Ecole superieur de la
statistique et de l’analyse de
l’information

Projet de fin d’année

PRÉSENTÉ PAR :

Hamzaoui Ahlem

2023/2024

Sommaire:

Introduction:

Les cartes de crédit permettent aux clients d'emprunter de l'argent jusqu'à une limite définie, mais elles représentent aussi un risque pour les banques. Pour éviter ces risques, les banques doivent prédire avec précision la probabilité de défaut de paiement.

Retour à l'ordre du jour

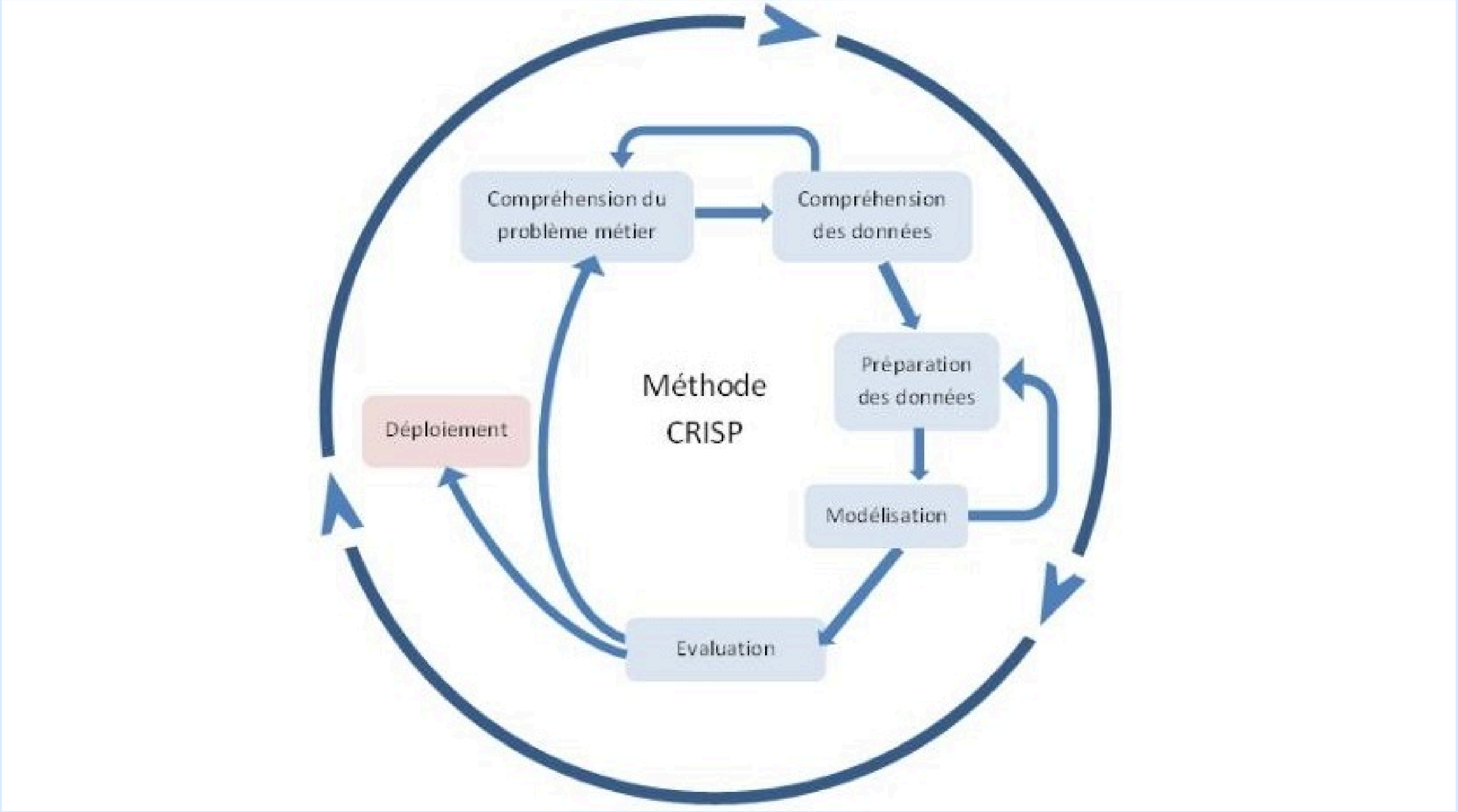
Objectifs:

Analyser les données pour trouver les facteurs clés qui affectent les probabilités de défaut des cartes de crédit.

Prédire la probabilité de défaut pour les clients des cartes de crédit de la banque.

