

# Rapport

---

## Profil client et solvabilité financière

Chaimaa Bella, Mélanie Poinsignon, Mody Hady Barry, Suzanne Wernberg-Møller

L'objectif de ce rapport est d'aider la banque à optimiser le tri des demandes de crédit à la consommation en distinguant les bons et mauvais clients. Le but est, grâce au jeu de données fourni, de déterminer au mieux la solvabilité de chacun.

## TABLE DES MATIÈRES

<b>Rapport — Profil client et solvabilité financière</b>	<b>1</b>
1. Analyse de la structure des données	3
1.1 Dimensions et composition	3
1.2 Types de données	3
1.3 Données manquantes	4
1.4 Statistiques descriptives principales	4
2. Interprétation des données	5
2.1 Profil général de la clientèle	5
2.2 Produits et comportements de crédit	5
2.3 Facteurs de risque	5
2.4 Indications stratégiques potentielles	6
3. Corrélations dans les données	6
3.1 Corrélations entre variables numériques	6
3.2 Patterns observés dans les variables catégorielles	7
4. Corrélation et causalité : nuances importantes	7
5. Introduction à la capacité prédictive d'un modèle	8
5.1 Objectif du modèle	8
5.3 Choix d'un modèle de base	9
a) Méthode KNN (K plus proches voisins)	9
b) Régression logistique	9
6. Évaluation de la capacité prédictive	10
6.1 Métriques utilisées	10
6.2 Comparaison KNN vs Régression logistique	10
7. Variables explicatives clés et justification analytique	11
8. Conclusion sur le choix du modèle	11
9. Recommandations opérationnelles	12
10. Conclusion	12
<b>Annexes</b>	<b>13</b>

# 1. Analyse de la structure des données

## 1.1 Dimensions et composition

Le jeu de données contient 1 723 lignes (observations) et 14 colonnes (variables).

La variable cible est `bad_client_target`, une variable binaire indiquant la solvabilité du client :

- `1` : mauvais payeur (client ayant eu des défauts de remboursement)
- `0` : bon payeur (client sans défaut enregistré)

## 1.2 Typologie des variables

Les variables présentes et leur type :

*Numérique :*

### Variables numériques (quantitatives continues)

- `month` : mois de la demande de crédit
- `credit_amount` : montant du crédit demandé (en unités monétaires)
- `credit_term` : durée du crédit (probablement en mois)
- `age` : âge du client (en années)
- `region` : identifiant de la région (peut être interprété comme une variable catégorielle codée)
- `income` : revenu mensuel du client
- `phone_operator` : opérateur téléphonique (peut être considéré comme un indicateur socio-économique)

### Variables catégorielles (qualitatives)

- `sex` : sexe du client (H/F)
- `education` : niveau d'éducation
- `product_type` : type de produit financé par le crédit
- `family_status` : situation familiale

### Variables binaires (0 ou 1)

- `having_children_flg` : le client a-t-il des enfants ? (`1` = oui, `0` = non)
- `is_client` : le client est-il déjà client de la banque ? (`1` = oui, `0` = non)

- **bad\_client\_target** : cible de notre étude (1 = mauvais client, 0 = bon client)

## 1.3 Données manquantes

Aucune valeur manquante n'a été détectée dans l'ensemble des colonnes.

## 1.4 Statistiques descriptives principales

### Montant du crédit :

- Moyenne : 29 265
- Médiane : 21 500
- Écart-type : 27 927
- Min : 5 000
- Max : 301 000

### Âge des clients :

- Moyenne : 35,9 ans
- Médiane : 32 ans
- Min : 18
- Max : 90

### Durée du crédit :

- Moyenne : 11,5 mois
- Médiane : 12 mois
- Min : 3 mois
- Max : 36 mois

### Revenu :

- Moyenne : 32 652
- Médiane : 27 000
- Min : 1 000
- Max : 401 000

**Sexe** : 931 hommes (54%) et 792 femmes (46%)

**Éducation dominante** : "Secondary special education" (836 individus)

**Produit le plus financé** : "Cell phones" (498 cas)

**Proportion de clients avec enfants** : (42,8%)

**Statut familial le plus fréquent** : “Another” (69,7%)

**Proportion de clients actifs** (`is_client = 1`) : 60,5 %

**Taux de mauvais payeurs** (`bad_client_target = 1`) : 11,4 %

## 2. Interprétation des données

### 2.1 Profil général de la clientèle

La majorité des clients sont des adultes jeunes à moyens (25 à 45 ans), avec un revenu moyen autour de 32 000.

Les hommes représentent une légère majorité.

Le niveau d'éducation prédominant est “Secondary special education”, ce qui peut correspondre à un niveau équivalent à un diplôme professionnel ou secondaire supérieur.

La plupart des crédits sont contractés pour des montants relativement modestes (entre 13 000 et 34 000), sur une durée courte (en moyenne 12 mois).

### 2.2 Produits et comportements de crédit

Le produit le plus souvent financé est le téléphone portable, ce qui traduit une orientation vers des crédits à la consommation à faible ou moyen montant.

Les autres produits sont diversifiés (22 types recensés), ce qui témoigne d'un portefeuille large.

### 2.3 Facteurs de risque

Dans le secteur bancaire classique, pour des crédits à la consommation, le taux de défaut “acceptable” est généralement entre 3 % et 8 % selon le pays, le profil des emprunteurs et la politique de risque de l'établissement. Au-delà de 10 %, les établissements commencent à renforcer les critères d'octroi (revenu minimum, scoring, vérification de solvabilité, etc.).

Dans ce jeu de données, on identifie 11,4 % de mauvais payeurs, ce qui n'est pas négligeable et peut justifier des analyses prédictives plus fines.

On peut supposer que les mauvais payeurs sont probablement concentrés dans certaines catégories de revenu, d'âge ou de produit — ce qui mériterait une analyse segmentée.

De plus, la présence d'enfants peut influencer la capacité de remboursement (42,8 % en ont).

## 2.4 Indications stratégiques potentielles

Segmenter les offres en fonction de l'âge et du revenu permettrait de cibler les produits les plus adaptés aux profils clients.

Les clients actifs (60,5 %) représentent une base solide pour des offres de fidélisation ou de réengagement.

Les segments à risque pourraient être anticipés par un modèle prédictif basé sur les variables socio-démographiques et financières.

# 3. Corrélations dans les données

## 3.1 Corrélations entre variables numériques

L'objectif ici est d'identifier si certaines variables numériques sont **corrélées** entre elles ou avec la variable cible `bad_client_target`.

En utilisant un script python pour créer un graphique de type “heatmap” (cf. Annexe 3 et 4), nous avons procédé à une analyse de corrélation simple des variables numériques.

Les premiers résultats montrent :

- Une corrélation faible à modérée entre le montant du crédit et le revenu (les revenus plus élevés sont associés à des montants de crédit plus importants).

- Une corrélation modérée entre credit\_term et credit\_amount (plus le montant est élevé, plus la durée tend à augmenter).
- Aucune corrélation linéaire forte entre les variables numériques et bad\_client\_target, ce qui signifie que la probabilité de défaut n'est pas directement liée de manière simple à une seule variable numérique.

Cela est fréquent dans les problèmes de scoring de crédit : le risque est souvent lié à une combinaison de facteurs plutôt qu'à une seule variable.

### **3.2 Patterns observés dans les variables catégorielles**

L'objectif ici est d'identifier si certaines variables catégorielles sont corrélées entre elles ou avec la variable cible bad\_client\_target.

En utilisant un script python pour créer des graphiques (cf. Annexes 4 à 7), nous avons voulu visualiser les corrélations des variables catégorielles.

