



# Détection de fraude par carte de crédit

Présenté par Fatou Lagnane  
Diop

# Sommaire

01

Introduction

02

Modèle de  
classification

03

Application sur les  
données réelles

04

Analyses exploratoire et  
Multivariée

05

Modélisation prédictive  
et Evaluation

06

Discussion

07

Conclusion et Perspectives

# Introduction

Face à l'augmentation du volume de transactions électronique, le potentiel d'exploitation des vulnérabilités des systèmes de paiement par les acteurs frauduleux ne cesse d'augmenter. La fraude à la carte bancaire représente un enjeu croissant pour les entreprises du monde entier. Selon un rapport Nilson, les pertes associées se sont élevées à 28.65 milliards de dollars en 2021, d'où la nécessité pour les entreprises d'apprendre à détecter et à prévenir ce type de fraude.

# Introduction

## Contexte

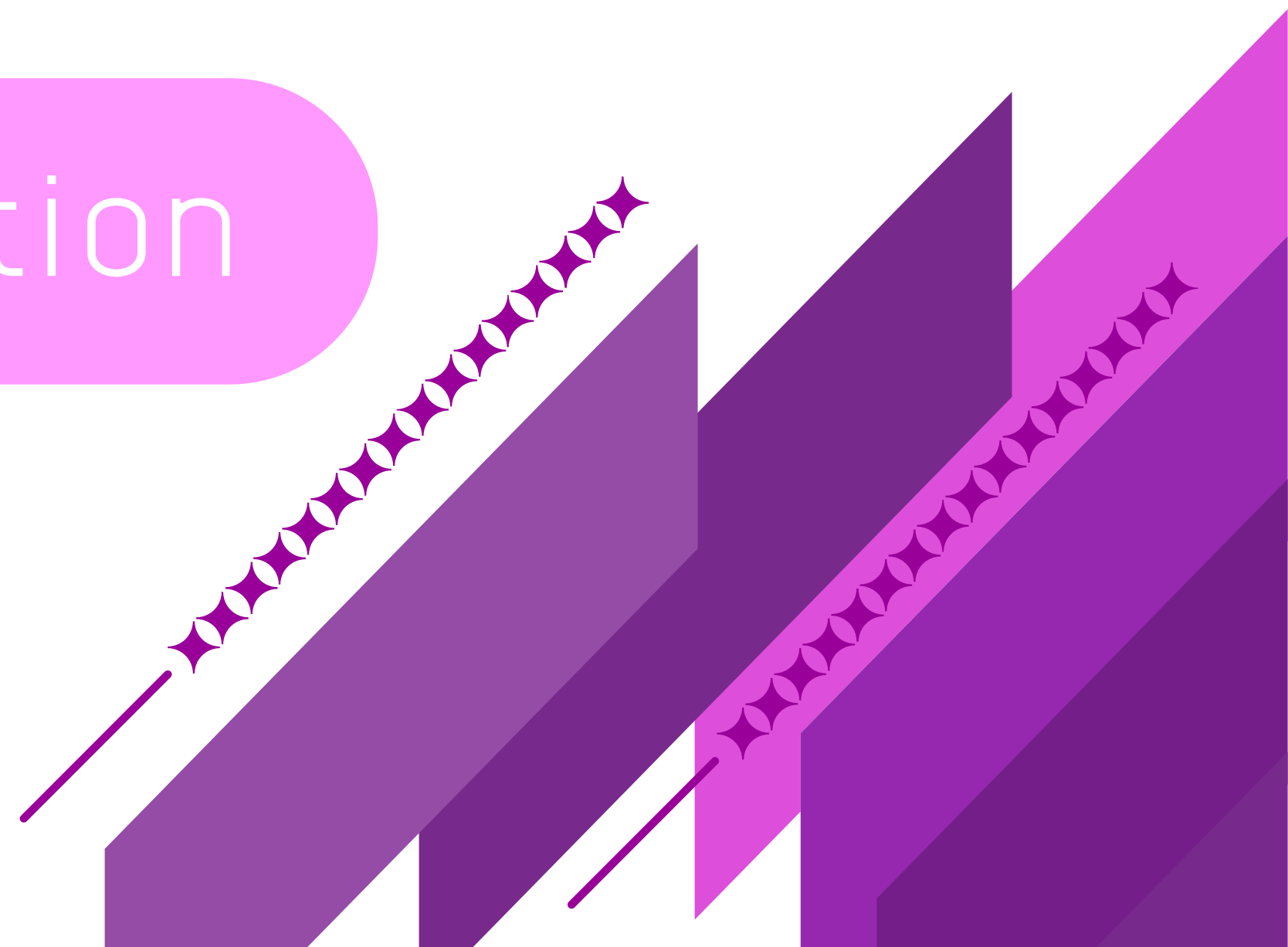
L'évolution et la complexité des risques de fraude nécessitent le recours à des moyens technologiques puissants pour renforcer les contrôles et assurer la précision des processus innovants en matière de détection et du filtrage des opérations suspectes. Sur ce, il est important que les sociétés de cartes de crédit soient en mesure de reconnaître les transactions frauduleuses par carte de crédit afin que les clients ne soient pas facturés pour des articles qu'ils n'ont pas achetés.

# Introduction

## Objectif

Comment peut-on classer les transactions frauduleuses et légitimes afin de proposer un mécanisme de prévention des risques de fraude au niveau des sociétés de cartes de crédit?