Méthodologie :

Retour à l'ordre du jour



Compréhension des données :

Retour à l'ordre du jour

24 Variables :

- **ID** identifiant client
- **LIMIT_BAL** Montant du crédit
- **SEX** (1=homme, 2=femme)
- **EDUCATION** Niveau d'éducation (1=école doctorale, 2=université, 3=lycée, 4=autres, 5=inconnu, 6=inconnu)
- **MARRIAGE** État civil (1=marié, 2=célibataire, 3=autres)
- **AGE** Âge en années
- **PAY_0 - PAY_6** Historique des paiements passés (en dollars NT) (d'avril à septembre 2005) (-1=paiement ponctuel, 1=délai de paiement d'un mois, 2=délai de paiement de deux mois... 8=délai de paiement de huit mois, 9=délai de paiement de neuf mois et plus)
- **BILL_AMT1 - BILL_AMT6** Montant des factures (en dollars NT) (d'avril à septembre 2005)
- **PAY_AMT1 - PAY_AMT6** Montant du paiement précédent (en dollars NT) (d'avril à septembre 2005)
- **DEFAULT.PAYMENT.NEXT.MONTH** DÉFAUT DE PAIEMENT (1=OUI, 0=NON)

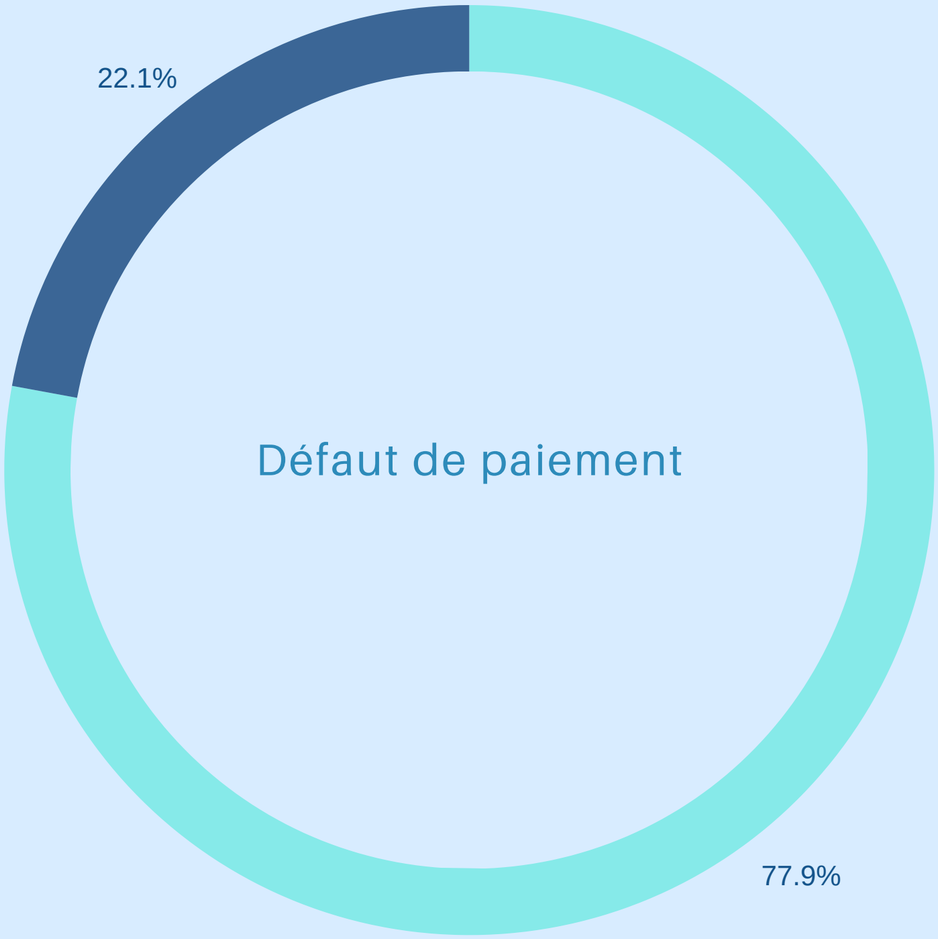
Nettoyage des données :

Retour à l'ordre du jour

- 3 fait référence à “autre” pour MARRIAGE (1=marié, 2=célibataire, 3=autres)
- On a regroupé les valeurs 4, 5, 6 sous la valeur 0 pour EDUCATION
- -1 indique l'utilisation du crédit renouvelable dans PAY_0 - PAY_6.
- On n’a pas des valeurs manquantes .
- vérification du type de chaque variables et les modifier .

Analyse univariée et bivariée des données :

1.Distribution de la variable cible :