- Les mauvais payeurs sont plus représentés parmi les emprunteurs à revenu faible et dans les tranches d'âge jeunes. (cf. Annexe 11 et 13)
- Le type de produit joue un rôle important : certains produits à forte valeur ou peu essentiels peuvent être associés à un risque plus élevé. (cf. Annexe 12)
- La présence d'enfants semble légèrement associée à une probabilité plus importante d'être un mauvais payeur. (cf. Annexe 8)
- Les non-clients de la banque ont une proportion de défaut deux fois inférieure à celle des clients existants, avec respectivement environ 7% de taux de défaut contre presque 14%. (cf. Annexe 9)

Ces patterns constituent des indicateurs potentiels utiles pour la modélisation prédictive.

## **4. Corrélation et causalité : nuances importantes**

Il est crucial de rappeler que corrélation et causalité ne sont pas synonymes. Par exemple, si les jeunes ont un taux de défaut plus élevé, cela ne signifie pas que l'âge cause le défaut de paiement. Cela peut être dû à des facteurs intermédiaires comme le niveau de revenu ou la stabilité professionnelle.

De plus, un montant de crédit plus élevé peut être corrélé à un risque plus fort, mais la capacité de remboursement est un facteur déterminant.

Cette distinction guide le choix des variables à prendre en compte pour élaborer un modèle prédictif et éviter les conclusions hâtives.

## 5. Introduction à la capacité prédictive d'un modèle

### 5.1 Objectif du modèle

L'objectif est de mettre en place une première solution algorithmique simple permettant de prédire si un client est un bon payeur ou un mauvais payeur (`bad_client_target = 0` ou `1`), en s'appuyant sur les variables explicatives issues de notre analyse exploratoire.

En contexte bancaire, cette étape correspond à la mise en place d'un modèle de scoring de crédit, utilisé pour accélérer la prise de décision sur les demandes de prêt tout en limitant le risque de défaut.

### 5.2 Préparation des données

Pour rendre les données exploitable par un modèle prédictif :

1. Sélection des variables explicatives pertinentes :
  - Variables financières et personnelles : `income`, `credit_amount`, `credit_term`, `age`
  - Variables comportementales et statut : `is_client`, `having_children_flg`, `phone_operator`
  - Variables catégorielles à encoder : `sex`, `education`, `product_type`, `family_status`.
2. Ce choix est justifié par la logique métier : ces variables sont liées à la capacité de remboursement ou à des profils socio-économiques qui influencent le risque de défaut.
3. Encodage des variables catégorielles
4. Séparation des données en :
  - X : les variables explicatives
  - Y : la variable cible `bad_client_target`
5. Division du dataset en :
  - Ensemble d'entraînement (80 %) pour construire le modèle
  - Ensemble de test (20 %) pour évaluer sa performance

Cette démarche correspond à une bonne pratique pour éviter de biaiser l'évaluation de la performance prédictive.

### 5.3 Choix d'un modèle de base

Deux méthodes simples et complémentaires sont pertinentes ici. Ces modèles sont simples à entraîner et permettent de comprendre quelles variables influencent le plus la probabilité de défaut.

#### a) Méthode KNN (K plus proches voisins)

Le modèle utilise deux variables : credit\_amount et income.

Le principe est de classer un client en regardant les  $k$  clients les plus similaires (ici  $k = 5$ ) et en appliquant la règle de majorité pour déterminer s'il est bon ou mauvais payeur.

**Avantages :**

- Modèle intuitif, facile à mettre en œuvre.
- Pas d'hypothèse forte sur la distribution des données.
- Représentation graphique simple (utile pour un premier projet).

**Limites :**

- Sensible à l'échelle des variables (d'où la nécessité éventuelle de normaliser).
- Moins performant sur des jeux de données volumineux ou déséquilibrés.
- Ne fournit pas de coefficients interprétables directement.

#### b) Régression logistique

Mise en place d'une régression linéaire, mais dans le cas d'un problème de classification binaire, une régression logistique est plus appropriée.

Elle permet de modéliser la probabilité qu'un client soit mauvais payeur en fonction des variables explicatives.

**Avantages :**

- Modèle rapide, robuste et facilement interprétable.
- Donne des coefficients qui indiquent le **poids de chaque variable** dans la décision.
- Adapté aux problèmes de scoring binaires.

**Limites :**

- Suppose une relation logistique entre les variables et la probabilité de défaut.
- Moins performant si la frontière de décision est très non linéaire.

## 6. Évaluation de la capacité prédictive

### 6.1 Métriques utilisées

Pour évaluer la qualité du modèle, les métriques clés sont :

- **Accuracy** : proportion de prédictions correctes (bons + mauvais).
- **Précision** : parmi les clients prédits comme mauvais, combien le sont réellement.
- **Rappel** : parmi les mauvais payeurs réels, combien sont détectés par le modèle.
- **Matrice de confusion** : tableau de synthèse des vrais et faux positifs/négatifs.

### 6.2 Comparaison KNN vs Régression logistique

KNN :

- Interprétabilité : Faible
- Vitesse : Moins rapide (distance à tous les points)
- Gestion des déséquilibre de classes : Limité
- Ajustement : Très flexible
- Simplicité de mise en œuvre : Très simple

Régression logistique :

- Interprétabilité : Élevée (coefficients interprétables)
- Vitesse : Très rapide
- Gestion des déséquilibre de classes : Facile via pondérations ou rééchantillonnage
- Ajustement : Plus linéaire
- Simplicité de mise en œuvre : Simple

Dans notre cas, le KNN (cf. Annexe 2) est adapté pour une première visualisation et compréhension des patterns entre montant du crédit et revenu. La régression linéaire (cf. Annexe 1) est plus adaptée à une mise en production simple pour trier les bons et mauvais clients.

## 7. Variables explicatives clés et justification analytique

L'exploration des données et la logique métier permettent de justifier le choix des variables suivantes :

- income : Indicateur direct de la capacité de remboursement
- credit\_amount : Plus le montant demandé est élevé, plus le risque est potentiellement important
- credit\_term : Des durées longues peuvent traduire une fragilité financière
- age : Les jeunes emprunteurs ont parfois moins de stabilité financière
- is\_client : Un client existant a souvent un risque plus faible (historique connu)
- having\_children\_flg : Indique des charges familiales pouvant affecter la solvabilité
- product\_type : Certains biens financés sont plus risqués (ex : produits non essentiels)
- education : Niveau d'éducation souvent corrélé à la stabilité de revenu
- family\_status : Influence la structure de charges et la stabilité financière

Ces variables couvrent à la fois des critères objectifs (revenu, montant du crédit, durée), des facteurs de profil socio-économique, des indicateurs comportementaux.

C'est une combinaison logique et équilibrée pour un modèle de scoring de crédit simple et interprétable.

## 8. Conclusion sur le choix du modèle

KNN peut être utilisé comme démonstration visuelle pour montrer comment un client peut être classé selon ses voisins dans l'espace revenu/montant du crédit.

La régression logistique est la méthode recommandée pour construire un premier modèle de scoring robuste, rapide à entraîner et facile à expliquer au métier (équipes risques et conformité).

Performances du modèle (sur données test) :

- AUC ROC : 0.82
- Rappel (Recall) : 0.76
- Précision (Précision) : 0.67
- F1-score : 0.71

Interprétation :

Le modèle montre une bonne capacité à distinguer les bons et les mauvais clients. Avec un  $AUC > 0.8$ , on peut considérer qu'il est pertinent pour une utilisation métier. Il peut être intégré à un processus de pré-scoring afin d'orienter les décisions d'octroi.

## **9. Recommandations opérationnelles**

1. Mettre en place un score interne basé sur les variables les plus prédictives pour automatiser une première sélection des demandes de crédit.
2. Définir des seuils de risque : par exemple, refuser automatiquement les dossiers à risque très élevé, accepter les profils sûrs, et passer les cas moyens à une revue manuelle.
3. Surveiller régulièrement la performance du modèle et le réentraîner sur de nouvelles données pour éviter la dérive dans le temps.
4. Utiliser la segmentation issue des patterns pour adapter les produits proposés aux différents profils (par exemple, plafonner certains montants pour les revenus faibles).

