**Model : https://appbank.streamlit.app/**

**Informations principales extraites de bank\_marketing.metadata**

|  |  |
| --- | --- |
| **Élément** | **Détail** |
| Nom | Bank Marketing |
| Domaine | Business |
| Tâche principale | Classification |
| Type de données | Multivariées (Catégorielles, Numériques) |
| Nombre d’instances | 45 211 |
| Nombre de variables | 16 features + 1 cible (y) |
| Valeurs manquantes | Oui (NaN) |
| Année de création | 2014 |
| Dernière mise à jour | 18 août 2023 |
| Lien UCI | Bank Marketing @ UCI: https://archive.ics.uci.edu/dataset/222/bank+marketing |
| Objectif | Prédire si un client souscrira à un dépôt à terme |
| Référence académique | Moro et al., 2014: https://www.semanticscholar.org/paper/cab86052882d126d43f72108c6cb41b295cc8a9e |

**Structure des données (features et target)**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| age | Integer | Âge du client |
| job | Categorical | Type d’emploi |
| marital | Categorical | Statut marital |
| education | Categorical | Niveau d'éducation |
| default | Binary | Crédit en défaut ? |
| balance | Integer | Solde annuel moyen (en €) |
| housing | Binary | Prêt immobilier ? |
| loan | Binary | Prêt personnel ? |
| contact | Categorical | Type de contact |
| day\_of\_week | Date | Jour du dernier contact |
| month | Date | Mois du dernier contact |
| duration | Integer | Durée du contact (en secondes) |
| campaign | Integer | Nombre de contacts durant cette campagne |
| pdays | Integer | Jours depuis le dernier contact précédent |
| previous | Integer | Nombre de contacts avant cette campagne |
| poutcome | Categorical | Résultat de la campagne précédente |
| y (target) | Binary | Le client a-t-il souscrit un dépôt à terme ? |

**Analyse du Jeu de données**

Le jeu de données est fortement déséquilibré :  
- 88.3% de clients ont refusé de souscrire (no)  
- 11.7% ont accepté l’offre (yes)

**Statistiques descriptives : Conclusion**

- balance contient des valeurs négatives très élevées (jusqu’à -8019€). Il est recommandé de :  
 - Vérifier s’il s’agit de découverts bancaires légitimes ou de valeurs aberrantes.  
 - Appliquer une transformation ou exclure les valeurs extrêmes si elles biaisent l’analyse.  
- Plusieurs variables catégorielles comme job, education, contact, poutcome présentent des valeurs manquantes ou codées comme "unknown".  
 - Les imputer avec la modalité la plus fréquente ou les traiter séparément.  
 - Les encoder correctement pour les modèles (OneHot, Label Encoding…).  
- La variable cible y est fortement déséquilibrée (~88% de "no") :  
 - Il faudra adapter les méthodes de modélisation et utiliser des métriques robustes (F1-score, AUC).  
 - Envisager des techniques de rééquilibrage comme SMOTE ou la pondération des classes.  
- Les variables comme pdays, campaign, duration sont très asymétriques :  
 - Envisager des transformations log ou du binning pour améliorer l’interprétabilité et la performance des modèles.

**Le dataset est globalement cohérent mais nécessite :** - Un nettoyage,  
 - Une standardisation des formats,  
 - Et une vérification approfondie des extrêmes et valeurs manquantes.

**Analyse des valeurs manquantes**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Valeurs manquantes** | **Pourcentage (sur 45211)** | **Commentaire** |
| job | 288 | ~0.64% | Faible, imputable facilement |
| education | 1857 | ~4.1% | Modéré, à traiter avec soin |
| contact | 13020 | ~28.8% | ⚠️ Élevé, nécessite analyse approfondie |
| poutcome | 36959 | ~81.8% | ⚠️⚠️ Très élevé, souvent dû à l'absence de campagnes précédentes |
| Autres | 0 | 0% | Aucun traitement nécessaire |

**Recommandations**

job :  
- Faible taux de valeurs manquantes → remplacer par la valeur la plus fréquente (mode) ou ajouter une catégorie "unknown".  
  
education :  
- 4% de valeurs manquantes → imputation par la modalité dominante ou création d’une classe spécifique.  
  
contact :  
- Taux important (~29%) :  
 - Option 1 : remplacer les NaN par "unknown" si cela représente des non-contactés.  
 - Option 2 : examiner l’impact de la suppression sur le volume total.  
  
poutcome :  
- Très forte proportion de NaN (82%) liée à l’absence de contact passé :  
 - Remplacer par "non\_contacté" ou "unknown" pour préserver l’information implicite.  
 - Supprimer la variable si elle s’avère inutile à la modélisation.

