et random forest

4.1 Regression logistique

	precision	recall	f1-score	support
0	0.92	0.97	0.95	9981
1	0.64	0.34	0.45	1322
accuracy			0.90	11303
macro avg	0.78	0.66	0.70	11303
weighted avg	0.89	0.90	0.89	11303

Matrice de confusion :

```
[[9728  253]
 [ 868  454]]
```

La régression logistique est utilisée comme modèle de référence en raison de sa simplicité et de sa forte interprétabilité. Elle permet d'estimer la probabilité de souscription à un dépôt à terme en fonction des caractéristiques du client, en supposant une relation linéaire entre les variables explicatives et le logarithme des chances de souscription.

Les résultats montrent que ce modèle fournit des performances correctes, mais reste limité dans sa capacité à capturer des relations complexes et non linéaires entre les variables. En particulier, la détection des clients réellement susceptibles de souscrire peut être insuffisante, ce qui est problématique dans un contexte bancaire où le coût d'un faux négatif est élevé. Ainsi, la régression logistique constitue une base de comparaison pertinente, mais non optimale pour ce problème.

4.2 Random forest

```
              precision    recall  f1-score   support

           0       0.93      0.97      0.95      9981
           1       0.65      0.43      0.52      1322

    accuracy                           0.91     11303
   macro avg       0.79      0.70      0.73     11303
weighted avg       0.90      0.91      0.90     11303

Matrice de confusion :
[[9680  301]
 [ 756  566]]
```

Le Random Forest est un modèle ensembliste basé sur l'agrégation de plusieurs arbres de décision, ce qui lui permet de capturer efficacement les interactions complexes et les relations non linéaires présentes dans les données. Il est également plus robuste face aux valeurs extrêmes et à l'hétérogénéité des profils clients.

Les performances obtenues avec le Random Forest sont globalement supérieures à celles de la régression logistique, notamment en termes de rappel et de précision sur la classe des clients souscripteurs. Cette amélioration est particulièrement importante dans le cadre du ciblage marketing, car elle permet de réduire les faux négatifs et d'optimiser l'efficacité des campagnes. Le Random Forest apparaît ainsi comme le modèle le plus adapté pour ce cas d'usage.

4.3 Analyse comparative

En comparaison, la régression logistique offre une bonne lisibilité des résultats mais montre des limites face à la complexité des données, tandis que le Random Forest privilégie la performance prédictive au détriment de l'interprétabilité. Le choix final du modèle dépend donc d'un compromis entre compréhension métier et efficacité opérationnelle.

Conclusion

Cette étude montre que l'exploitation des données clients permet d'améliorer significativement l'efficacité des campagnes de télé-marketing bancaire. L'analyse met en évidence que tous les clients ne présentent pas le même potentiel de souscription et que certaines caractéristiques, notamment liées aux interactions commerciales, jouent un rôle déterminant dans la décision finale.

La comparaison des modèles révèle que l'utilisation d'outils d'intelligence artificielle avancés, comme le Random Forest, permet de mieux identifier les clients à fort potentiel par rapport à des approches classiques. Cela se traduit concrètement par une réduction des appels inutiles, une meilleure allocation des ressources commerciales et une augmentation du taux de conversion des campagnes.

D'un point de vue décisionnel, ces résultats confirment l'intérêt d'intégrer des modèles prédictifs dans le processus de ciblage afin de soutenir les équipes commerciales et d'orienter les actions vers les clients les plus susceptibles de générer de la valeur.

Pistes d'amélioration

- **Ciblage plus fin des clients :** Prioriser les clients avec une forte probabilité de souscription afin d'optimiser le temps des centres d'appels.
- **Réduction des coûts opérationnels :** Limiter les contacts auprès des clients à faible potentiel pour diminuer les dépenses liées aux campagnes.
- **Intégration des coûts d'erreur :** Ajuster les décisions en tenant compte du coût plus élevé des clients intéressés non contactés.
- **Pilotage en temps réel :** Utiliser les prédictions pour ajuster les campagnes au fil de leur déroulement.
- **Aide à la décision commerciale :** Fournir aux équipes des scores de probabilité simples à interpréter plutôt que des résultats techniques.
- **Amélioration continue :** Mettre à jour régulièrement les modèles afin de s'adapter à l'évolution du comportement des clients.

En conclusion, l'adoption d'un système de prédiction basé sur le machine learning constitue un levier stratégique de performance, permettant à la banque de concilier efficacité commerciale, maîtrise des coûts et amélioration de la relation client.