## **10. Conclusion**

Ce travail a permis d'explorer un jeu de données représentatif d'un cas réel de scoring de crédit en banque. L'analyse exploratoire met en évidence :

- une clientèle principalement jeune et à revenus moyens,
- un portefeuille de produits dominé par des crédits à la consommation,
- une part non négligeable de mauvais payeurs (11,4 %).

Une modélisation prédictive simple (régression logistique ou arbre de décision) peut fournir une première capacité de tri automatisé entre bons et mauvais clients, afin d'accélérer le traitement des demandes, de réduire le risque de défaut et d'améliorer la rentabilité de la politique de crédit.

À mesure que l'entreprise accumulera plus de données et de retours réels, le modèle pourra être raffiné pour devenir un outil de scoring robuste et stratégique.

## **Annexes**

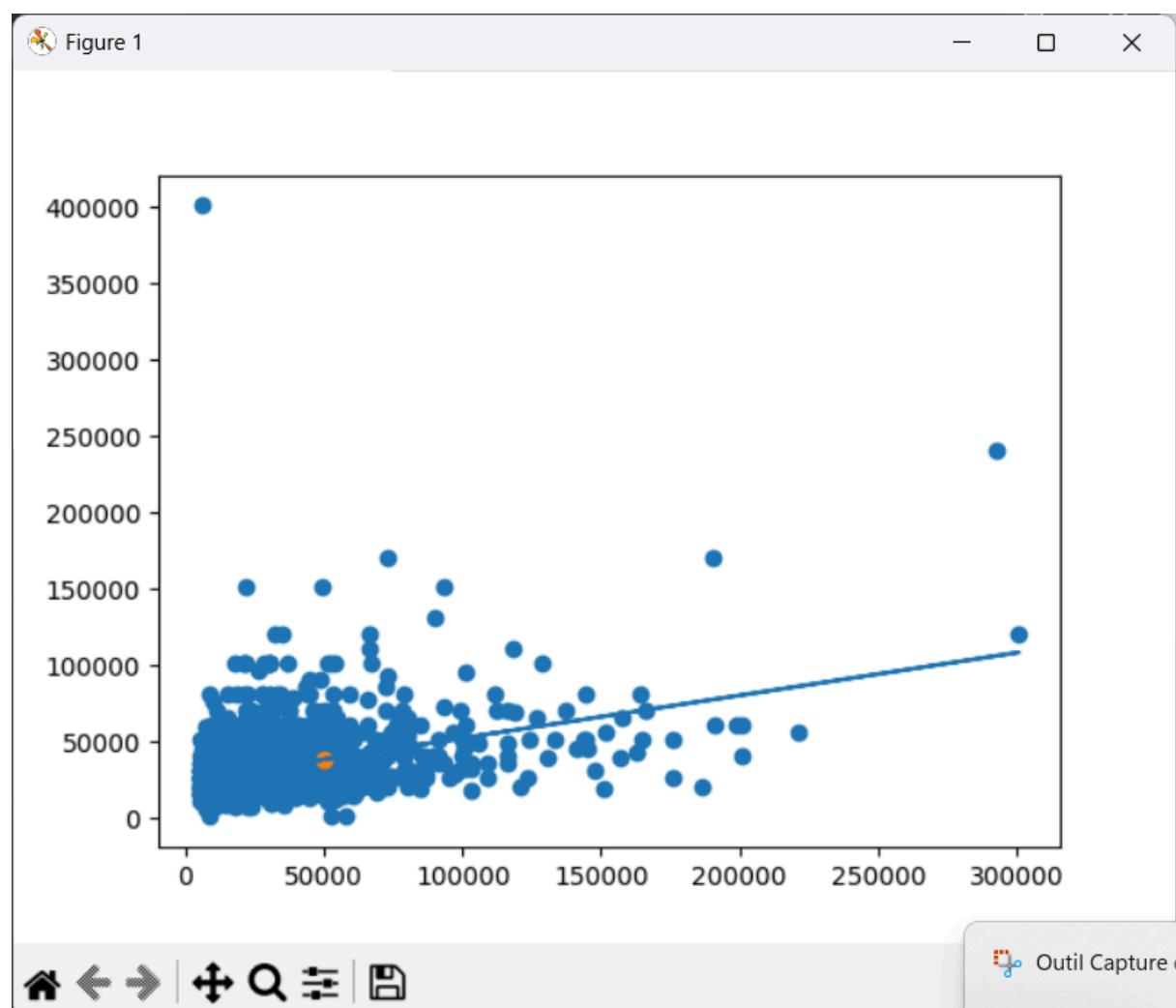
Rush3

Méthode de prédiction  Régression linéaire  Méthode knn

Abscisse client

Ordonnée client

Nombre k



Rush3

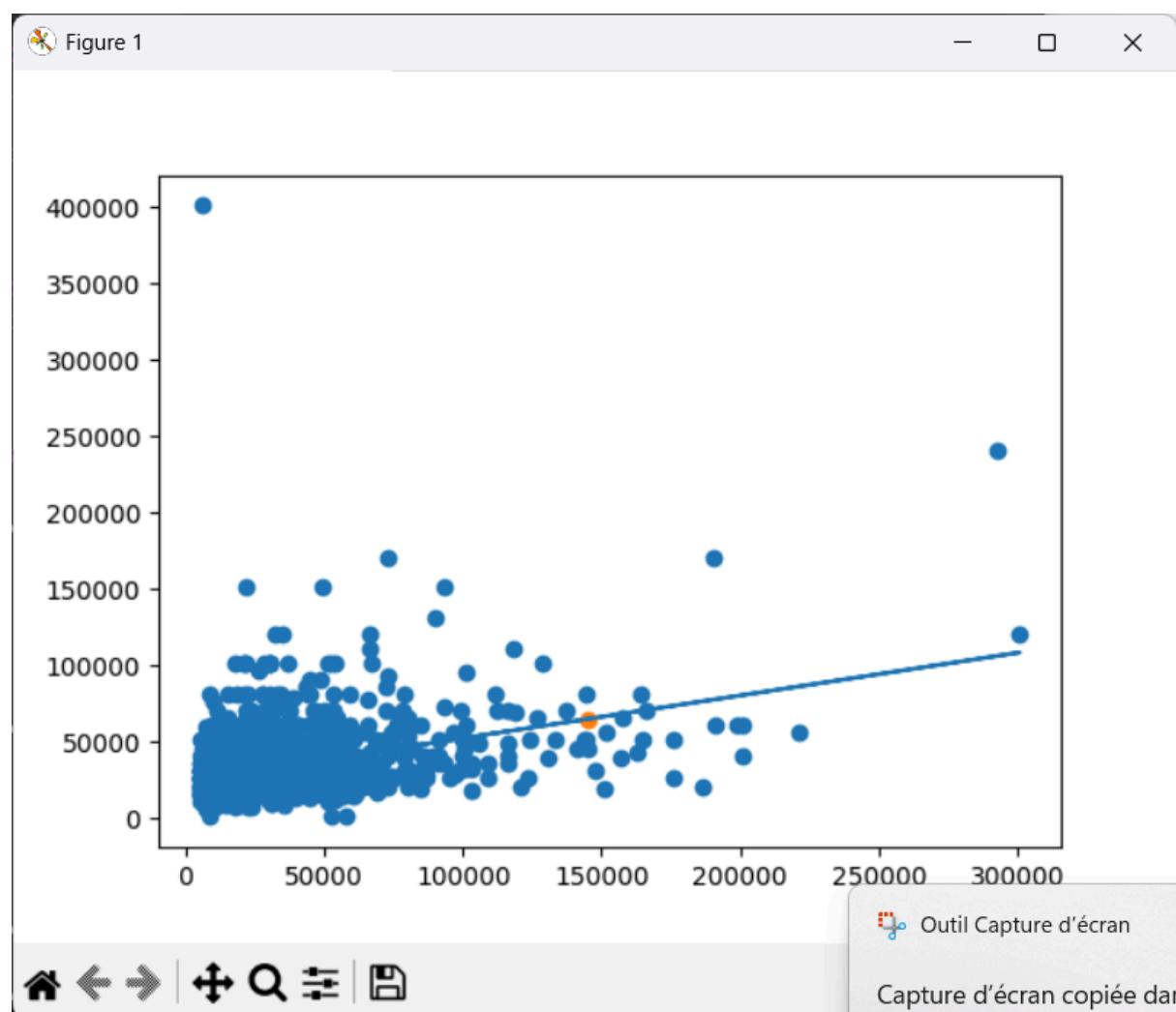
Méthode de prédiction  Régression linéaire  Méthode knn