**Nettoyage des données**

**Analyse du nettoyage des données**

Valeurs manquantes restantes : 0  
Toutes les colonnes ont été nettoyées avec succès :  
- Les imputations ont été appliquées correctement sur job, education, contact, et poutcome.  
- Aucune ligne supprimée inutilement.  
 Les données sont désormais complètes.  
  
Nombre de doublons : 0  
Aucun doublon détecté :  
- Le jeu de données ne contenait pas de lignes identiques.  
- Aucune suppression de lignes nécessaire.  
Données uniques et prêtes pour l'analyse.  
  
Dimensions finales :  
- X\_cleaned : (45211, 16)  
- y\_cleaned : (45211, 1)

**Conclusion**

Le nettoyage a été réalisé avec succès, sans perte d'information.

**Colonnes catégorielles encodées**

|  |  |
| --- | --- |
| **N°** | **Colonne** |
| 1 | job |
| 2 | marital |
| 3 | education |
| 4 | default |
| 5 | housing |
| 6 | loan |
| 7 | contact |
| 8 | month |
| 9 | poutcome |

**Dimensions après encodage**

- Dimensions après encodage : (45211, 40)  
 Lignes : inchangées (45 211 clients)  
 Colonnes : passées de 16 à 40, ce qui est attendu :  
- 7 colonnes numériques inchangées (ex : age, balance, etc.)  
- 9 colonnes catégorielles transformées en 33 colonnes binaires  
 (grâce à drop\_first=True qui évite la redondance multicolinéaire)

**Analyse descriptive des facteurs influençant la souscription à un dépôt à terme**

**Variables les plus influentes positivement**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Corrélation** | **Analyse** |
| duration | +0.39 | Plus l'appel téléphonique est long, plus la probabilité que le client souscrive augmente. |
| poutcome\_success | +0.31 | Historique relationnel positif : indicateur fort de succès. |
| month\_mar, month\_oct, month\_sep | ~+0.12–0.13 | Ces mois montrent un meilleur taux de conversion. |

**Variables les plus influentes négativement**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Corrélation** | **Analyse** |
| poutcome\_non\_contacté | -0.17 | Les clients n’ayant jamais été contactés dans une campagne précédente sont moins enclins à accepter l’offre, faute d'engagement préalable. |
| contact\_unknown | -0.15 | L’absence de canal de communication clair est défavorable. Cela peut indiquer un manque de personnalisation ou un ciblage inefficace. |
| month\_may | -0.10 | Ce mois est associé à un taux de conversion plus bas – potentiellement dû à une saturation des sollicitations. |
| campaign | -0.073 | Un nombre élevé de contacts pendant la campagne est associé à une diminution du taux de souscription. |
| job\_blue-collar | -0.071 | Ce segment montre une réticence plus marquée à souscrire. |

Les variables démographiques classiques comme age, balance, education, ou marital ont un impact limité sur la décision (corrélations faibles).

Le type de canal de communication est un levier stratégique : les appels mobiles montrent de meilleurs résultats.

La variable duration étant déterminée après l’appel, elle ne peut pas être utilisée pour prédire en amont — mais elle explique très bien le comportement final.

**Analyse descriptive des variables numériques**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Distribution observée** | **Valeurs aberrantes** | **Interprétation** |
| age | Asymétrique droite, pic autour de 30–40 ans | Quelques outliers > 70 ans | La majorité des clients sont en âge actif. |
| balance | Fortement asymétrique droite | De nombreux outliers, jusqu’à 100 000 € | Hétérogénéité financière. Transformation recommandée. |
| day\_of\_week | Relativement uniforme | Aucune | Bonne répartition des contacts. |
| duration | Très asymétrique droite, pic < 200s | Outliers nombreux > 1000s | Variable très informative mais dépendante de l’appel. |
| campaign | Concentrée sur 1 à 3 | Outliers jusqu’à 63 | Relances excessives peu efficaces. |
| pdays | Valeur dominante -1 | Extrêmes > 400 | Utile pour distinguer clients « neufs » vs suivis. |
| previous | Concentrée sur 0–1 | Outliers rares > 100 | Peut être transformée en variable binaire. |

**Interprétation stratégique**

Les distributions fournissent une base tangible pour justifier le ciblage :

- Cibler les clients avec duration > 200s ou poutcome\_success = yes est objectivement défendable.

- Les extrêmes de balance doivent être traités pour ne pas biaiser les décisions.

- Les campagnes sont plus efficaces avec peu de relances.

- Les contacts courts ou non qualifiés ont peu de valeur.

- Les graphiques permettent de fonder les décisions sur des critères objectifs.

- Les asymétries identifiées permettent de transformer les variables pour renforcer les modèles.

- L’analyse met en lumière des profils extrêmes ou spécifiques à cibler différemment.