Modèle de classification





# Modèle de classification

## Régression logistique

La régression logistique est couramment utilisée pour résoudre les problèmes en matière de prédiction et de classification. Elle permet d'étudier les relations entre:

- $Y$  -> variable qualitative dichotomique ( prend 0 absence d'occurrence ou 1 présence d'occurrence)
- $X$  -> variables qualitatives binaires ou quantitatives



# Régression logistique

## Modèle

Le modèle de régression logistique consiste donc à caractériser la loi de  $Y|X=x$  par une loi de Bernoulli de paramètre  $p(x)$  telque :

$$P(Y = 1/X = x) = \pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)} \quad (1)$$

$$\text{logit}(p(x)) = \log \left( \frac{p(x)}{1 - p(x)} \right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (2)$$





# Régression logistique

Test: métriques d'évaluation

**True positive**

On prédit vrai  
et la réponse est  
vraie

**TP**

**True negative**

On prédit faux et  
la réponse est  
fausse

**TN**

**False positive**

On prédit vrai et  
la réponse est  
fausse

**FP**

**False negative**

On prédit faux et  
la réponse est  
vraie

**FN**

# Régression logistique

Test: métriques d'évaluation

## Matrice de confusion

Elle permet de mesurer la qualité d'un problème de classification à l'aide d'une matrice.

		<i>Reality</i>	
<b>Confusion matrix</b>		<b>Negative : 0</b>	<b>Positive : 1</b>
<i>Prediction</i>	<b>Negative : 0</b>	<b>True Negative : TN</b>	<b>False Negative : FN</b>
	<b>Positive : 1</b>	<b>False Positive : FP</b>	<b>True Positive : TP</b>



# Régression logistique

Test: métriques d'évaluation

## Accuracy

Est le quotient entre le nombre de prédictions correctes et le nombre total de prédictions

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

## Precision

Permet de connaître le pourcentage de prédictions positives bien effectuées

$$Precision = \frac{TP}{TP + FP}$$

## Recall

Permet de connaître le pourcentage de positifs bien prédits par le modèle

$$Recall = \frac{TP}{TP + FN}$$

## F1\_score

Est la moyenne harmonique entre la precision et le rappel

$$F_1 = 2 \cdot \frac{Precision \cdot recall}{Precision + recall}$$

## Receiver Operating Characteristic (ROC) Curve

Courbe qui trace le taux de vrais positifs en fonction du taux de faux positifs.

Si le seuil de prédictions augmente le taux de TP augmente de même que le taux de FP.

## Area Under Curve (AUC)

C'est une métrique de classification et permet de représenter l'aire sous la courbe ROC ou de Precision-Recall.

L'AUC est un nombre réel entre 0 et 1. Plus cette aire est élevée meilleur est le modèle.

## Precision-Recall Curve

La courbe de précision-rappel est un outil essentiel pour comprendre comment un modèle de classification binaire équilibre la précision et le rappel à travers différents seuils de décision, offrant ainsi une perspective approfondie sur ses performances, surtout dans des situations de déséquilibre des classes.

Application



# Problème:

Comment peut-on classer les transactions frauduleuses et légitimes afin de proposer un mécanisme de prévention des risques de fraude au niveau des sociétés de cartes de crédit?





# Présentation des données

L'ensemble de données contient les transactions effectuées par carte de crédit en septembre 2013 par des titulaires de cartes européens.

Cet ensemble de données présente les transactions qui ont eu lieu en deux jours, où nous avons **492** fraudes sur **284 807** transactions, soit environ **0.173%**, ce qui montre un déséquilibre de classes.

La base est composée de **31** variables dont la variable **Class** qui est notre variable réponse et prend la valeur 1 en cas de fraude et 0 sinon. Et le reste sont des variables numériques et explicatives.



# Base de données

	Time	V1	V2	V3	V4	V5	V6	V7	\
0	0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	
1	0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	
2	1	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	
3	1	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	
4	2	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	

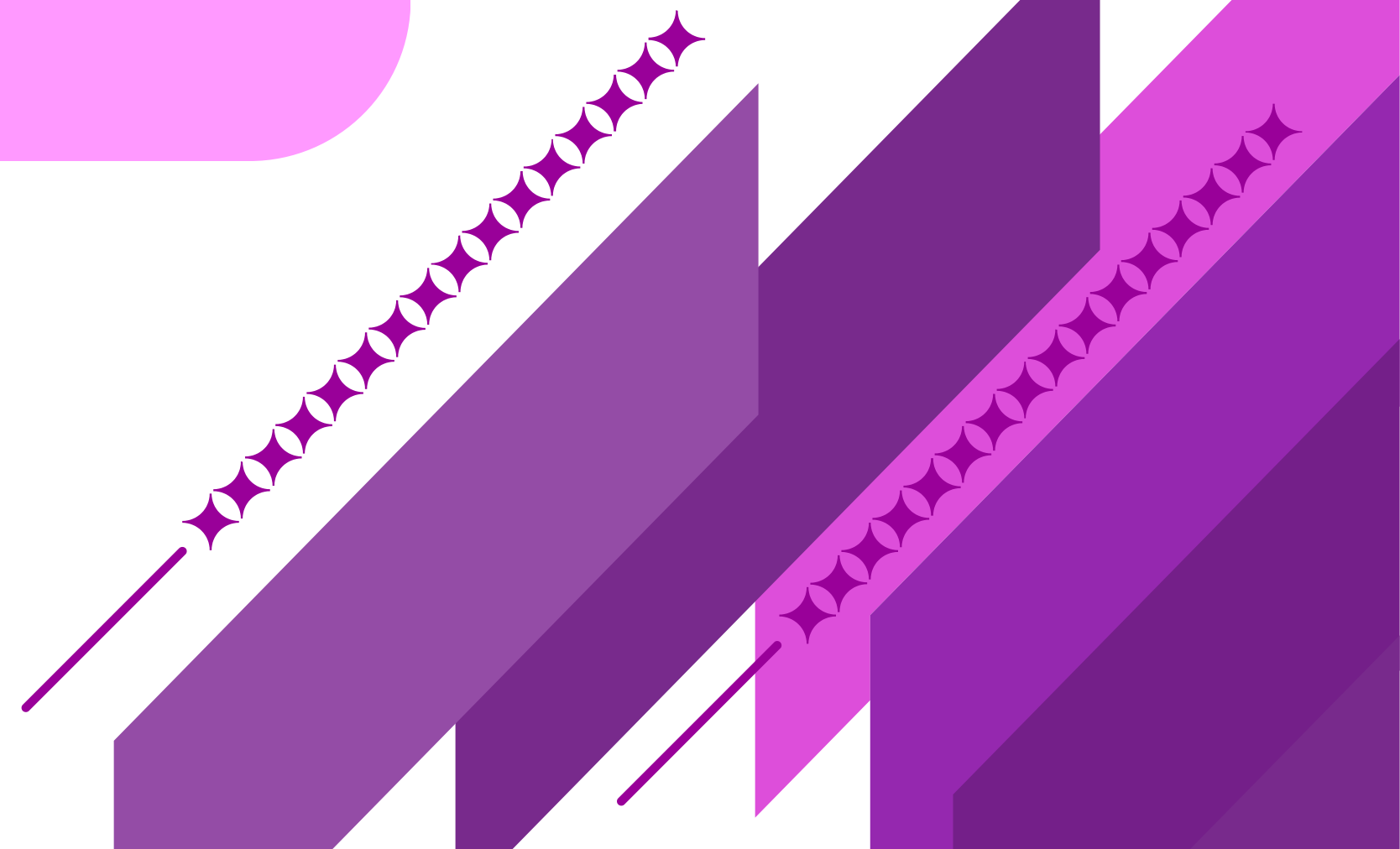
	V8	V9	...	V21	V22	V23	V24	V25	\
0	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	
1	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	
2	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	
3	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	
4	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	