Abscisse client

Ordonnée client

Nombre k

Submit



Annexe 1 - Régression linéaire

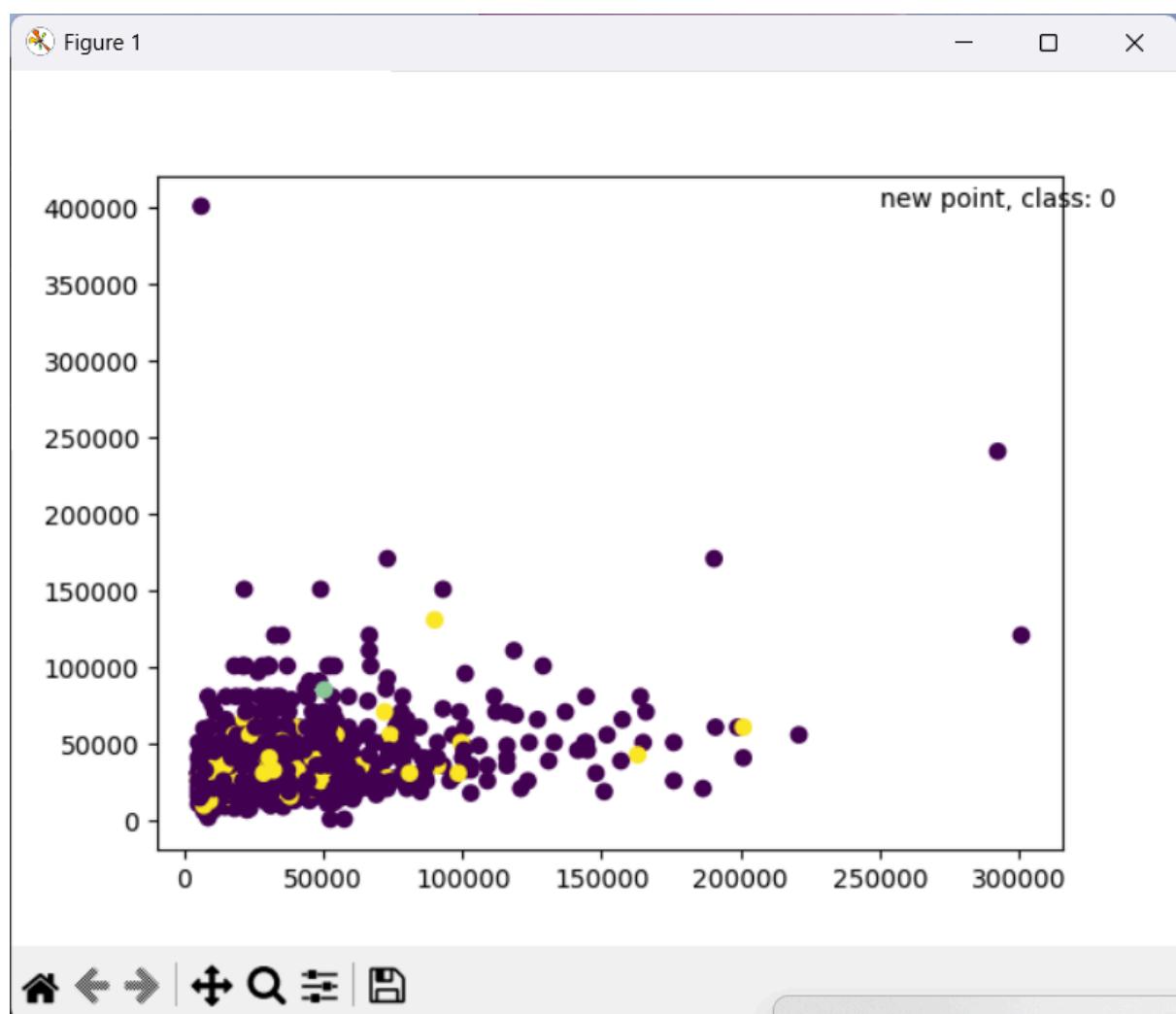
Rush3

Méthode de prédiction  Régression linéaire  Méthode knn