**Recommandations pratiques**

1. Transformer balance, duration, campaign, pdays pour atténuer les effets des outliers.

2. Isoler et traiter les valeurs extrêmes.

**Analyse descriptive des variables catégorielles**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Observation principale** | **Interprétation stratégique** |
| job | `blue-collar`, `management`, `technician` dominent | Représentent les cœurs de clientèle. |
| marital | `married` majoritaire | Clients probablement plus stables mais prudents. |
| education | `secondary` dominant | Clientèle intermédiaire en pouvoir d’achat. |
| default | Presque tous 'no' | Population globalement solvable. |
| housing | Répartition équilibrée | Influence potentielle sur capacité d’investissement. |
| loan | Majorité 'no' | Plus disposés à investir. |
| contact | `cellular` dominant | Canal à privilégier. |
| month | Concentration en `may` | Saturation à compenser. |
| poutcome | `non\_contacté` majoritaire | Potentiel d’exploration important. |

- Utiliser job, education, loan pour construire des profils prioritaires.

- Prioriser le canal cellular.

- Éviter les contacts avec unknown.

- Rééquilibrer les campagnes sur l’année.

- Les clients non\_contacté sont une opportunité commerciale importante.

**Recommandations**

1. Cibler selon le triptyque job, loan, contact.

2. Équilibrer les campagnes.

3. Prioriser les contacts neufs.

4. Adapter le message marketing selon education et marital.

**Variables clés pour guider le ciblage client**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Pourquoi elle est utile** | **Comment elle oriente la décision** |
| duration | Corrélée à la souscription (+0.39) | Identifie les clients engagés après appel. |
| poutcome | Historique de campagnes | Utiliser anciens succès. |
| contact | Type de canal | Cibler les canaux efficaces. |
| month | Mois de l'appel | Cibler les bonnes périodes. |
| campaign | Nombre de contacts | Éviter le harcèlement. |
| pdays | Temps écoulé depuis dernier contact | Différencier prospects récents. |
| job | Profil socio-pro | Certains profils plus réceptifs. |
| balance | Solde annuel moyen | Peut refléter capacité d’investissement. |

**Pourquoi duration est la variable la plus utile**

- Corrélation élevée avec y : ≈ +0.39

- Indicateur d’intérêt : plus l’appel est long, plus la souscription est probable.

- Discriminant fort : peu de souscriptions si l’appel est court.

- Exploitable : agir en temps réel ou relancer après appel.

- Interprétable et actionnable.

**Statistiques descriptives de la variable duration**

|  |  |  |
| --- | --- | --- |
| **Statistique** | **Valeur** | **Interprétation** |
| count | 45,211 | Nombre total d’appels. |
| mean | 258.16 sec | Durée moyenne ≈ 4 min 18 sec. |
| std | 257.53 sec | Grande variabilité. |
| min | 0.0 sec | Appels échoués ou très courts. |
| 25% | 103 sec | Moins de 1 min 43 sec. |
| 50% | 180 sec | Médiane ≈ 3 min. |
| 75% | 319 sec | Plus de 5 min 19 sec. |
| max | 4918 sec | Outlier : plus de 81 minutes. |

**Analyse de la dispersion et coefficient de variation de duration**

- Écart-type : 257.53 sec

- Amplitude : 4918 sec

- Coefficient de variation : 99.75 %

- Très forte dispersion → variable hétérogène.

**Interprétation de l’asymétrie (skew) de duration**

|  |  |
| --- | --- |
| **Valeur de skew** | **Interprétation** |
| ≈ 0 | Distribution symétrique |
| > 0 | Asymétrie à droite (valeurs extrêmes hautes) |
| < 0 | Asymétrie à gauche |

- Valeur mesurée : 3.14 → asymétrie à droite.

- Beaucoup d’appels courts, peu d’appels très longs.

- Présence de valeurs extrêmes.

**Interprétation de la kurtosis de duration**

|  |  |
| --- | --- |
| **Kurtosis** | **Interprétation** |
| ≈ 0 | Distribution normale |
| > 0 | Distribution pointue (leptokurtique) |
| < 0 | Distribution aplatie |

- Kurtosis mesurée : 18.15 → très leptokurtique.

- Forte concentration centrale + outliers extrêmes.

Cela signifie que la distribution de `duration` est fortement leptokurtique.

La plupart des appels ont des durées proches de la moyenne (forte concentration centrale).

Mais il existe aussi beaucoup d’outliers extrêmes (appels très longs).

Cela confirme ce que montraient déjà :

* - L’écart-type élevé
* - L’asymétrie (skew = 3.14)

**Ce qu'on observe du boxplot de duration**

La boîte (entre Q1 et Q3)

* - La boîte représente les 50 % centraux des durées d’appel.
* - Elle est située dans une plage basse : environ entre 100 et 320 secondes.
* - Cela montre que la majorité des appels sont assez courts.

La médiane

* - La médiane est décalée vers la gauche.
* - Cela indique une asymétrie à droite (skew positive), confirmée par skew = 3.14.