	V26	V27	V28	Amount	Class
0	-0.189115	0.133558	-0.021053	149.62	0
1	0.125895	-0.008983	0.014724	2.69	0
2	-0.139097	-0.055353	-0.059752	378.66	0
3	-0.221929	0.062723	0.061458	123.50	0
4	0.502292	0.219422	0.215153	69.99	0

[5 rows x 31 columns]



# Analyses exploratoire et multivariée





# Description des données

	Time	Class	Amount	V1	V2	V3
<b>count</b>	284807.000000	284807.000000	284807.000000	2.848070e+05	2.848070e+05	2.848070e+05
<b>mean</b>	94813.859575	0.001727	88.349619	1.168375e-15	3.416908e-16	-1.379537e-15
<b>std</b>	47488.145955	0.041527	250.120109	1.958696e+00	1.651309e+00	1.516255e+00
<b>min</b>	0.000000	0.000000	0.000000	-5.640751e+01	-7.271573e+01	-4.832559e+01
<b>25%</b>	54201.500000	0.000000	5.600000	-9.203734e-01	-5.985499e-01	-8.903648e-01
<b>50%</b>	84692.000000	0.000000	22.000000	1.810880e-02	6.548556e-02	1.798463e-01
<b>75%</b>	139320.500000	0.000000	77.165000	1.315642e+00	8.037239e-01	1.027196e+00
<b>max</b>	172792.000000	1.000000	25691.160000	2.454930e+00	2.205773e+01	9.382558e+00