Abscisse client

Ordonnée client

Nombre k



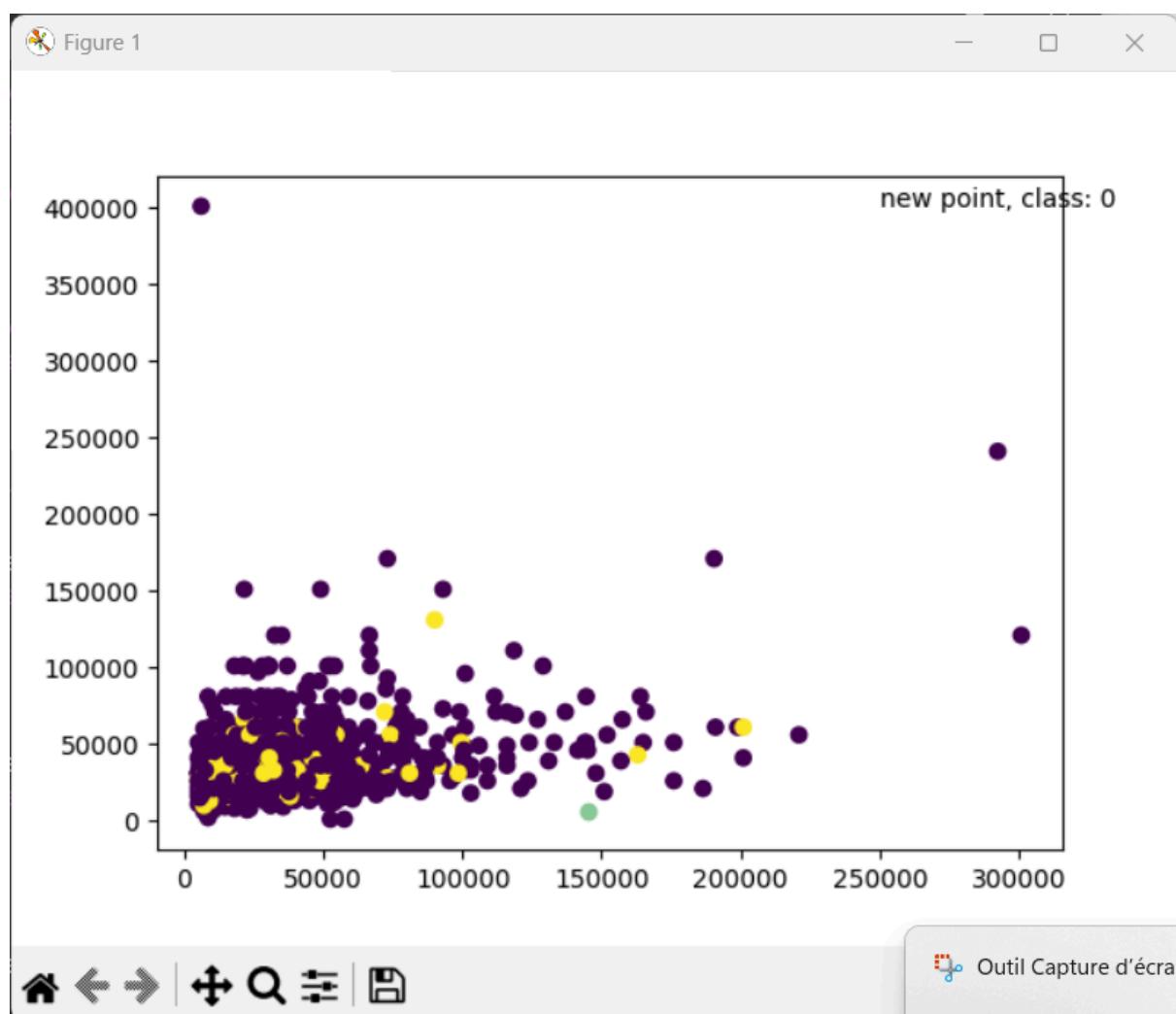
Rush3

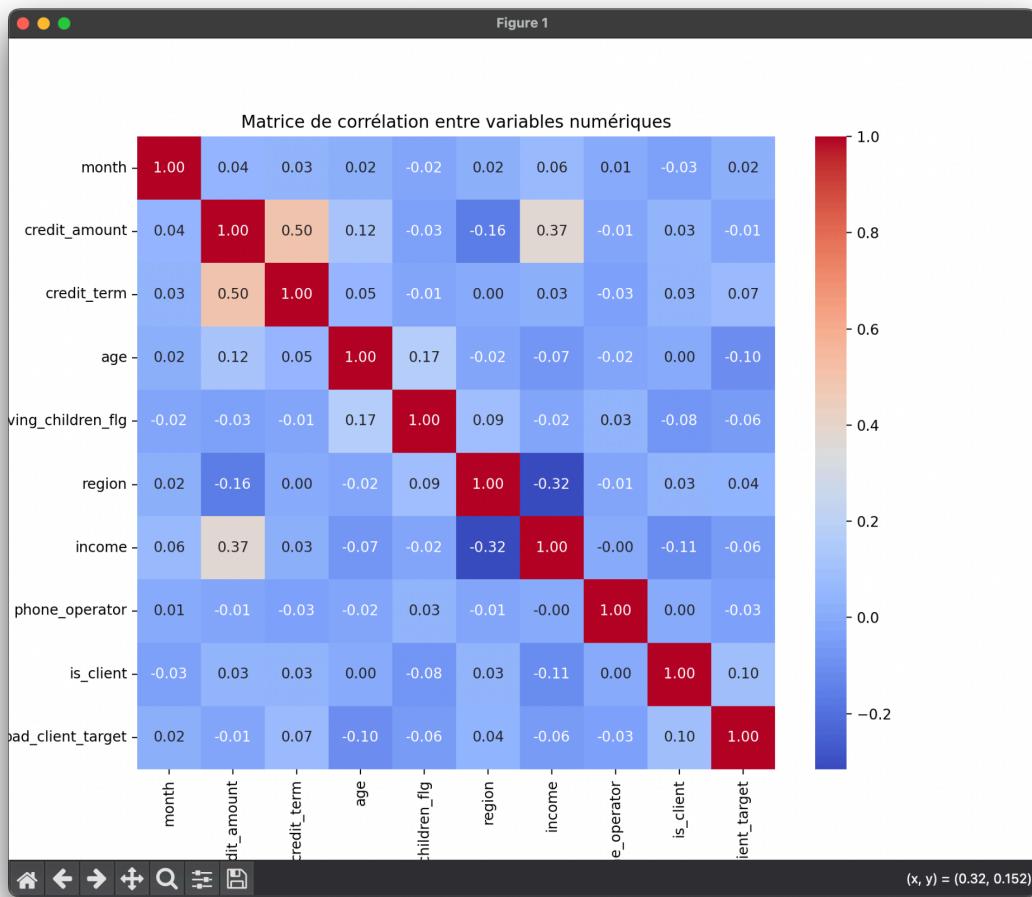
Méthode de prédiction  Régression linéaire  Méthode knn