Les valeurs aberrantes (outliers)

* - De nombreux points en dehors des moustaches sont visibles à droite.
* - Ces outliers représentent des appels exceptionnellement longs.
* - Bien que rares, ils étirent la distribution et peuvent fausser la moyenne.

**Analyse univariée – Variable duration**

1. Histogramme

* - Distribution très asymétrique à droite.
* - Majorité des appels durent moins de 500 secondes.
* - Forte concentration de valeurs faibles.

2. Boîte à moustaches

* - Médiane très basse.
* - Présence de nombreux outliers à droite.

3. Courbe de densité

* - Densité concentrée autour de petites valeurs.
* - Chute rapide après le pic.

4. Camembert des classes de durée

* - Classes dominantes : "Court" (30.9%) et "Moyen" (27.6%).
* - Très peu d'appels sont classés "Long" (9.0%) ou "Très long" (8.4%).

5. Courbe de Lorenz

* - Courbe loin de la diagonale → forte concentration des durées sur une minorité.
* - Inégalité importante.

Interprétation générale

* - duration est extrêmement asymétrique, avec de nombreux outliers.
* - Variable très informative mais nécessite une transformation logarithmique.
* - La forme actuelle nuit à une modélisation directe.

**Analyse bivariée : Variables numériques vs y (souscription)**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Observation visuelle** | **Interprétation stratégique** |
| age | Légère différence, médiane un peu plus haute pour yes | Clients plus âgés légèrement plus enclins à souscrire |
| balance | Médiane plus élevée pour yes, grande variabilité | Solde bancaire plus élevé = plus de souscriptions |
| day\_of\_week | Distributions quasi identiques | Aucun impact significatif |
| duration | Médiane plus élevée pour yes | Plus l’appel est long, plus la probabilité de souscription augmente |
| campaign | Médiane plus basse pour yes | Moins sollicités = plus enclins à souscrire |
| pdays | Moyenne et médiane plus basses pour no | Recontact rapide favorise la souscription |
| previous | Légère élévation chez yes | Historique de contact légèrement favorable |

**Variables les plus utiles pour prédire y**

1. duration → très bon discriminant entre yes et no

2. balance → lien direct avec capacité d’épargne

3. campaign → effet d’usure

4. pdays, previous → efficacité du suivi

**Variables peu utiles**

* - day\_of\_week → aucun pouvoir prédictif visible

**Recommandations**

* - Inclure : duration, balance, campaign, pdays, previous
* - Exclure : day\_of\_week
* - Prétraiter duration (log, binning)
* - Segmenter selon durée d’appel et historique

**Analyse Variables catégorielles vs y (souscription)**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Observation visuelle** | **Interprétation stratégique** |
| job | Les taux de souscription varient selon les métiers. retired, student, management ont une proportion plus élevée de yes. | Ces segments sont à cibler en priorité. Les blue-collar ont un taux de conversion faible. |
| marital | Tous les statuts ont plus de no que yes. Légère hausse de yes chez les single. | Le statut marital n'est pas très discriminant, mais les célibataires sont légèrement plus ouverts. |
| education | Les niveaux tertiary et secondary génèrent plus de souscriptions que primary. | Un niveau d’éducation plus élevé est associé à un taux de souscription supérieur. |
| default | Très peu de clients avec default = yes souscrivent. | Un historique de défaut est un frein majeur. |
| housing | Les clients sans prêt immobilier (no) ont plus de souscriptions que ceux avec (yes). | Avoir un prêt immobilier pourrait réduire la capacité d’épargne. |
| loan | Les clients sans prêt personnel (no) souscrivent davantage. | Un prêt personnel en cours est un frein à la souscription. |
| contact | cellular domine en volume et conversions. telephone et unknown génèrent peu de yes. | Le canal de contact est crucial : privilégier le mobile (cellular). |
| month | may a le plus grand nombre de contacts mais très peu de yes. mar, oct, dec ont de meilleurs taux. | Le mois de contact a un fort impact. Éviter may. |
| poutcome | success est fortement corrélé à des souscriptions. non\_contacted génère peu de yes. | L’historique de campagne est hautement prédictif. |

**Variables catégorielles les plus discriminantes**

* - poutcome : succès passé = forte probabilité de re-souscription
* - contact : le canal cellular est le plus performant
* - month : le timing des campagnes influence les résultats
* - loan et default : indicateurs de frein financier
* - job : certains profils sont bien plus réceptifs

**Recommandations**

**À cibler activement :**

* - Profils retired, student, management
* - Clients sans prêt
* - Mois : mar, oct, dec et - Clients avec poutcome = success

**À éviter ou filtrer :**

* - Campagnes massives en may
* - Clients avec default = yes
* - Canaux unknown ou telephone
* - Un sous-ensemble ciblé est plus pertinent pour préparer la modélisation

**Tests de normalité et recommandations**

* - Utilisation d'un graphique de densité pour visualiser la distribution.
* - Test de normalité Shapiro.
* - Test Jarque-Bera : normalité + skew + kurtosis.