DEFAUT

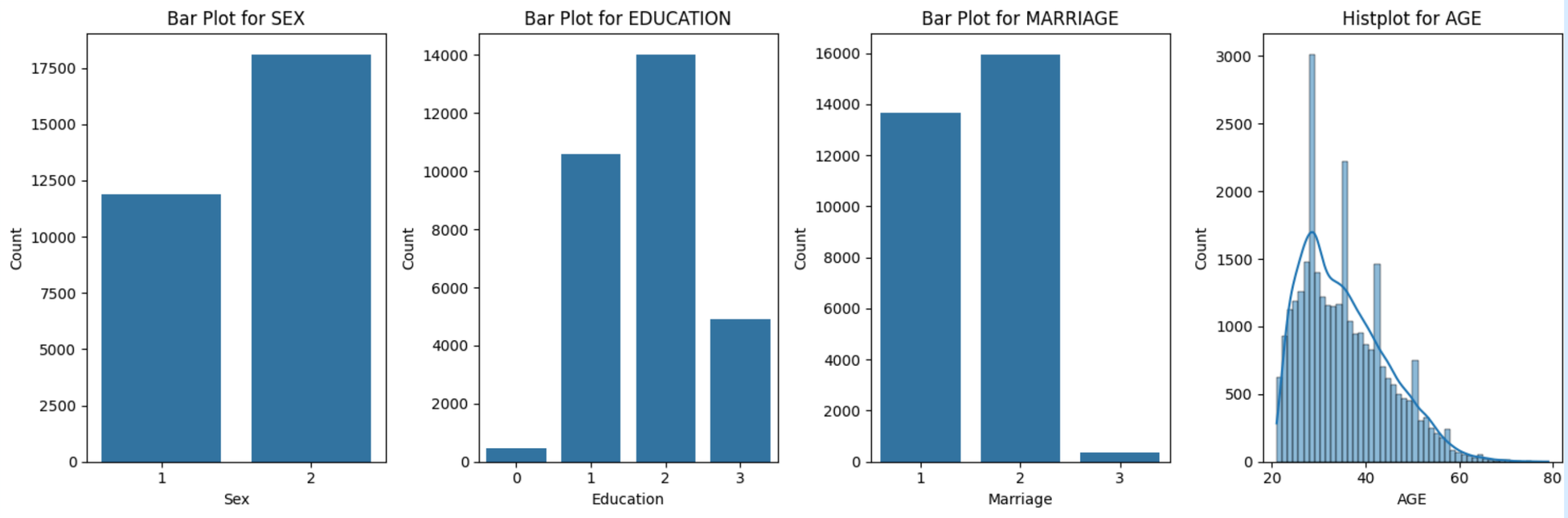
PAS DE DEFAUT

Conclusion :

Les données ne sont pas distribuées de manière égale.
Le nombre des clients qui ne font pas défaut est bien plus important que le nombre des clients qui font défaut.

Analyse univariée et bivariée des données :

2.Distribution univariée du Sexe,Education,Marriage,Age :