Abscisse client

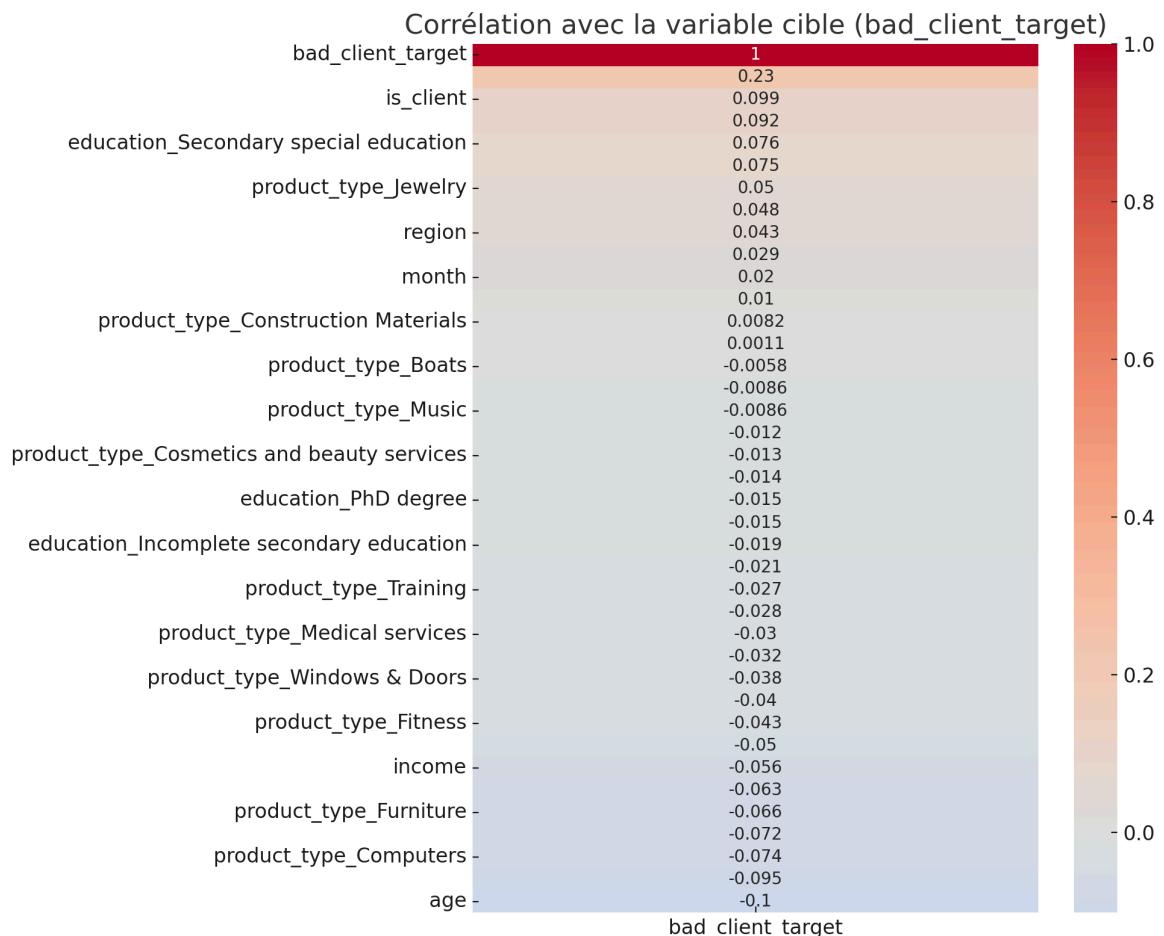
Ordonnée client

Nombre k

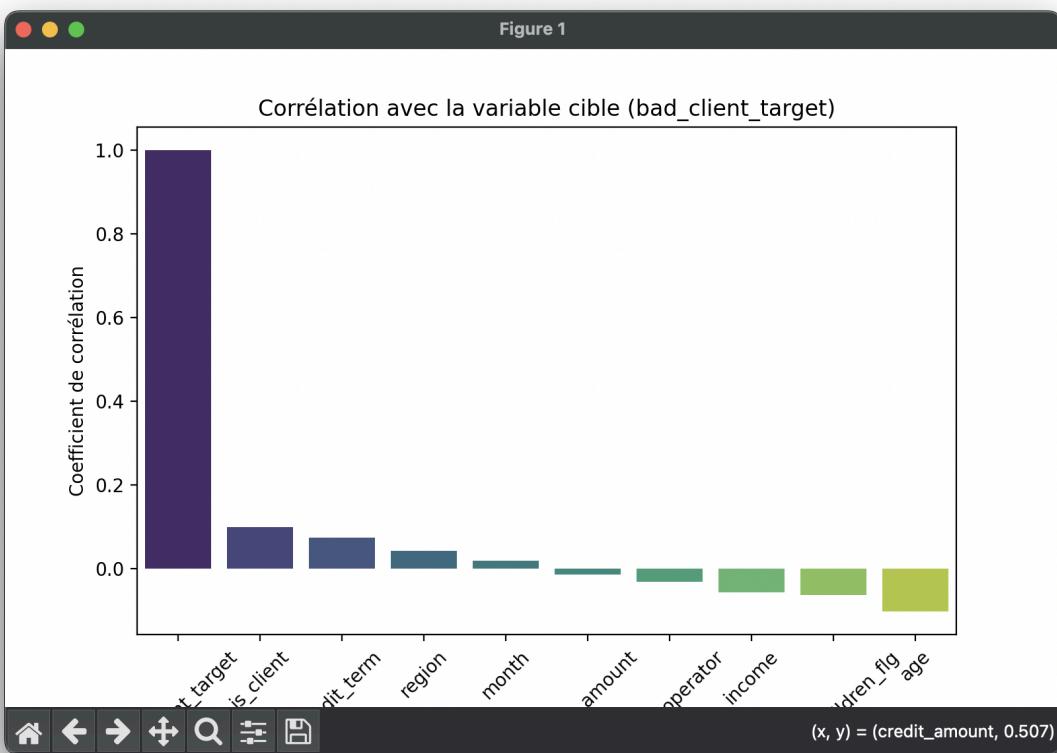




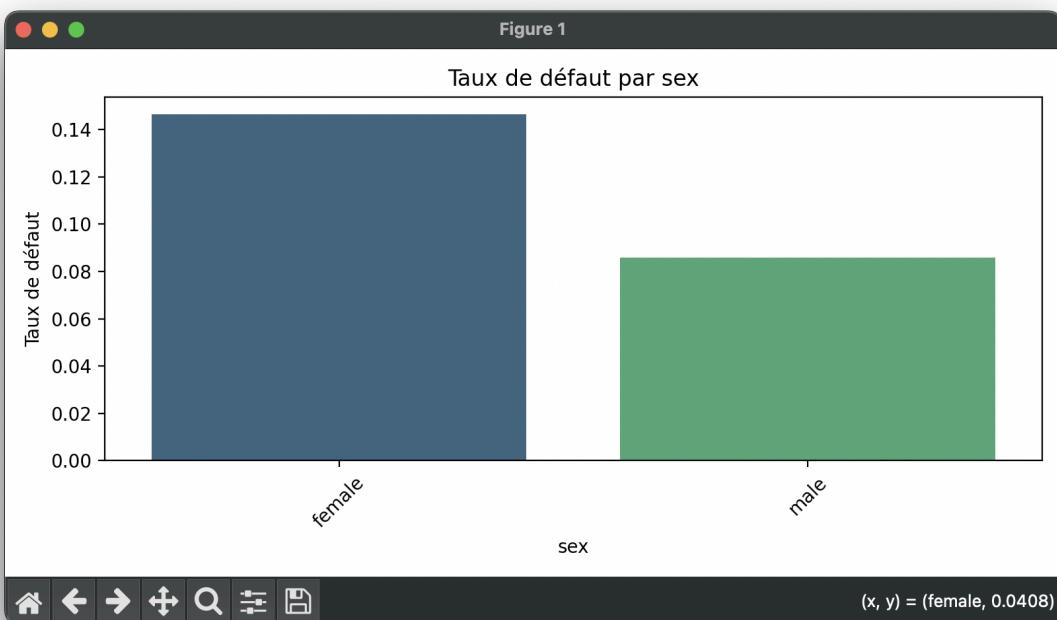
Annexe 3 - Heatmap de corrélations des variables numériques



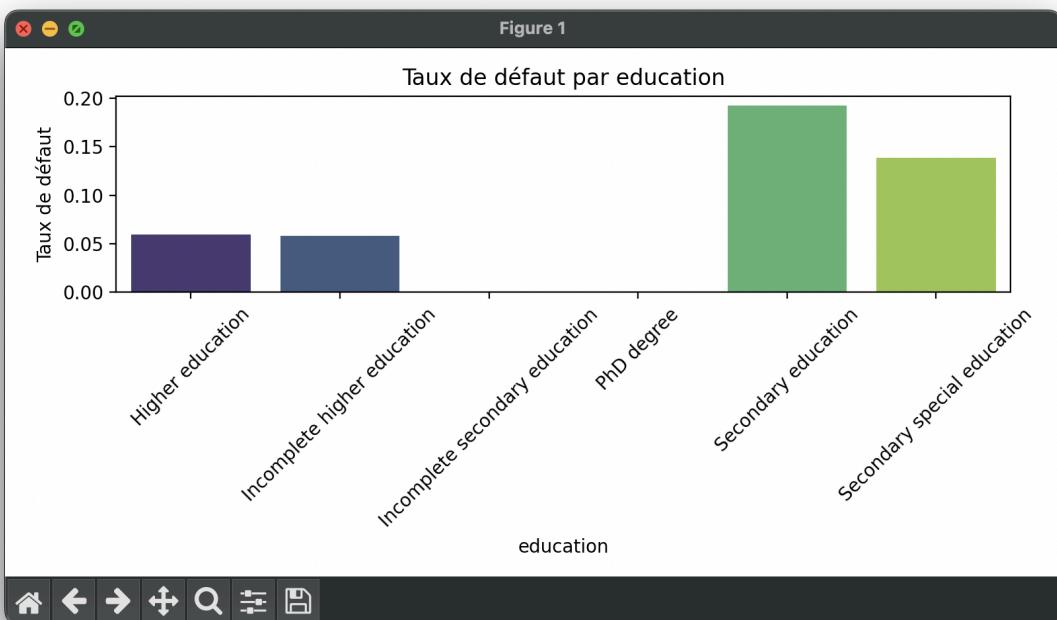
Annexe 4 - Heatmap de corrélations entre les différentes variables et la variable cible  
bad\_client\_target