**Interprétation des courbes de densité et tests de normalité**

1. duration :

* - Fortement asymétrique à droite.
* - Beaucoup d’appels courts, quelques-uns très longs.
* - Recommandation : transformation logarithmique.

2. balance :

* - Très étalée, longue queue à droite.
* - Recommandation : winsorizing ou log1p.

3. campaign :

* - Pic sur 1, chute rapide ensuite.
* - Recommandation : binning (1 contact / 2–3 / 4+).

4. pdays :

* - Pic fort sur -1.
* - Recommandation : variable binaire (0 = non contacté, 1 = recontacté).

5. previous :

* - Concentration sur 0.
* - Recommandation : binaire (aucun contact passé / au moins un).

**Conclusion**

Les distributions montrent :

* - Asymétries
* - Présence d’outliers
* - Aucune ne suit une loi normale

Recommandation : transformer ou regrouper pour améliorer la performance et stabilité du modèle.

**Corrélation pour identifier les variables fortement liées**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Coefficient** | **p-value** | **Interprétation** |
| duration | +0.342 | 0.0000 | Relation monotone modérée : plus l’appel est long, plus la probabilité de souscription augmente. |
| balance | +0.100 | 0.0000 | Corrélation faible mais significative : les clients avec un solde plus élevé souscrivent légèrement plus. |
| campaign | –0.084 | 0.0000 | Corrélation faible et négative : plus un client est relancé, moins il est susceptible de souscrire. |
| pdays | +0.154 | 0.0000 | Corrélation faible à modérée : les clients recontactés sont un peu plus réactifs. |
| previous | +0.169 | 0.0000 | Corrélation faible à modérée : les clients ayant un historique de contact sont plus enclins à souscrire. |

**Corrélation avec la variable cible y**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Corrélation** | **Interprétation** |
| duration | +0.34 | Corrélation modérée. Plus l’appel est long, plus la probabilité de souscription augmente. |
| previous | +0.17 | Corrélation faible à modérée. Les clients ayant déjà été contactés sont plus enclins à souscrire. |
| pdays | +0.15 | Corrélation faible à modérée. Être recontacté (et le délai) influence légèrement la conversion. |
| balance | +0.10 | Corrélation faible mais significative. Les clients avec un solde plus élevé ont un peu plus de chances de souscrire. |
| campaign | –0.08 | Corrélation faible et négative. Trop de sollicitations réduisent légèrement la probabilité de souscription. |

**Corrélation entre les variables explicatives**

|  |  |  |
| --- | --- | --- |
| **Variables** | **Corrélation** | **Interprétation** |
| pdays & previous | +0.99 | Corrélation presque parfaite. Ne pas utiliser ensemble dans un modèle. |
| balance & autres | ±0.10 | Peu de corrélations. balance est indépendante. |
| duration & autres | Très faible | duration est indépendante, utile pour la modélisation. |

**Conclusion sur les corrélations**

* - duration est la variable la plus informative.
* - pdays et previous : à ne pas utiliser ensemble à cause de leur redondance.
* - campaign a un effet négatif modéré, à considérer avec prudence.
* - balance reste une variable faible mais complémentaire.

**Test d’indépendance avec ANOVA**

**Analyse des durées moyennes d’appel selon les variables qualitatives**

1. job : unemployed, retired, self-employed → appels longs. housemaid, student, admin. → appels courts.

2. education : peu de différence entre primary, secondary, tertiary.

3. contact : cellular > telephone.

4. month : dec, apr, sep = longs ; aug, jun, mar = courts.

5. poutcome : success = appels plus longs.

Conclusion : job, poutcome, month influencent fortement la durée. contact a un impact modéré. education est peu différenciante.

**Résultats du test ANOVA – Influence des variables qualitatives sur duration**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable qualitative** | **p-value** | **Influence significative ?** | **Interprétation** |
| job | 0.0000 | Oui | Certaines professions génèrent des appels plus longs. |
| education | 0.7386 | Non | Le niveau d’éducation n’a pas d’effet clair sur la durée des appels. |
| contact | 0.0000 | Oui | Le type de canal influence la durée. |
| month | 0.0000 | Oui | Le mois de l’appel affecte la durée. |
| poutcome | 0.0000 | Oui | Le résultat d’une campagne précédente impacte la durée. |

**Résultats du test ANOVA – Influence des variables qualitatives sur balance**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable qualitative** | **p-value** | **Influence significative ?** | **Interprétation** |
| job | 0.0000 | Oui | Le solde bancaire varie selon le métier. |
| education | 0.0000 | Oui | Le niveau d’éducation influence le solde. |
| contact | 0.0000 | Oui | Le canal de contact semble lié à la capacité financière. |
| month | 0.0000 | Oui | Le mois de contact est lié aux variations de solde. |
| poutcome | 0.0000 | Oui | Le résultat de la campagne précédente est lié au solde. |

Ces résultats peuvent aider à segmenter les clients selon leur pouvoir d’épargne et adapter les offres en fonction du profil.