# Analyse univariée

Description de variable réponse

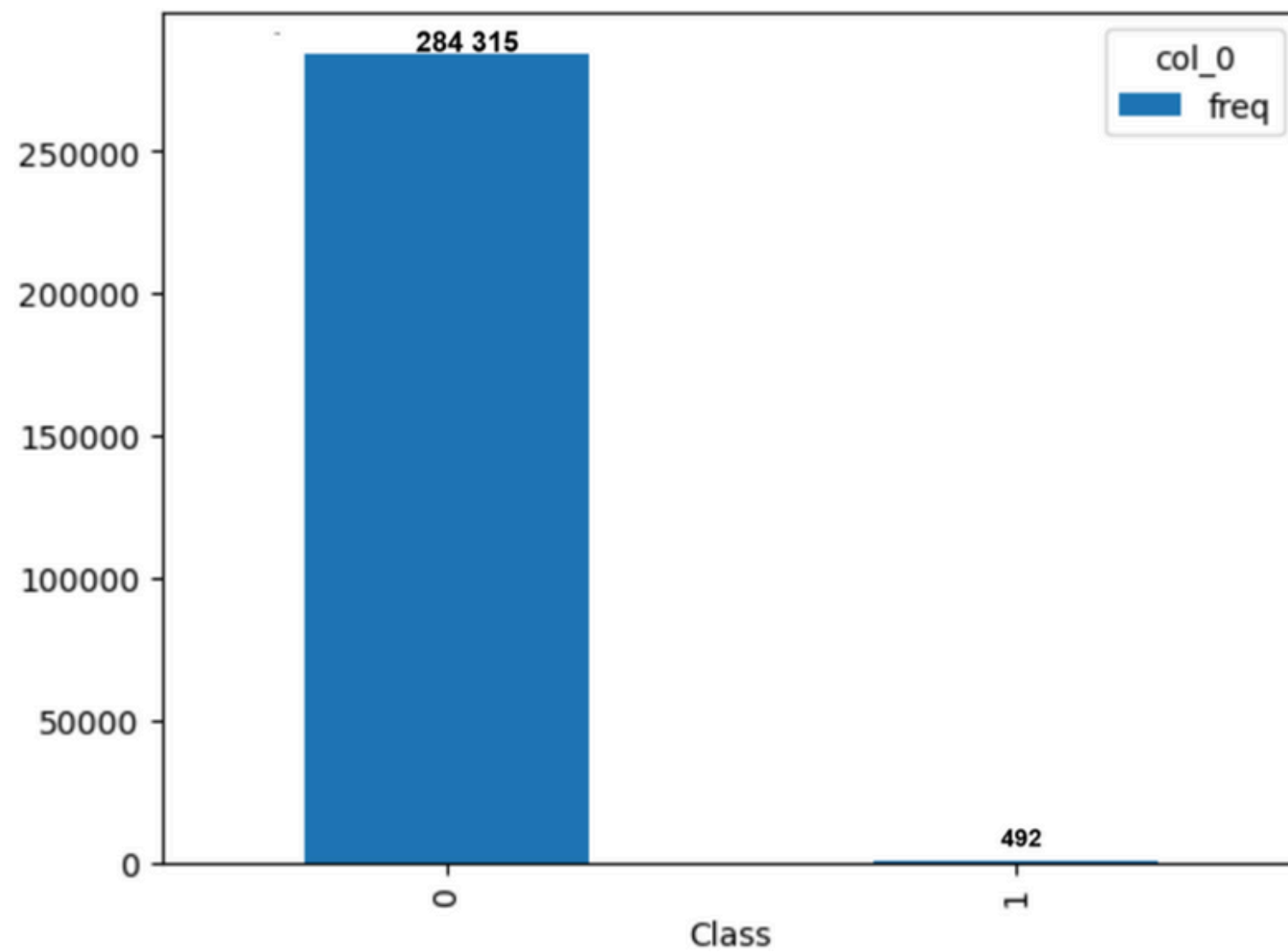


Figure 1

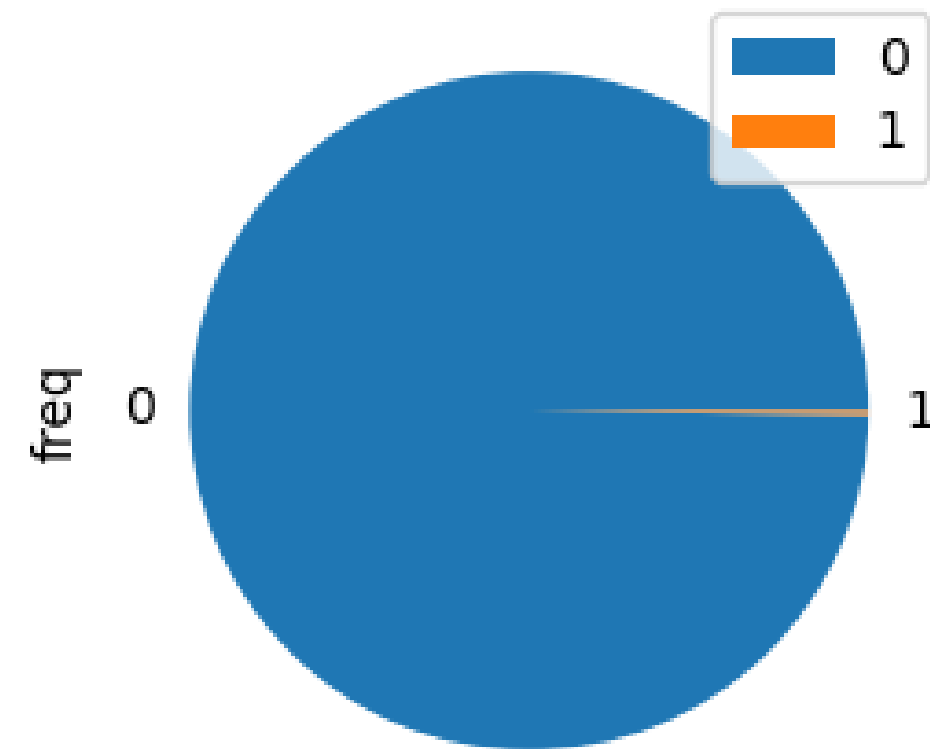
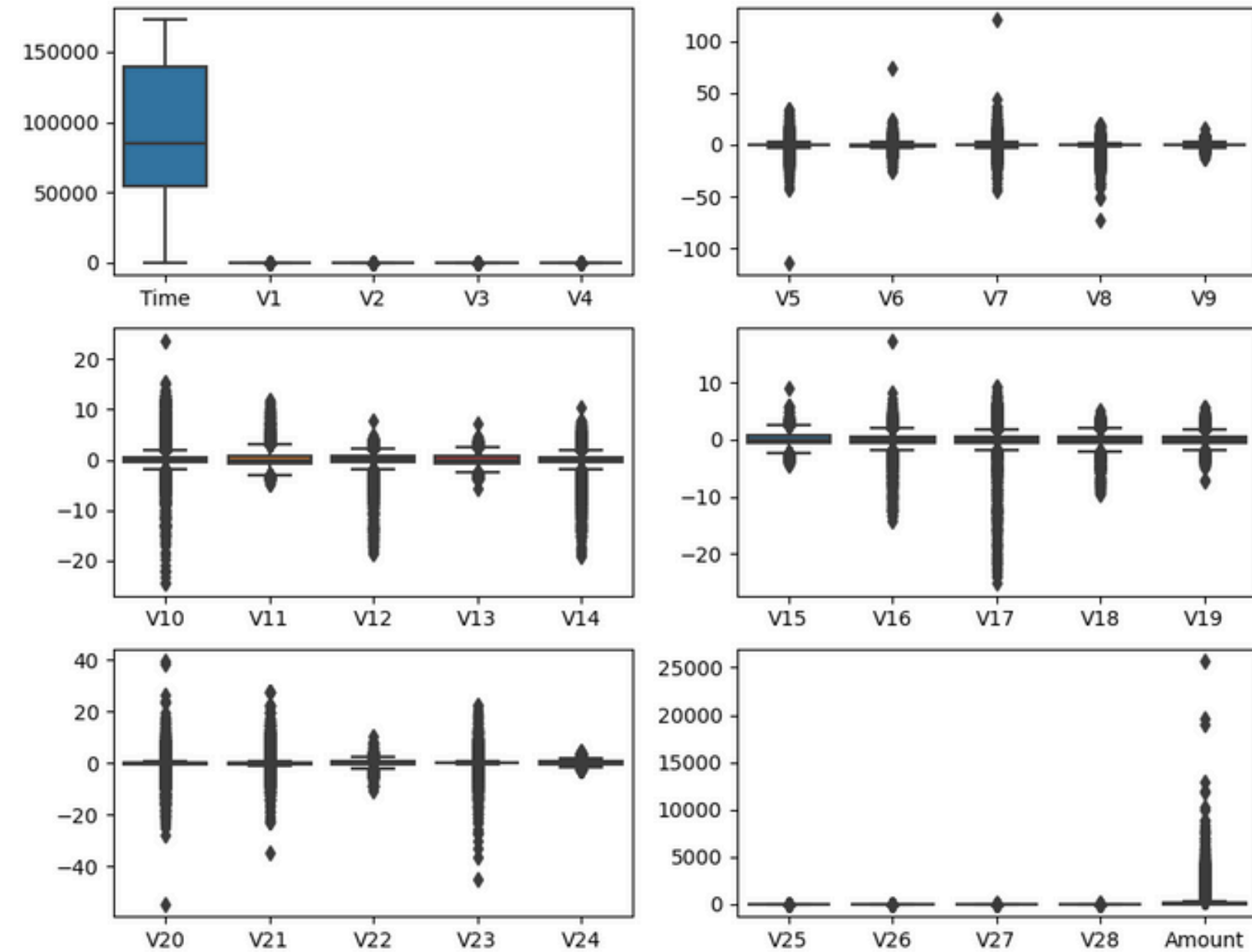