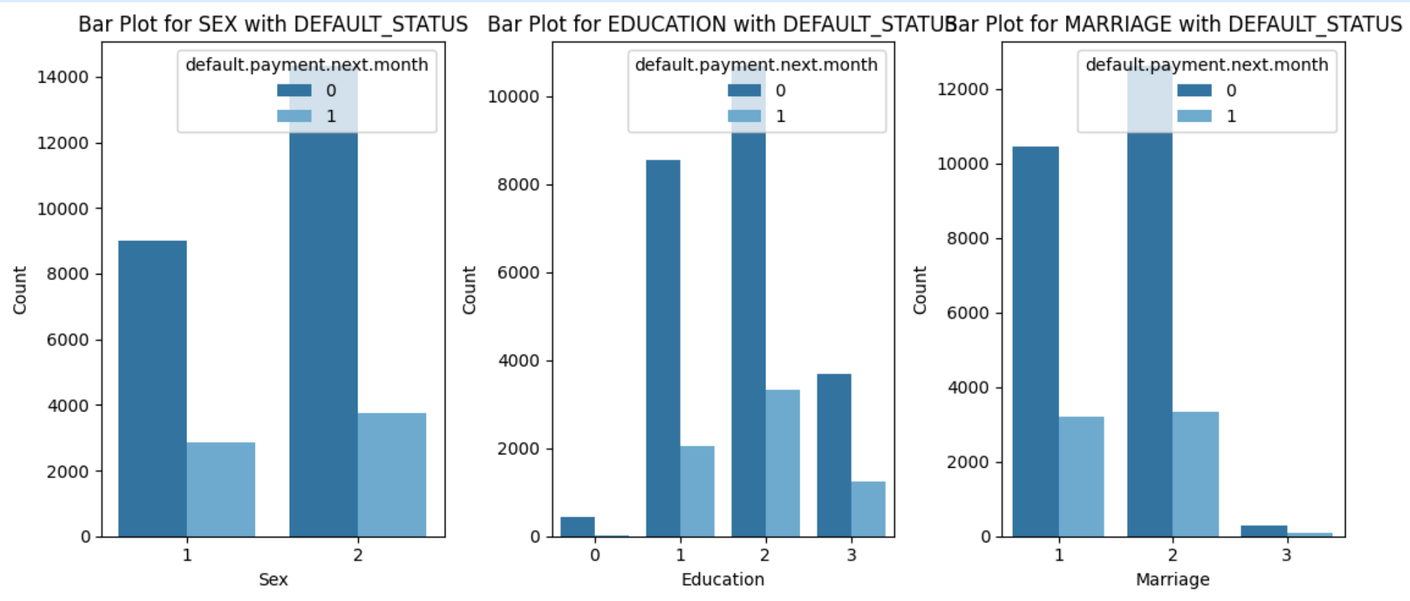


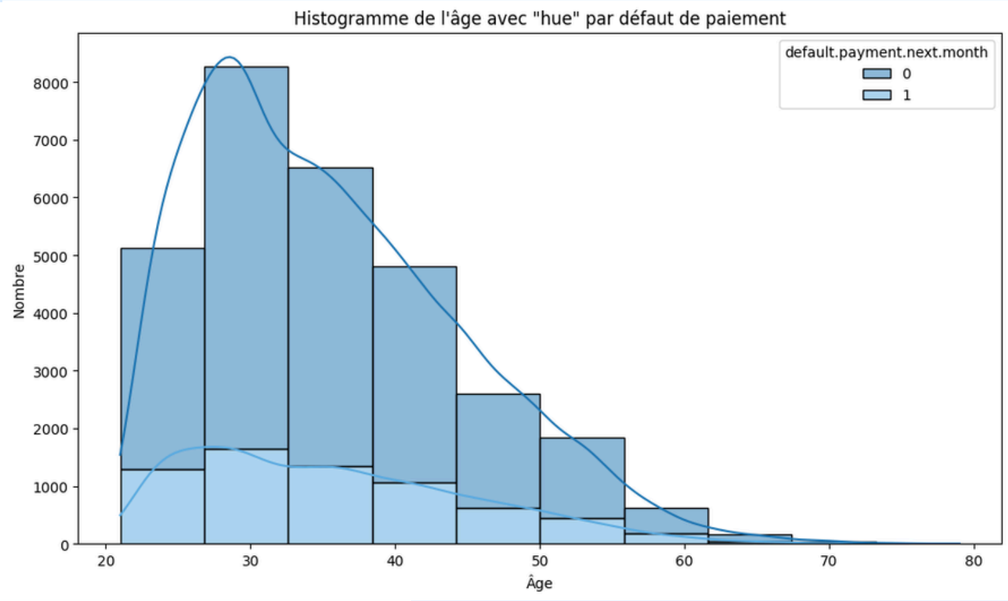
Conclusion :

Les caractéristiques de la base de données sont principalement des femmes, diplômées de l'université, célibataires agées entre 20ans et 45ans

Analyse univariée et bivariée des données :

2.Distribution bivariée du Sexe,Education,Marriage,Age :





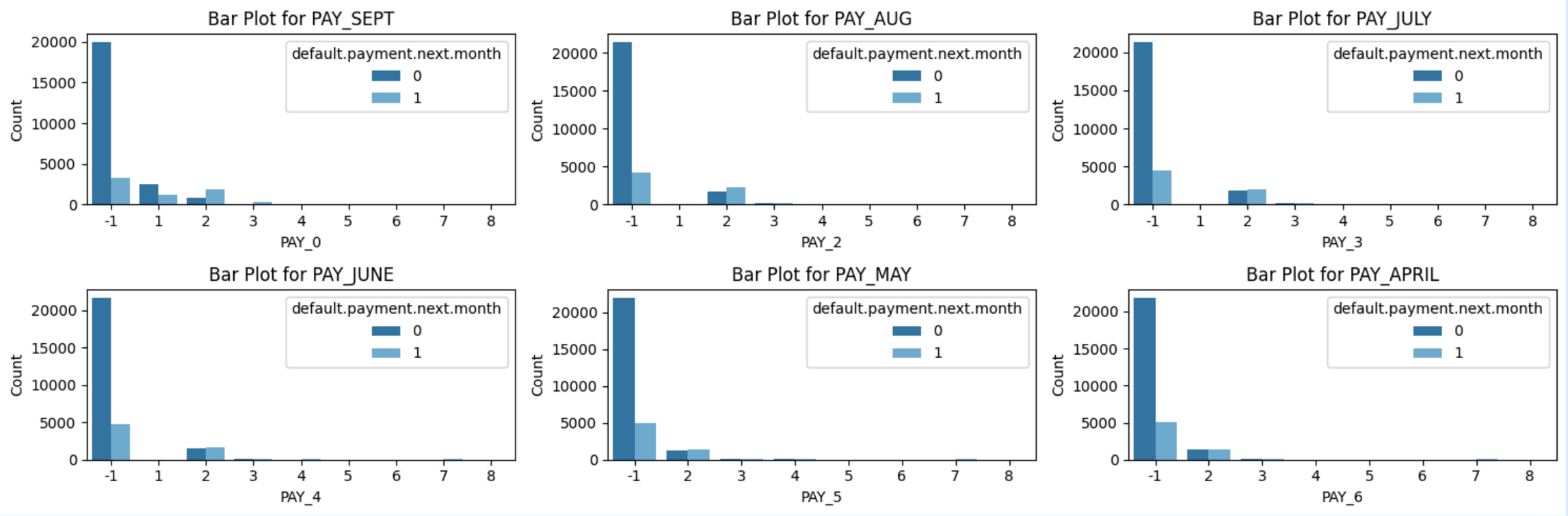
Conclusion :

Les clients qui n'ont pas de diplôme d'études supérieures, universitaires ou secondaires ont tendance à montrer un taux de défaut plus élevé, se situant entre 30% et 40%, quel que soit leur état civil.

Les clients diplômés catégorisés comme 'Autre' présentent un risque significatif de 50% sur leurs paiements de carte de crédit.

Analyse univariée et bivariée des données :

2.Distribution bivariéé du PAY_APRIL - PAY_APRIL :