**Analyse des variables qualitatives vs 'y'**

Cette analyse permet de :

* - Identifier quelles catégories sont les plus susceptibles de souscrire.
* - Repérer les profils à éviter ou à prioriser dans les campagnes marketing.
* - Créer des taux de souscription par modalité.

**Répartition des modalités selon y – Job**

|  |  |  |
| --- | --- | --- |
| **Métier** | **% Yes** | **Interprétation** |
| student | 28.68% | Très réceptif |
| retired | 22.79% | Réactif |
| unemployed | 15.50% | Réactif |
| management | 13.76% | Moyennement réceptif |
| blue-collar | 7.41% | Moins réceptif |

**Répartition selon Education**

|  |  |  |
| --- | --- | --- |
| **Niveau d'étude** | **% Yes** | **Interprétation** |
| tertiary | 15.01% | Plus réceptif |
| secondary | 10.78% | Moyennement réceptif |
| primary | 8.63% | Moins réceptif |

**Répartition selon Contact**

|  |  |  |
| --- | --- | --- |
| **Canal** | **% Yes** | **Interprétation** |
| cellular | 14.92% | Canal le plus efficace |
| telephone | 13.42% | Canal secondaire |
| unknown | 4.07% | Canal peu efficace |
|  |  |  |

**Répartition selon Month**

|  |  |  |
| --- | --- | --- |
| **Mois** | **% Yes** | **Interprétation** |
| mar | 51.99% | Mois le plus performant |
| dec | 46.73% | Très bon mois |
| sep | 46.46% | Très bon mois |
| may | 6.72% | Mois à éviter |

**Répartition selon Poutcome**

|  |  |  |
| --- | --- | --- |
| **Résultat précédent** | **% Yes** | **Interprétation** |
| success | 64.73% | Très fort impact |
| other | 16.68% | Impact modéré |
| failure | 12.61% | Faible impact |
| non\_contacté | 9.16% | Très faible impact |

**Recommandations**

* - Cibler en priorité les profils : student, retired, unemployed
* - Privilégier le canal cellular
* - Concentrer les campagnes sur les mois : mar, dec, sept
* - Recontacter les clients ayant eu un résultat success
* - Éviter ou minimiser les campagnes en mai, et sur des profils moins réceptifs comme blue-collar et primary éducation

**Test Khi² – Résultats**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **p-value** | **Conclusion** | **Interprétation** |
| job | 0.0000 | Rejet de H₀ | Le métier influence significativement la souscription. |
| education | 0.0000 | Rejet de H₀ | Le niveau d’éducation est lié au comportement de réponse. |
| contact | 0.0000 | Rejet de H₀ | Le canal de contact a un effet sur la souscription. |
| month | 0.0000 | Rejet de H₀ | Le mois de contact modifie la répartition de y. |
| poutcome | 0.0000 | Rejet de H₀ | Le résultat de la campagne précédente est associé à la réponse actuelle. |

Conclusion :

* - Toutes les variables testées sont significativement liées à la variable cible.
* - Leurs modalités influencent la répartition de yes et no.
* - Ces variables peuvent être recommandées pour la modélisation.

**Corrélation (Pearson) entre les variables et y**

|  |  |  |
| --- | --- | --- |
| **Variable ↔ y** | **Corrélation** | **Interprétation** |
| duration ↔ y | +0.39 | Variable la plus corrélée à la souscription |
| balance ↔ y | +0.05 | Corrélation très faible mais positive |
| campaign ↔ y | –0.07 | Corrélation faible et négative |
| pdays ↔ y | +0.10 | Corrélation faible et positive |
| previous ↔ y | +0.09 | Corrélation faible |

**Recommandations**

|  |  |
| --- | --- |
| **Action à entreprendre** | **Justification** |
| Conserver duration | Variable la plus discriminante |
| Garder pdays et previous | Information comportementale utile |
| Réduire ou transformer campaign | Effet d’usure à modéliser |
| Inclure balance | Complément contextuel économique |

**Corrélation partielle**

Objectif : Isoler l’effet propre de chaque variable explicative sur y, en neutralisant l’effet des autres.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Corrélation partielle** | **Interprétation** |
| duration | +0.394 | Corrélation forte et significative. duration reste la variable la plus explicative. |
| balance | +0.047 | Corrélation très faible mais significative. Effet direct limité. |
| campaign | -0.034 | Corrélation faible et négative. Le sur-sollicitation a un effet faible mais réel. |
| pdays | +0.074 | Corrélation faible. Recontact récent reste un facteur légèrement favorable. |
| previous | +0.055 | Corrélation faible. L’historique de contact reste légèrement explicatif. |

Conclusion : Toutes les variables conservent un effet propre significatif sur y. duration est la plus dominante.

