**Sujet du projet : On doit étudier le présent à la lumière du passé afin d’éclairer le futur » (John Maynar KEYNES)**

Le projet concerne un **scoring de crédit** pour des prêts hypothécaires. Tu as un jeu de données avec des informations sur 5 960 prêts hypothécaires (prêts où la maison est utilisée comme garantie). L'objectif est de **prédire si un client va faire défaut (BAD = 1)** sur son prêt ou non (BAD = 0), en utilisant les autres variables disponibles dans le jeu de données.

**BAD** :

* **1** signifie que le client a fait défaut ou a eu des retards de paiement importants.
* **0** signifie que le client a remboursé son prêt correctement.

**LOAN** : Montant du prêt demandé par le client.

Si LOAN = 50,000, cela signifie que le client a demandé un prêt de 50 000 euros pour financer quelque chose (ex. : des rénovations ou pour consolider ses dettes).

**MORTDUE** : Montant restant à payer sur l'hypothèque actuelle (autre prêt hypothécaire que le client pourrait avoir).

**Exemple** : Si MORTDUE = 75,000, cela signifie que le client doit encore 75 000 euros sur son prêt hypothécaire existant.

**VALUE** : Valeur actuelle de la propriété du client (la maison qui est utilisée comme garantie).

**Exemple** : Si VALUE = 200,000, cela signifie que la maison du client vaut 200 000 euros. C'est une information importante pour évaluer la sécurité du prêt, car la maison sert de garantie.

 **REASON** :

* **DebtCon** (consolidation de dettes) : Le client a pris ce prêt pour consolider d'autres dettes.
* **HomeImp** (amélioration de l'habitat) : Le client a pris ce prêt pour améliorer sa maison.

Exemple : Si REASON = DebtCon, cela signifie que le client a pris ce prêt pour consolider ses autres dettes et en rembourser une seule à la banque.

 **JOB** : Catégorie professionnelle du client (par exemple, fonctionnaire, employé, etc.).

**Exemple** : Si JOB = Manager, cela signifie que le client est un cadre dans une entreprise.

 **YOJ** : Nombre d'années que le client a passé dans son emploi actuel ("Years on Job").

Exemple : Si YOJ = 5, cela signifie que le client occupe son poste actuel depuis 5 ans. C’est un indicateur de stabilité financière.

**DEROG** : Nombre de rapports de , défauts de paiement majeurs dans le dossier de crédit du client (par exemple, défauts de paiement, faillites).

Exemple : Si DEROG = 2, cela signifie que le client a eu deux incidents majeurs de délinquance financière, comme des faillites dans le passé.

**DELINQ** : Nombre de lignes de crédit (cartes de crédit, prêts, etc.) du client ayant été en retard de paiement.

Une valeur élevée de **DELINQ** est généralement un indicateur de risque pour les prêteurs, car cela montre que le client a déjà eu des difficultés à respecter ses obligations financières.

Si un client a la variable DELINQ = 3, cela signifie que sur trois lignes de crédit différentes (prêts ou cartes de crédit), ce client a été en retard de paiement à un moment donné.

**CLAGE** : Âge de la plus ancienne ligne de crédit du client, en mois.

**Exemple** : Si CLAGE = 120, cela signifie que le client a une ligne de crédit qui existe depuis 120 mois, soit 10 ans. Plus la ligne de crédit est ancienne, plus cela montre que le client a de l'historique financier.

**NINQ** : Nombre de récentes demandes d'enquête de crédit (par exemple, lorsque le client a demandé un nouveau crédit ou un prêt).

**Exemple** : Si NINQ = 4, cela signifie que le client a fait quatre demandes récentes de crédit. Trop de demandes récentes peuvent indiquer un risque de surendettement.

**CLNO** : Nombre total de lignes de crédit ouvertes par le client.

**Exemple** : Si CLNO = 10, cela signifie que le client a 10 lignes de crédit ouvertes (peut-être 5 cartes de crédit et 5 prêts, par exemple).