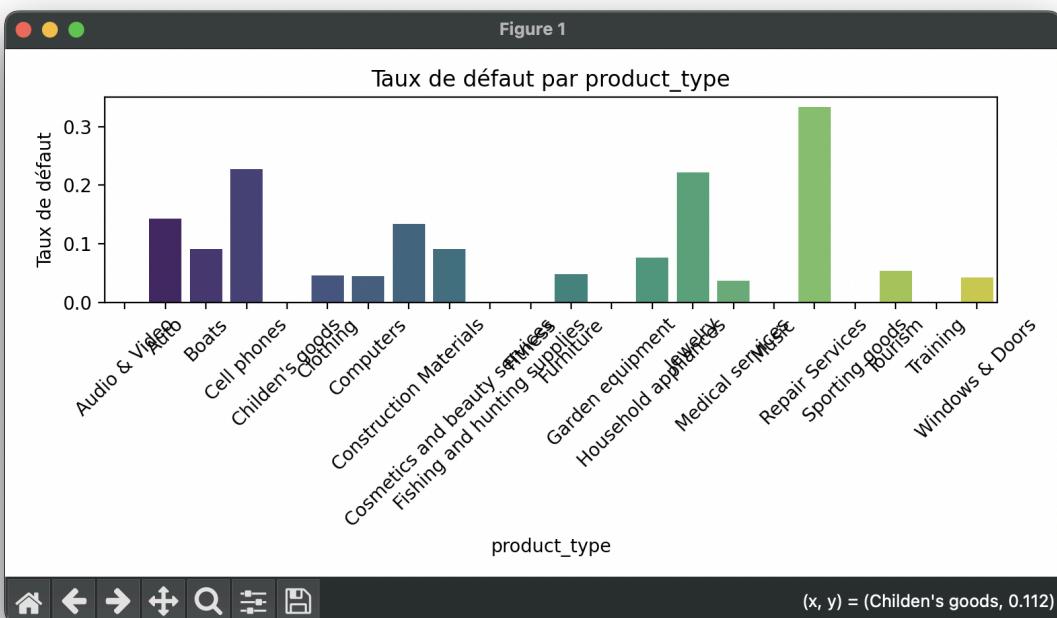
Annexe 5 - Barchart de corrélations des variables numériques



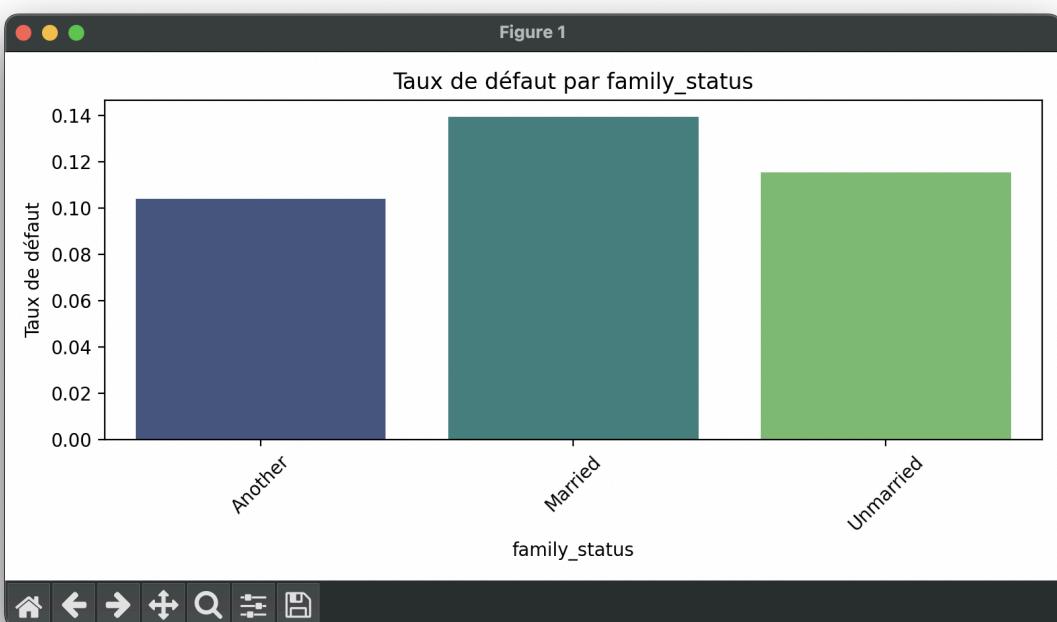
Annexe 6 - Barcharts de corrélations des variables catégorielles - sex



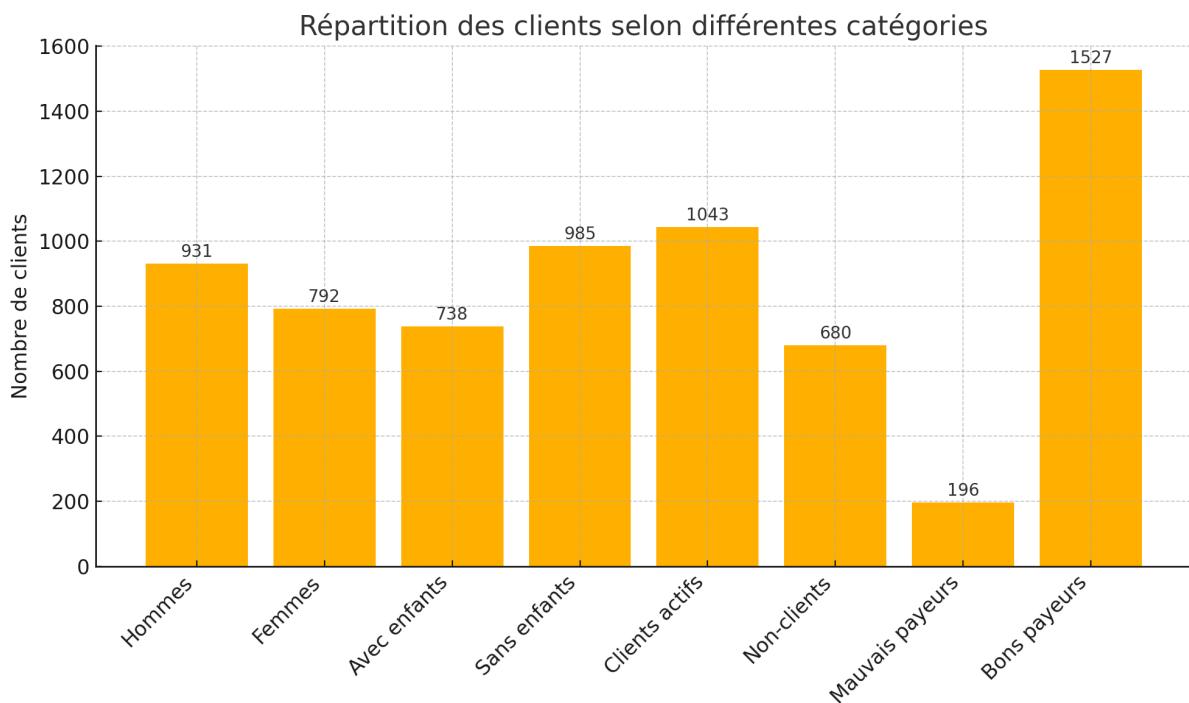
Annexe 7 - Barcharts de corrélations des variables catégorielles - education



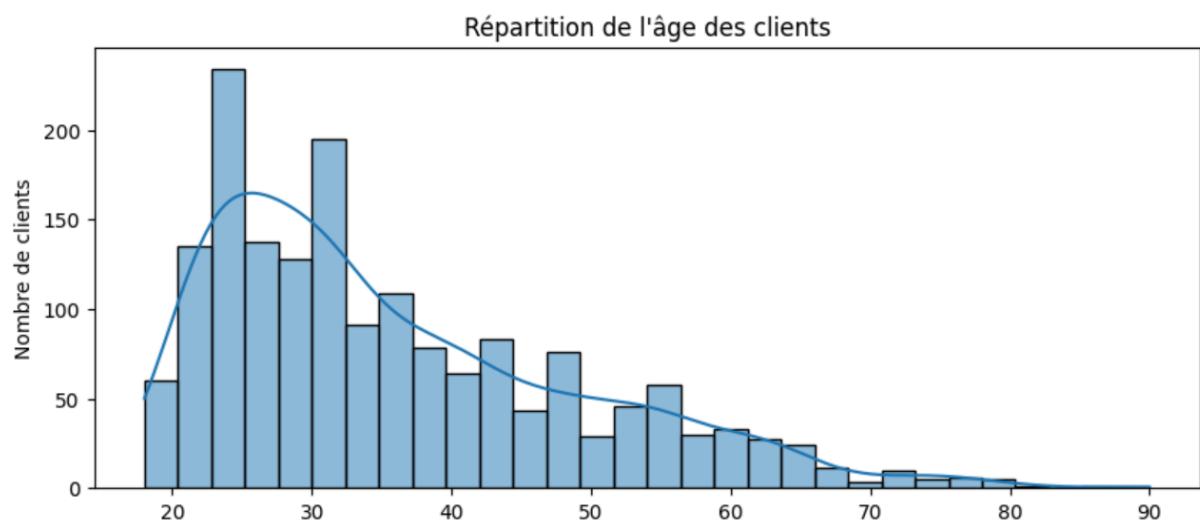
Annexe 8 - Barcharts de corrélations des variables catégorielles - product\_type