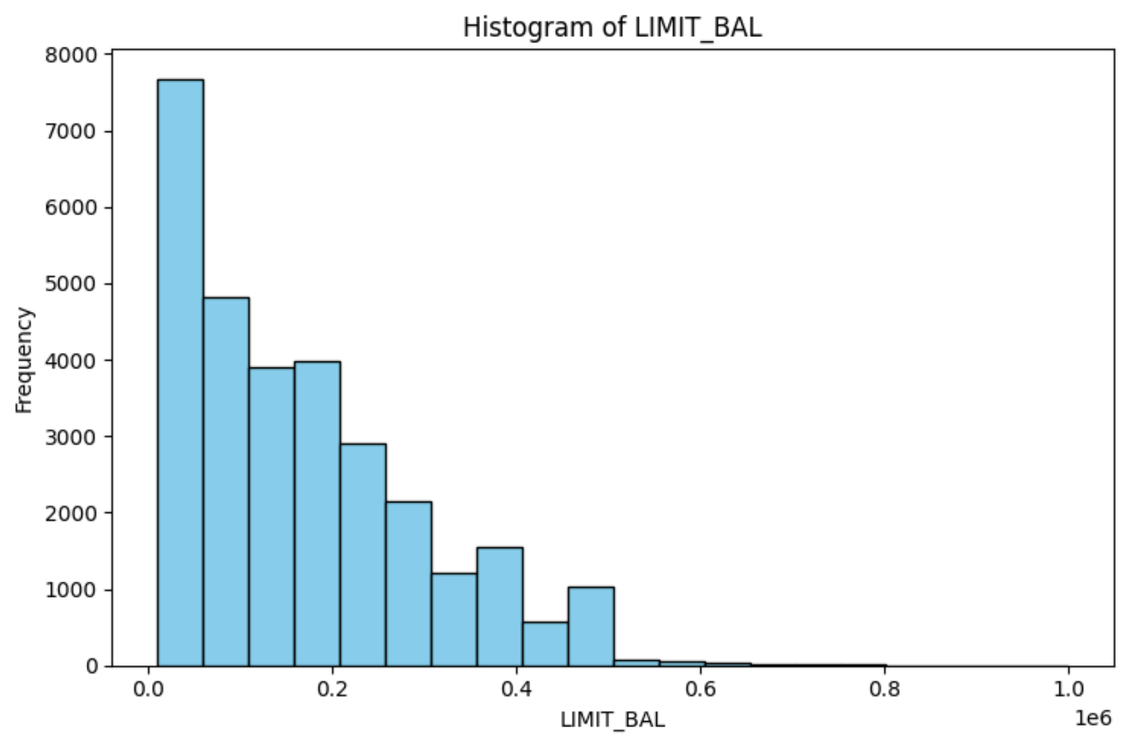


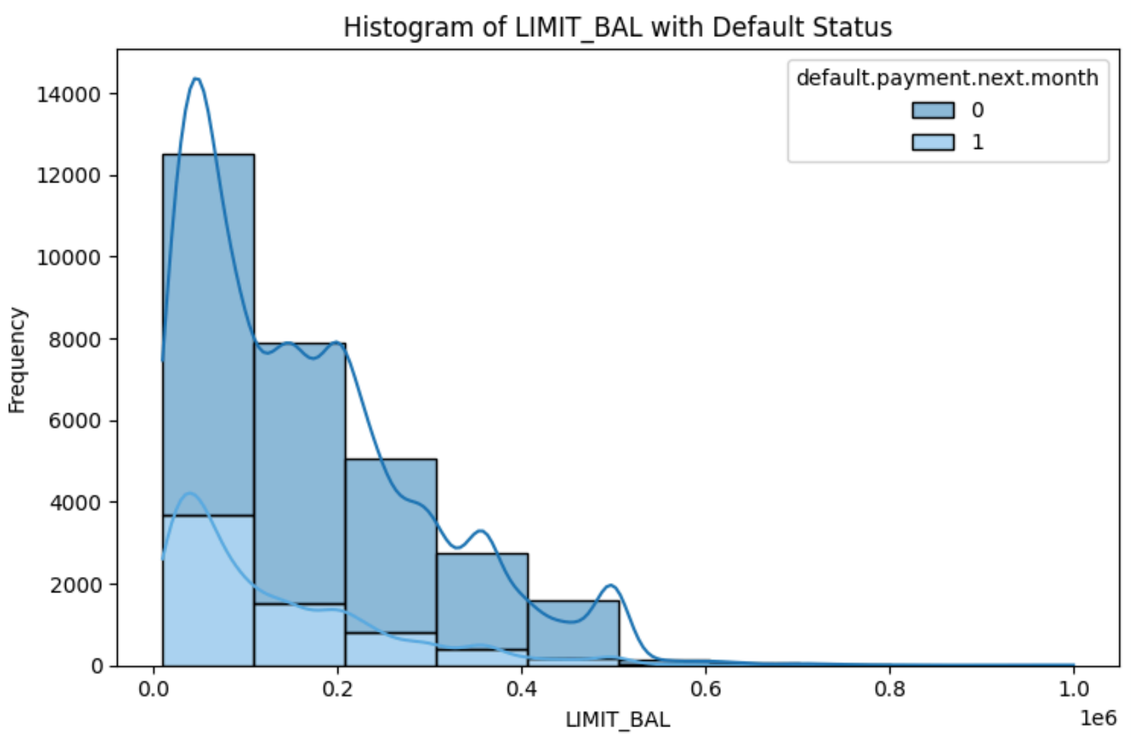
Conclusion :

PAY_SEPT - PAY_APRIL est similaire lorsqu'on les compare avec defaut de paiement : Il y a plus de non-défaillants que de défaillants. Cependant, nous avons remarqué que lorsque les détenteurs de cartes retardaient le paiement de 2 mois ou plus, il y avait légèrement plus de défaillants que de non-défaillants.