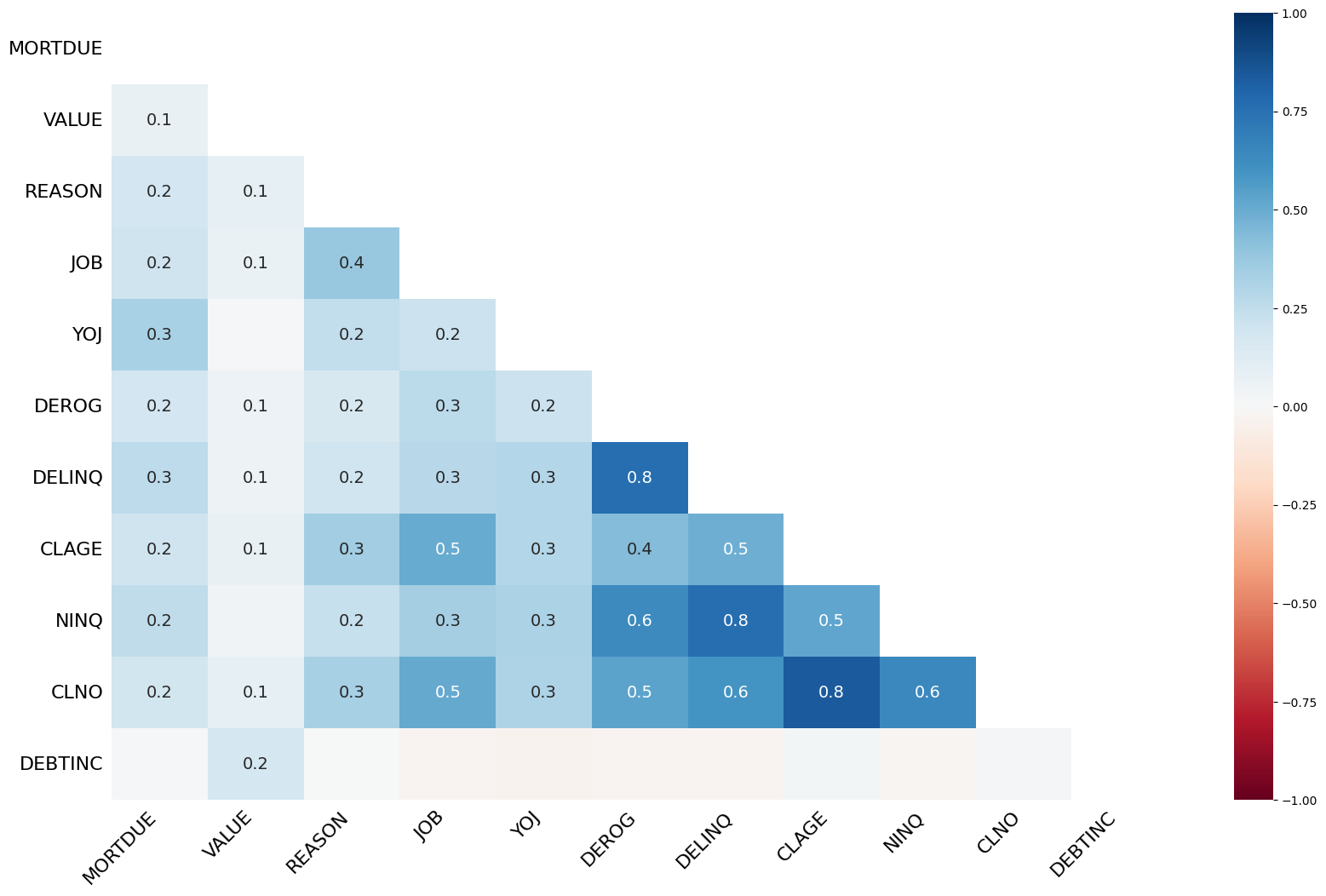
**DEBTINC** : Ratio dette/revenu du client (combien de dette il a par rapport à ses revenus).

**Exemple** : Si DEBTINC = 35, cela signifie que 35 % du revenu mensuel du client est utilisé pour rembourser ses dettes. Un ratio élevé peut être un signal de surendettement.

###########



* **Gestion des variables manquantes**

****

**La carte thermique de corrélation manquante aide à comprendre la corrélation nulle : dans quelle mesure la présence ou l'absence d'une variable affecte la présence d'une autre :**

Si deux variables sont fortement corrélées (proche de 1), cela signifie que, lorsque des données manquent dans l'une des variables, il est très probable qu'elles manquent aussi dans l'autre.

1. Il y a une forte corrélation entre les valeurs manquantes de **DEROG (nombre de rapports dérogatoires) et DELINQ (nombre de lignes de crédit en retard).** Cela signifie que si des valeurs manquent pour **DEROG,** elles manquent probablement aussi pour **DELINQ.**

**Action : l'imputation par régression ou MissForest, pour imputer ces deux variables ensemble. Elles ont probablement des dépendances sous-jacentes (par exemple, les clients ayant des rapports dérogatoires ont souvent des lignes de crédit en retard).**

1. **CLNO, NINQ, et CLAGE (Corrélations modérées) :** CLNO (nombre total de lignes de crédit), NINQ (nombre de demandes récentes de crédit) et CLAGE (âge de la ligne de crédit la plus ancienne) montrent des corrélations modérées entre elles (entre 0.5 et 0.6). Cela signifie que les valeurs manquantes dans ces colonnes sont souvent présentes ensemble, mais pas systématiquement.

**Action : l'imputation multiple ou l'imputation par KNN,**

1. **YOJ (Années à l'emploi actuel) et MORTDUE (Montant dû sur l'hypothèque)**

**Conclusion :**

**Actions globales à suivre pour traiter les valeurs manquantes :**

1. **Variables avec forte corrélation des valeurs manquantes** (ex. **DEROG** et **DELINQ**) :
   * Utilise des **méthodes d'imputation avancées** (imputation par régression, MissForest, KNN), car ces variables sont probablement liées par une relation sous-jacente.
2. **Variables avec corrélation modérée** (ex. **CLNO**, **NINQ**, **CLAGE**) :
   * Utilise des méthodes comme l'**imputation multiple** ou **KNN** pour imputer ces variables ensemble, car elles montrent des relations partielles.
3. **Variables avec faible ou aucune corrélation** (ex. **YOJ**, **MORTDUE**, **DEBTINC**) :
   * Ces variables peuvent être imputées de manière indépendante avec des méthodes plus simples, comme la **moyenne**, la **médiane**, ou la **mode** pour les variables catégorielles.
4. **Validation des imputations** :
   * Après l'imputation, vérifie les distributions des variables pour t'assurer que l'imputation n'a pas introduit de biais significatifs dans les données.