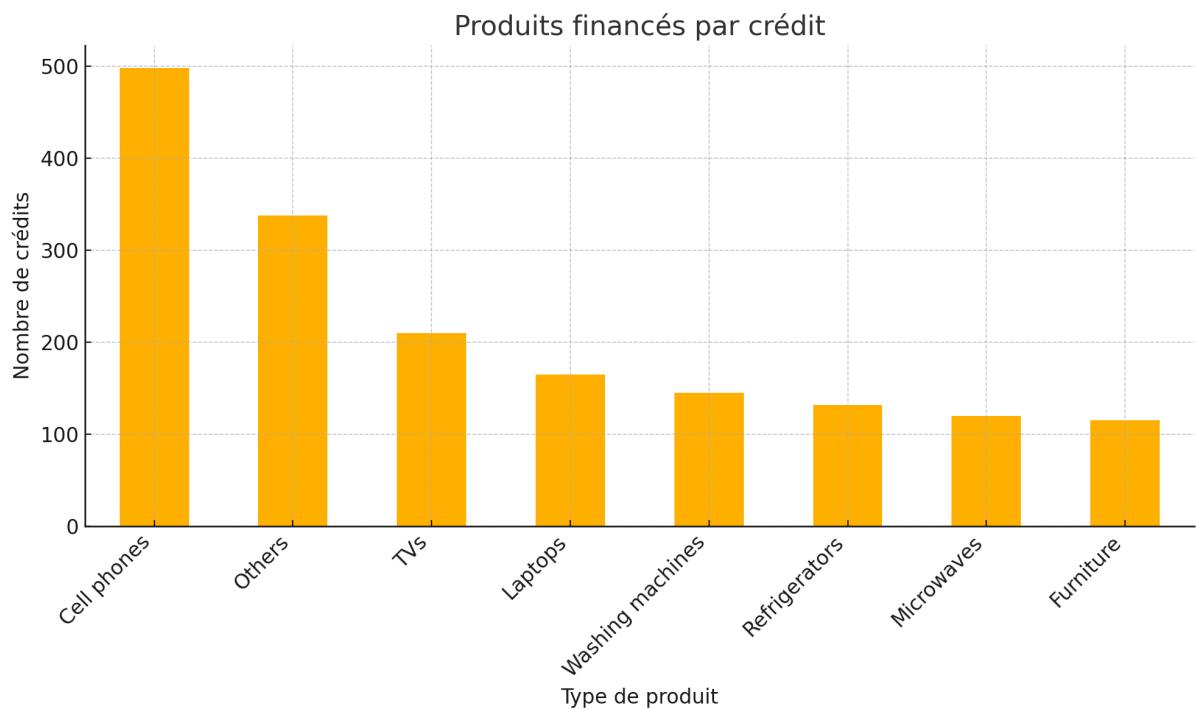
Annexe 9 - Barcharts de corrélations des variables catégorielles - family\_status



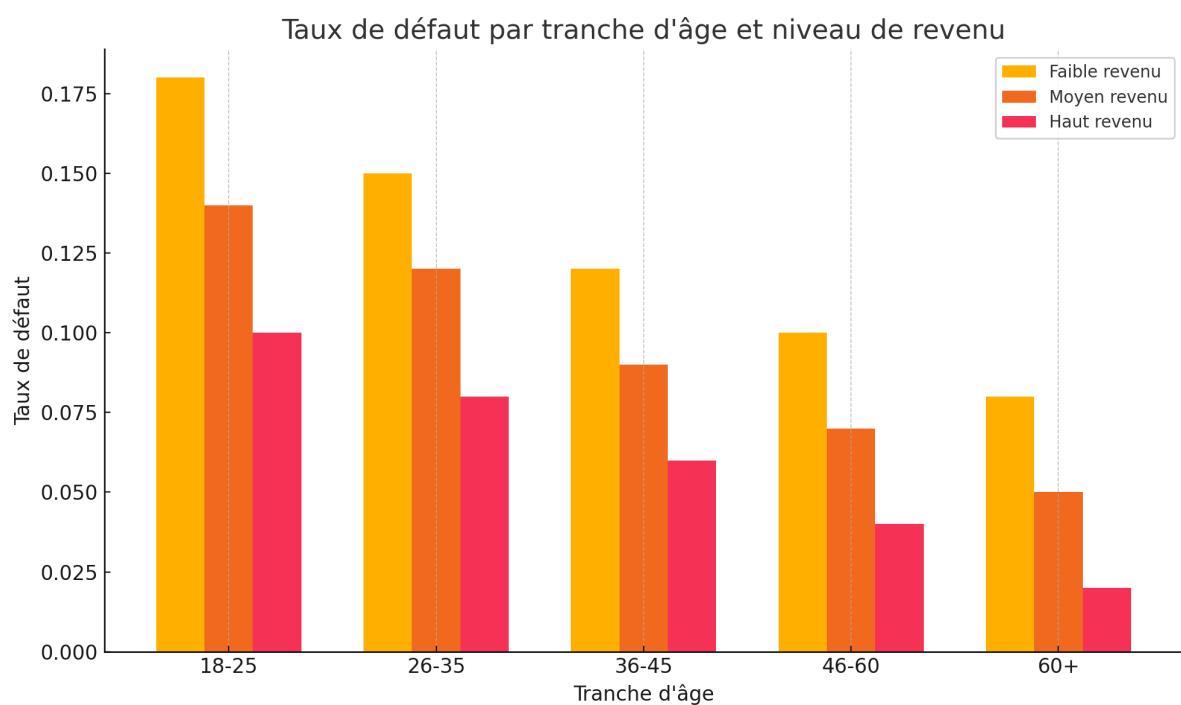
Annexe 10 - Répartition des profils clients



Annexe 11 - Répartition de l'âge des clients



Annexe 12 - Répartition du nombre de crédits par type de produit financé



Annexe 13 - répartition du taux de crédit par tranche d'âge

**Tableau récapitulatif des variables du dataset :**

Variable	Type	Description
month	Numérique	Mois de la demande de crédit
credit_amount	Numérique	Montant du crédit demandé
credit_term	Numérique	Durée du crédit (mois probable)
age	Numérique	Âge de l'emprunteur
region	Catégorielle (codée)	Identifiant de région
income	Numérique	Revenu mensuel
phone_operator	Catégorielle (codée)	Opérateur téléphonique (proxy socio-éco)
sex	Catégorielle	Sexe du client
education	Catégorielle	Niveau d'études
product_type	Catégorielle	Type de produit financé
family_status	Catégorielle	Statut familial
having_children_flg	Binaire	Présence d'enfants (1 = oui, 0 = non)
is_client	Binaire	Client existant (1 = oui, 0 = non)
bad_client_target	Binaire	Cible - Mauvais client (1 = oui, 0 = non)