Analyse univariée et bivariée des données :

2.Distribution univariée et bivariéé du LIMIT_BALL :



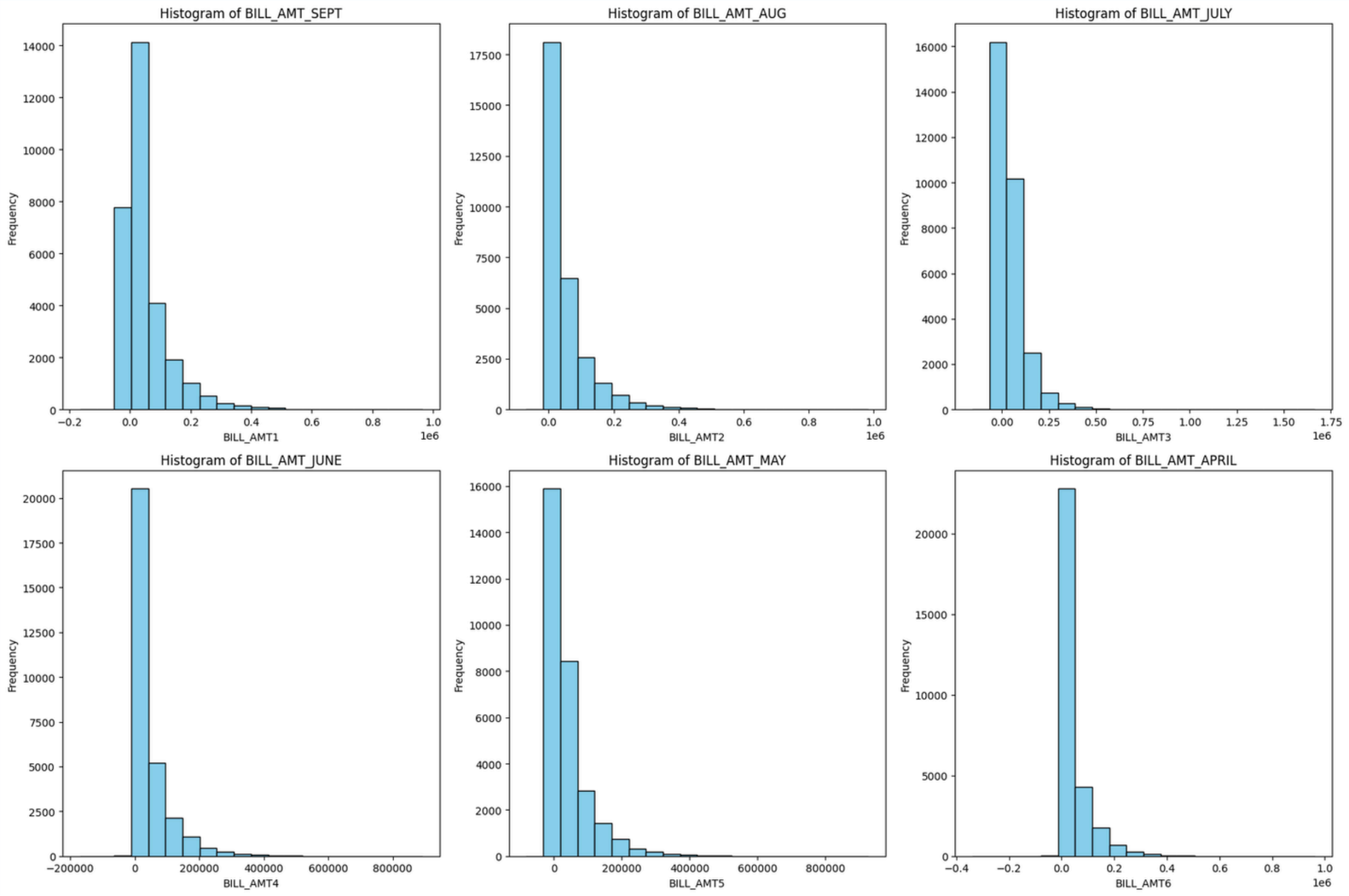


Conclusion :

On remarque que le pourcentage des défaillants diminue légèrement lorsque le montant de crédit diminue

Analyse univariée et bivariée des données :

2.Distribution univariée du BILL_AMTsep - BILL_AMTapril :