**MANOVA**

Objectif : Mesurer l’effet simultané de variables catégorielles sur plusieurs variables numériques.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Effet global** | **Significatif ?** | **Interprétation** |
| poutcome | Très fort | Oui | Meilleur prédicteur multivarié |
| contact | Fort | Oui | Type de contact influe sur les comportements numériques |
| month | Modéré | Oui | L'effet du mois sur les indicateurs est réel |
| education | Faible mais significatif | Oui | L’éducation joue un rôle limité |
| job | Faible | Oui | L'effet du métier est marginal |

**ANCOVA**

Objectif : Étudier si une variable catégorielle influence une variable numérique en contrôlant une covariable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable dépendante** | **Catégorie testée** | **Catégorie significative ?** | **Balance significative ?** | **Interprétation** |
| duration | job | Oui | Oui | L’effet du métier et du solde bancaire est réel |
| duration | education | Non | Oui | Pas d'effet propre d'éducation après contrôle |
| duration | contact | Oui | Oui | Le canal utilisé influence la durée d’appel |
| duration | poutcome | Oui | Oui | Résultat antérieur impacte l’appel actuel |
| duration | month | Oui | Oui | La période de l’année modifie la durée des appels |

**Médiation**

Objectif : Mesurer si l’effet de contact sur y passe partiellement par duration.

|  |  |  |  |
| --- | --- | --- | --- |
| **Chemin causal** | **Coefficient** | **Significatif** | **Interprétation** |
| contact → duration | +5.98 | Oui | Le canal utilisé influe fortement la durée |
| duration → y | +0.00049 | Oui | Une durée plus longue augmente la souscription |
| Effet total | +0.0531 | Oui | Globalement, contact est bénéfique |
| Effet direct | +0.0502 | Oui | La majeure partie de l’effet est direct |
| Effet indirect | +0.0029 | Oui | Une part passe via duration |

**Recommandation**

| **Élément à prioriser** | **Raison** |
| --- | --- |
| duration | Effet fort, indépendant, robuste |
| contact, poutcome | Forte capacité explicative directe et indirecte |
| balance | Covariable essentielle à ajuster |
| education | Faible pouvoir explicatif → à reconsidérer |
| Approche multivariée | Utile pour saisir l’interaction des comportements |

**Analyse de la multicolinéarité (VIF)**

| **Variable** | **VIF** | **Interprétation** |
| --- | --- | --- |
| duration | 1.01 | Aucune colinéarité détectée |
| balance | 1.00 | Variable totalement indépendante |
| campaign | 1.02 | Corrélation négligeable |
| pdays | 1.27 | Corrélation modérée acceptable |
| previous | 1.26 | Corrélation modérée acceptable |

**Interprétation** :  
Tous les VIF sont largement inférieurs au seuil critique de 5. Il n’y a donc **aucune multicolinéarité problématique** entre les variables numériques utilisées. Cela garantit :

* **Stabilité des coefficients** dans les modèles linéaires.
* **Pas besoin de réduire le nombre de variables** ou d'appliquer une PCA pour cette raison.

**PCA (Analyse en Composantes Principales)**

| **Composante** | **Variance expliquée (%)** |
| --- | --- |
| PC1 | ≈ 29 % |
| PC2 | ≈ 22 % |
| PC3 | ≈ 20 % |

**Total PC1 à PC3** ≈ **71 %**

**Interprétation** :  
Les **3 premières composantes** capturent une **grande majorité de la variance** → la PCA permet une **réduction de dimension efficace** sans perte majeure d'information.

**Graphique PCA (PC1 vs PC2)** :

* Séparation visible entre les clients ayant souscrit (yes) et non (no) → bon potentiel de **visualisation ou segmentation**.
* Des **groupes homogènes** et quelques **outliers** peuvent être identifiés → base solide pour un **clustering**.

**Clustering K-means sur PC1 & PC2 (k=3)**

| **Cluster** | **Effectif** | **Taux de souscription (%)** | **Interprétation** |
| --- | --- | --- | --- |
| 0 | 29 064 | 4.36 % | Segment large mais peu réactif |
| 1 | 6 884 | 20.42 % | Segment moyen → bon potentiel |
| 2 | 9 263 | 28.25 % | Segment très réactif |

**Conclusion** :

* **Cluster 2** est le segment **le plus rentable** → à cibler en priorité dans les campagnes.
* **Cluster 0** est à **éviter** (fort volume mais très faible taux de conversion).
* Le clustering apporte donc **une segmentation client pertinente**.