Figure 2

# Analyse univariée

## Description de features variables: Boxplots

- On constate des valeurs extrêmes au niveau des composantes V1 à V28 et du montant.
- Puisque ces variables ont été obtenues à partir d'une étude préalable en ACP, on préfère alors garder ces valeurs extrêmes.



# Analyse bivariable

Description conjointe des features

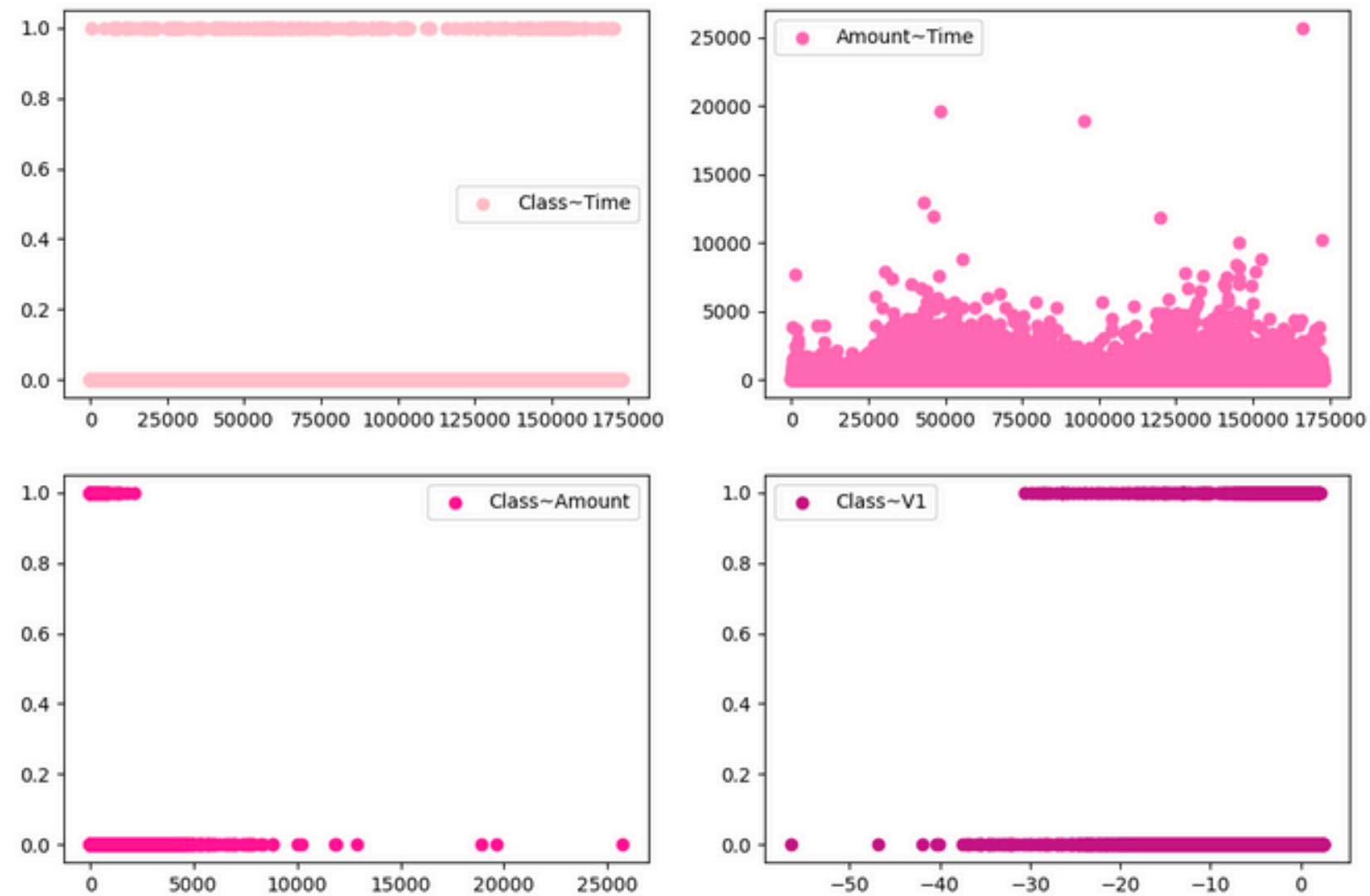


Figure 5

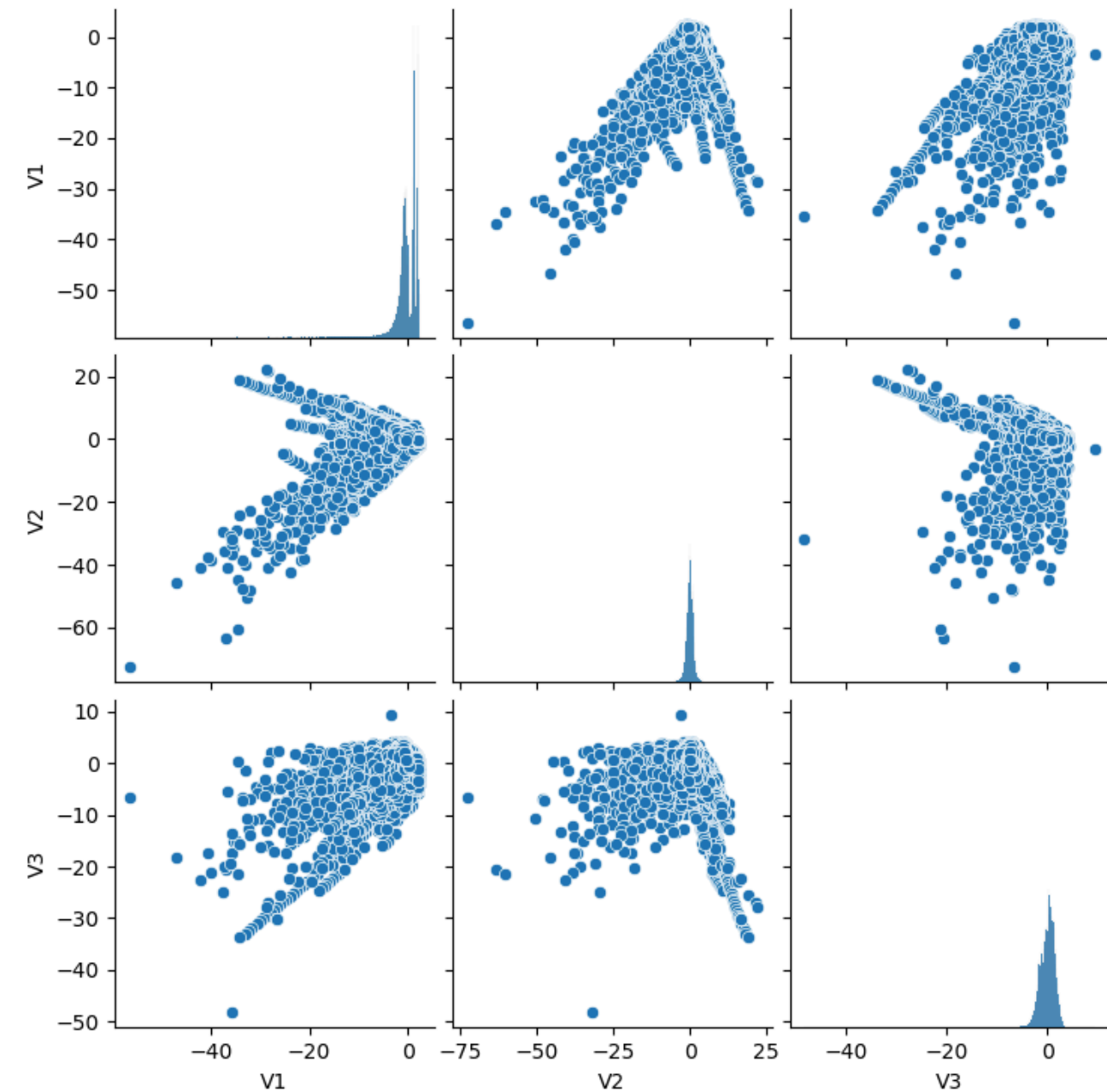
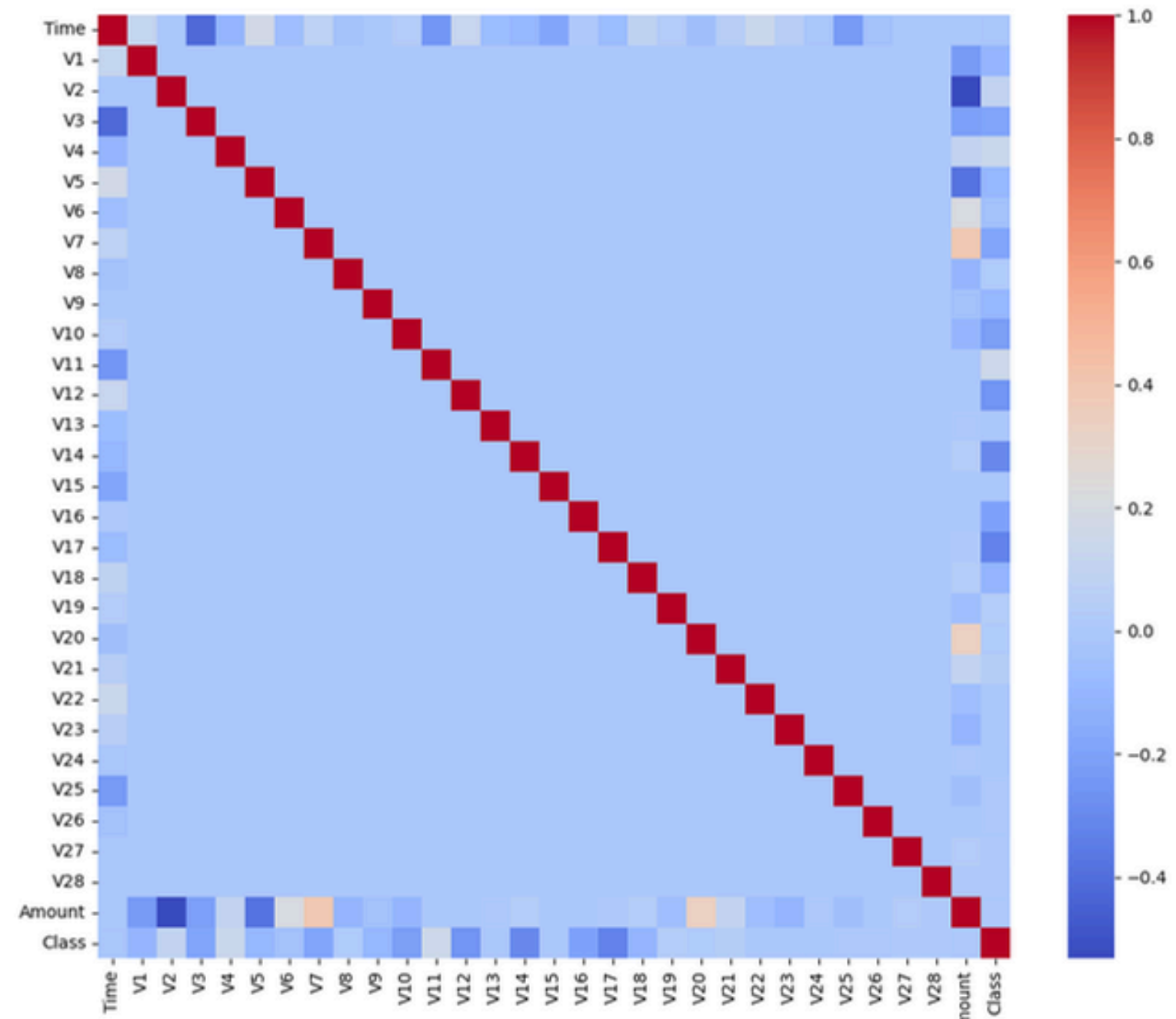