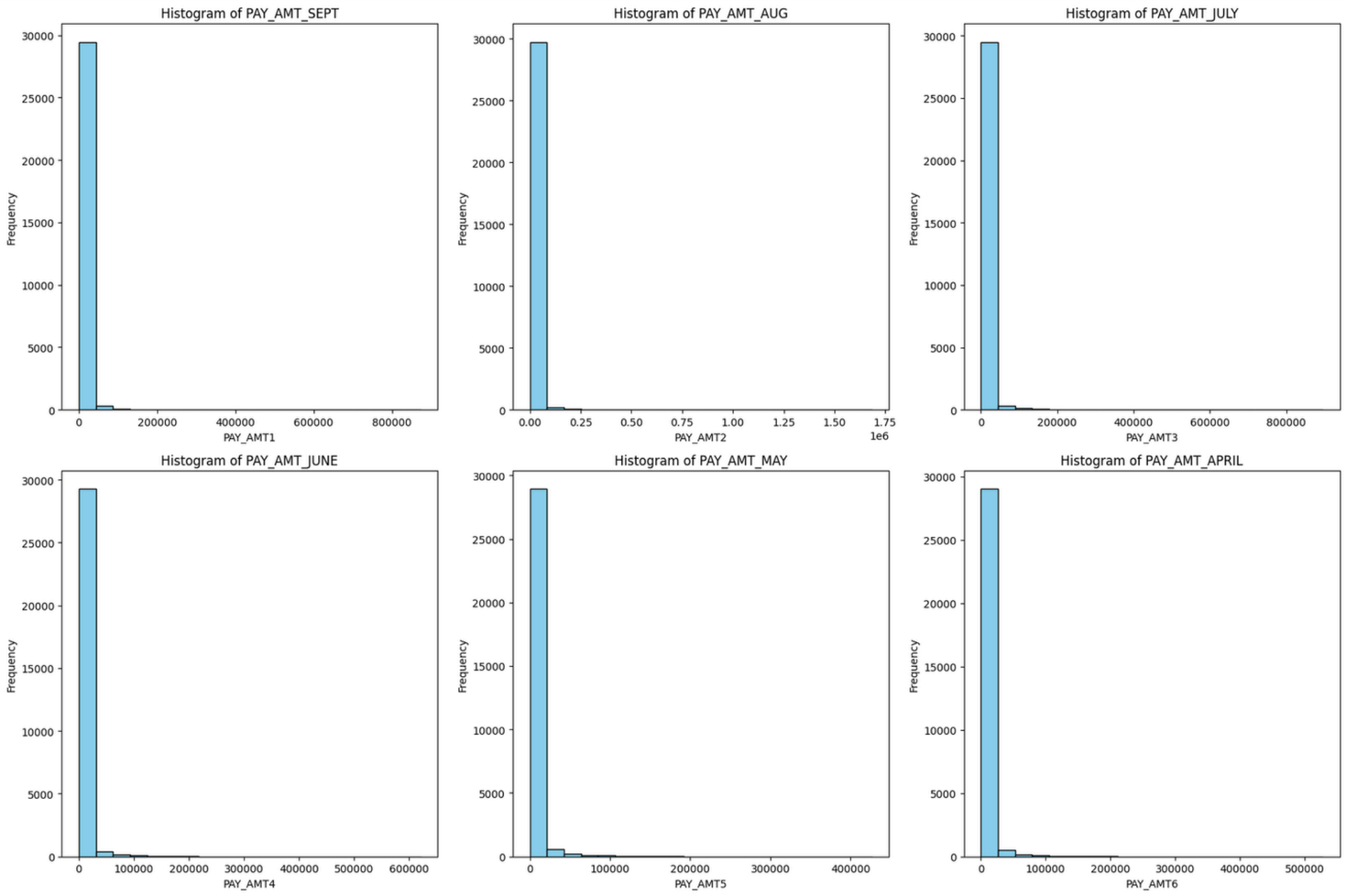


Conclusion :

on a des distributions décalées. certains montants de factures présentent des valeurs négatives, reflétant des soldes créditeurs ou des trop-payés

Analyse univariée et bivariée des données :

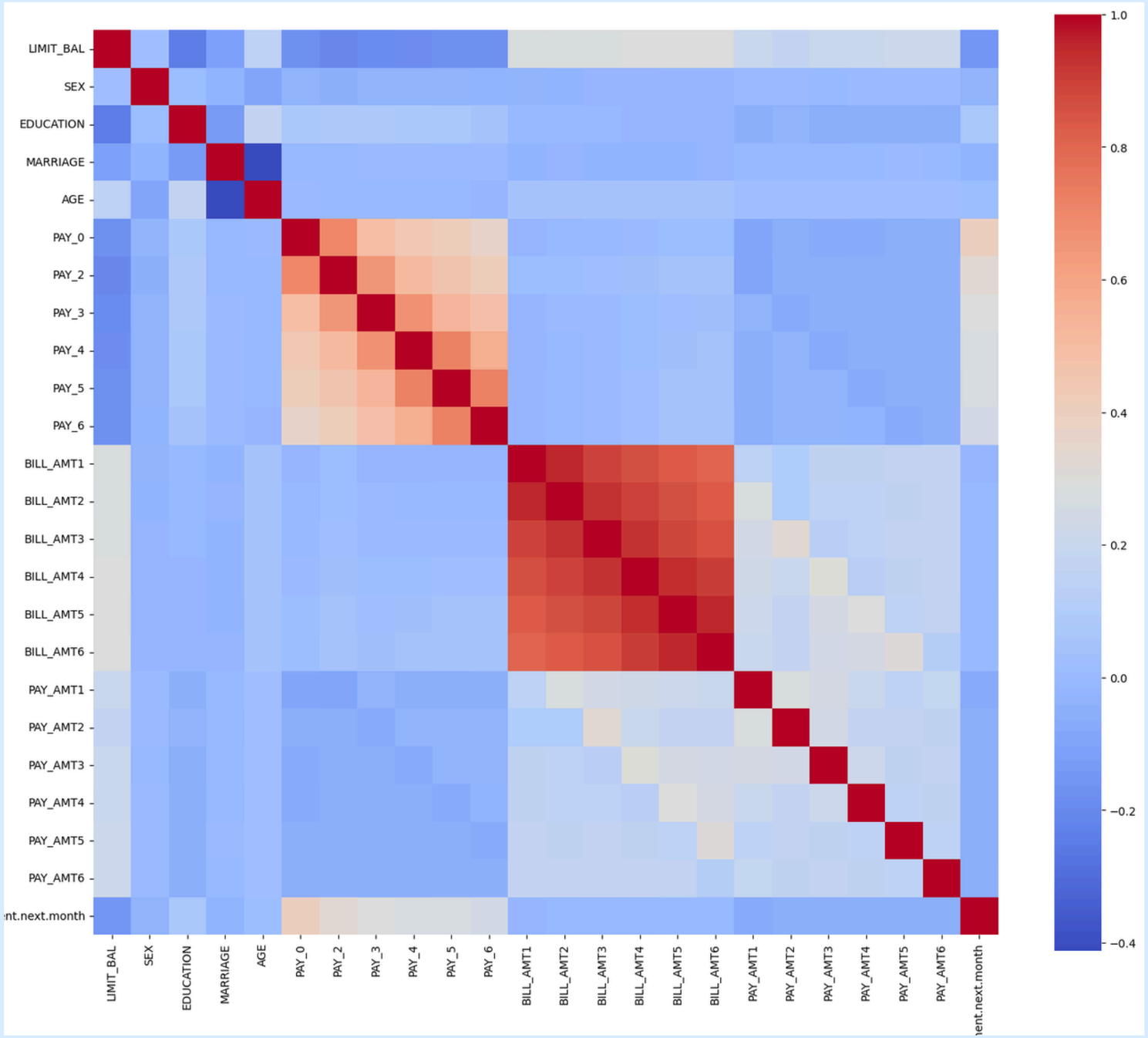
2.Distribution univariée du PAY_AMT_sep - PAY_AMT_april :