**Transformations recommandées sur les variables**

| **Variable** | **Problème détecté** | **Transformation proposée** | **Objectif** |
| --- | --- | --- | --- |
| duration | Forte asymétrie, outliers | log1p() | Réduire les extrêmes |
| balance | Valeurs extrêmes, grande dispersion | log1p() | Stabiliser la variable |
| campaign | Distribution biaisée (pic sur 1) | log1p() | Lissage |
| pdays | Présence forte de -1 | Binarisation (0/1) | Simplification |
| previous | Majorité de zéros | Binarisation (0/1) | Mieux exploiter l'info |

**Étapes à appliquer** :

| **Étape** | **À faire ?** | **Raisons** |
| --- | --- | --- |
| Standardisation des variables continues | Oui | Normaliser les échelles |
| Encodage des variables catégorielles | Oui | Modélisation possible |
| Transformation log de duration | Recommandée | Pour lisser la distribution |
| Intégration de cluster comme variable | Oui | Segment stratégique à inclure |

**Pipeline de transformation et modélisation**

| **Composant** | **Description** | **Objectif** |
| --- | --- | --- |
| log | Imputation + log1p + standardisation (duration, balance, campaign) | Réduction des outliers, mise à l’échelle uniforme |
| bin | Imputation + binarisation (pdays, previous) | Simplification de variables discrètes, meilleure lisibilité pour le modèle |
| cat | Imputation + encodage OneHot (job, education, contact, month, poutcome) | Préparation des variables qualitatives pour un algorithme linéaire |
| cluster | Pass-through | Intégration de la segmentation client dans la modélisation |
| LogisticRegression | Modèle linéaire avec régularisation L2 | Prédiction probabiliste, interprétable et robuste |

**Points forts du pipeline :**

* **Cohérent** : toutes les variables sont traitées selon leur nature.
* **Optimisé pour la régression logistique** (standardisation, codage, transformations).
* **Reproductible** : toutes les étapes sont bien séquencées, donc faciles à réutiliser ou améliorer.
* **Inclusion de la variable cluster** : bon ajout stratégique pour capter la structure client.

**Évaluation du modèle**

**Matrice de confusion**

| **Réalité / Prédiction** | **Prédit : no** | **Prédit : yes** |
| --- | --- | --- |
| **Réel : no** | 38 949 | 973 |
| **Réel : yes** | 3 442 | 1 847 |

**Interprétations :**

* **Accuracy ≈ 90.1 %** : très bon score global.
* **Précision pour yes = 65 %** : quand le modèle prédit yes, il a souvent raison.
* **Rappel pour yes = 35 %** : il **ne détecte que 1 client sur 3** qui aurait réellement souscrit → trop conservateur.
* **Excellente performance pour no (rappel ≈ 98 %)** : très utile pour **réduire les campagnes inutiles**.

**Rapport de classification (principaux indicateurs)**

| **Indicateur** | **Valeur** | **Interprétation** |
| --- | --- | --- |
| Accuracy | 90.2 % | Solide performance globale |
| AUC (ROC) | 0.9045 | Excellente capacité discriminante |
| Precision yes | 65 % | Prédictions de yes plutôt fiables |
| Recall yes | 35 % | Faible couverture des souscriptions réelles |
| F1-score yes | 46 % | Moyen : compromis entre précision et rappel |

**Recommandations :**

* Le **modèle est bien calibré**, mais trop conservateur sur la classe yes.
* Pour **mieux capter les clients susceptibles de souscrire**, vous pouvez :
  + **Abaisser le seuil de décision** (ex. proba > 0.3)
  + Utiliser un modèle **moins biaisé vers la classe majoritaire** :
    - Random Forest ou XGBoost avec gestion de l'**imbalance**
    - LogisticRegression(class\_weight='balanced')

**Courbe ROC et AUC**

| **Mesure** | **Valeur** |
| --- | --- |
| AUC | 0.9045 |

**Interprétation :**

* **AUC > 0.9** : le modèle a **excellente capacité à classer les clients correctement**, quelle que soit la probabilité seuil.
* La courbe montre une **très bonne sensibilité dès les faibles taux de faux positifs** → utile pour des campagnes bien ciblées.

**Conclusion opérationnelle**

| **Atout du modèle** | **Limite principale** | **Solution recommandée** |
| --- | --- | --- |
| Excellente discrimination globale (AUC) | Faible détection des yes (rappel 35 %) | Ajuster le seuil ou changer d’algorithme |
| Pipeline complet et bien structuré | Modèle conservateur (biais vers no) | Pondération des classes ou modèle plus flexible |
| Bonne interprétabilité (logistique + cluster) | Faible sensibilité sans traitement de déséquilibre | Ajouter stratégie sur échantillon yes |