Figure 6

# Analyse multivariée

## Corrélation

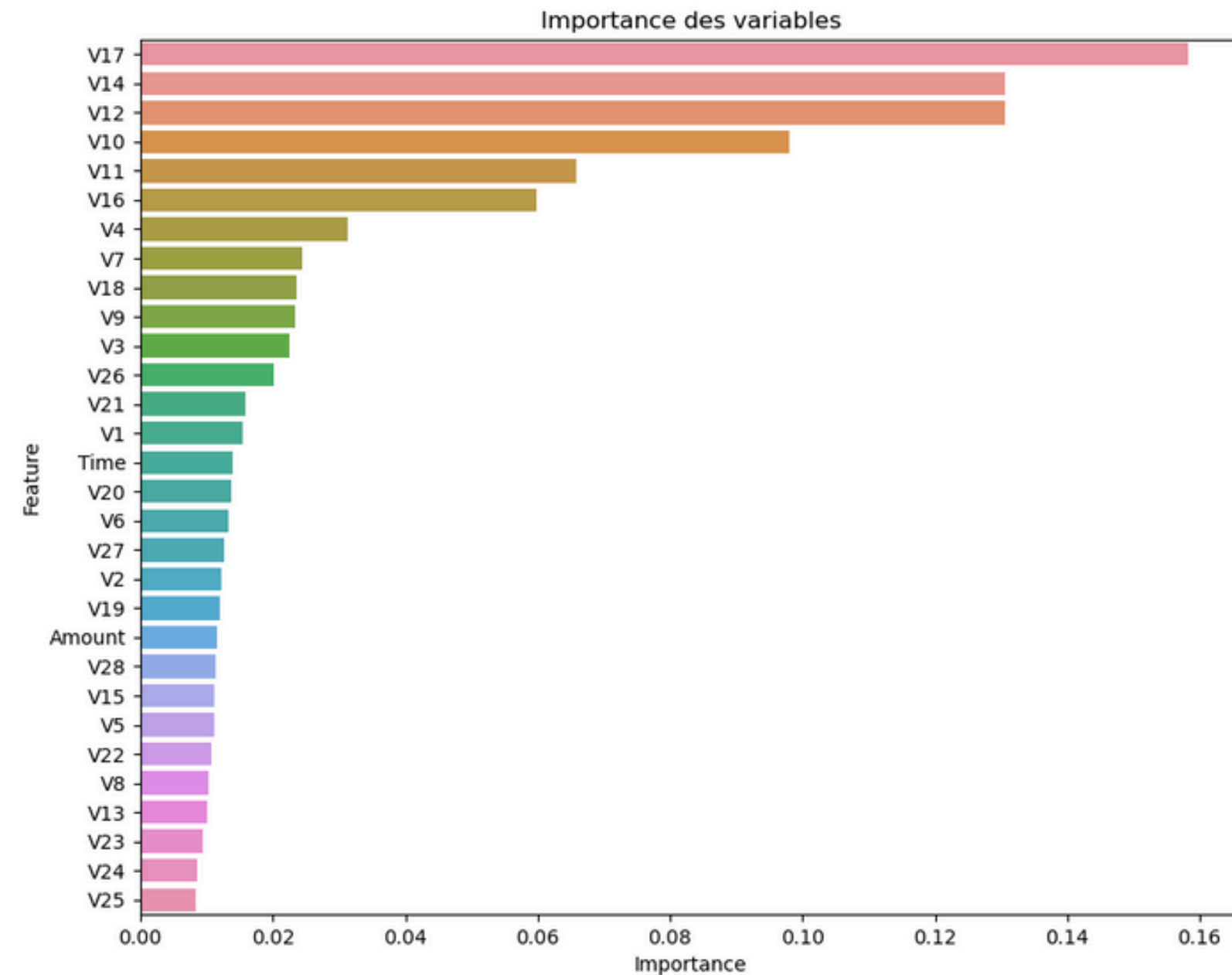
Les variables sont indépendantes les unes aux autres donc non corrélées.