Conclusion :

Les montants du paiement précédent montrent une forte asymétrie pour tous les mois, ce qui indique des distributions significativement décalées.

Matrice de corrélation :



Conclusion :

Les montants des factures sur plusieurs mois sont fortement corrélés entre eux. Les montants des paiements sur plusieurs mois sont légèrement corrélés entre eux.

Modélisation :

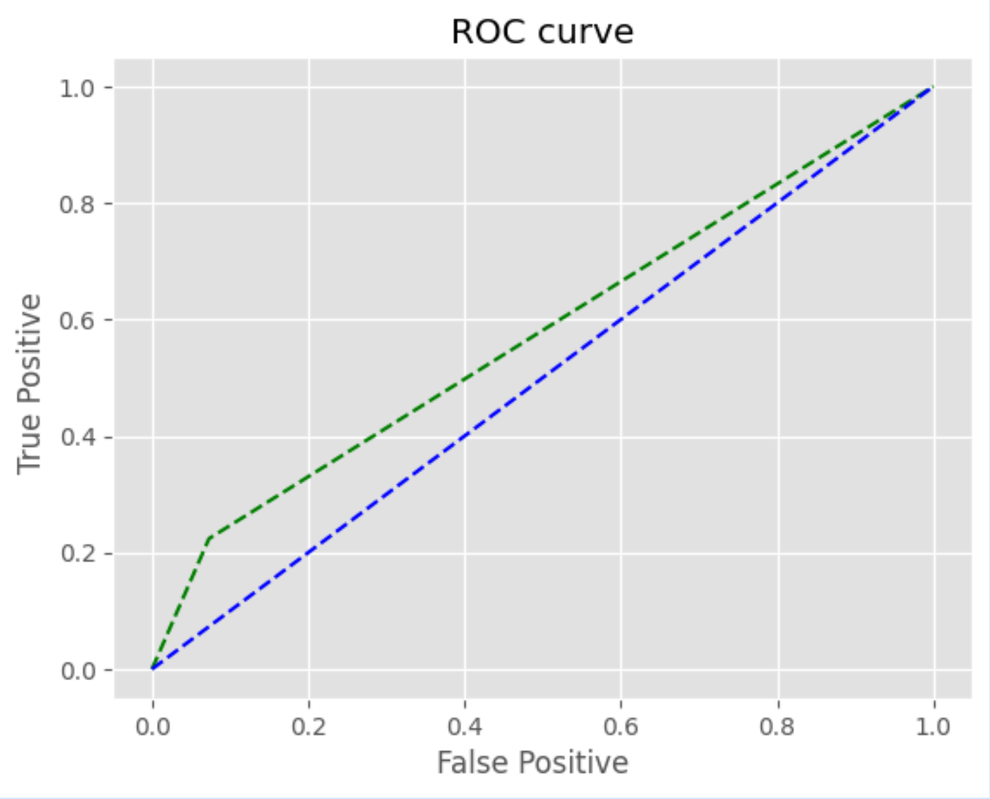
Retour à l'ordre du jour

		Régression logistique	Random forest	XGboost
Avant normali-sation	Accuracy	0.78	0.82	0.81
	Matrice de confusion	[[4686 1] [1311 2]]	[[4420 267] [865 448]]	[[4339 348] [847 466]]
Aprés normali-sation	Accuracy	0.78	0.78	0.77
	Matrice de confusion	[[4687 0] [1313 0]]	[[4433 254] [1041 272]]	[[4345 342] [1019 294]]

Conclusion:

Je vois que les deux bons modèles ici sont Random Forest et XGBoost qui ont atteint une précision de 82% (Sans normalisation) .

```
Gini Index (XGBoost): 0.4053486243406965
```



Retour à l'ordre du jour

MERCI DE VOTRE ATTENTION

Retour à l'ordre du jour