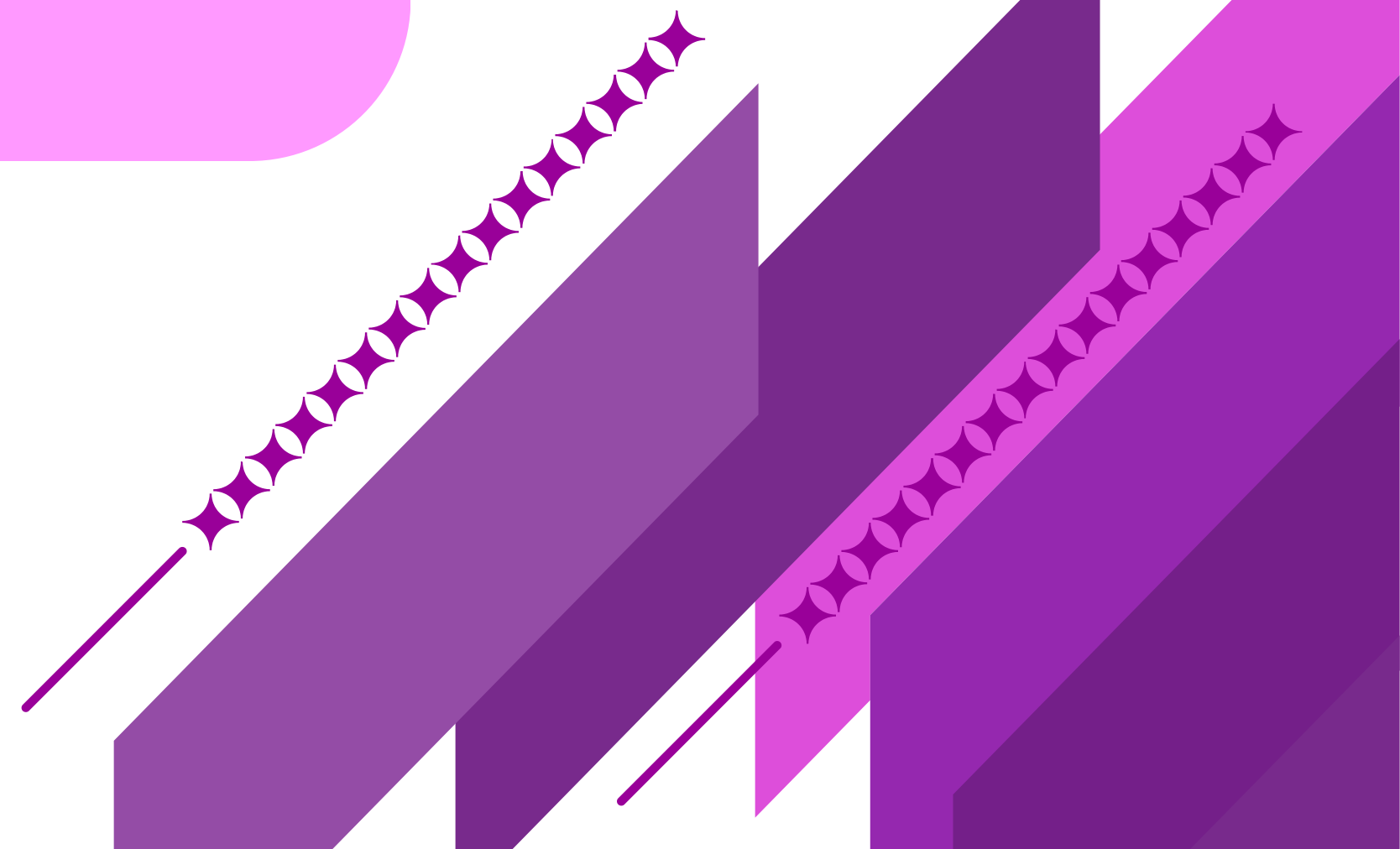
# Analyse multivariée

## Sélection de variables

- Les variables les plus pertinentes sont déterminées en utilisant l'algorithme d'un modèle d'arbre décisionnel.
- Pour chaque variable, on calcule une mesure de qualité de division pour déterminer à quel point elle est efficace pour séparer les classes ou réduire la variance(régression).
- Après avoir testées toutes les variables disponibles, on choisit celle qui maximise la qualité de la division (ou minimise un critère d'erreur).



# Modélisation prédictive et Evaluation







# Modélisation prédictive

## Régression logistique

$$\text{logit ( Class = 1 / X = x )} = \text{beta X ;}$$

avec **beta** un vecteur et **X** une matrice (284 807 , 31)

# Modélisation prédictive

## Régression logistique: Résumé du modèle

Il serait difficile de se baser sur ces coefficients pour interpréter notre modèle. Sur ce, il serait question d'évaluer la performance de notre modèle à partir des métriques de classification.

Logit Regression Results						
=====						
Dep. Variable:	Class	No. Observations:	227845			
Model:	Logit	Df Residuals:	227815			
Method:	MLE	Df Model:	29			
Date:	Thu, 09 Jan 2025	Pseudo R-squ.:	0.2202			
Time:	13:52:06	Log-Likelihood:	-2330.4			
converged:	True	LL-Null:	-2988.3			
Covariance Type:	nonrobust	LLR p-value:	8.975e-259			
=====						
	coef	std err	z	P> z	[0.025	0.975]
-----						
Time	-9.386e-05	1.78e-06	-52.652	0.000	-9.74e-05	-9.04e-05
V1	0.6225	0.037	16.723	0.000	0.550	0.696
V2	-1.3127	0.044	-30.101	0.000	-1.398	-1.227
V3	-1.5564	0.038	-41.343	0.000	-1.630	-1.483
V4	0.3460	0.033	10.448	0.000	0.281	0.411
V5	-0.7417	0.061	-12.249	0.000	-0.860	-0.623
V6	0.3668	0.047	7.859	0.000	0.275	0.458
V7	1.3396	0.068	19.798	0.000	1.207	1.472
V8	-0.5401	0.026	-21.072	0.000	-0.590	-0.490
V9	-0.7137	0.058	-12.238	0.000	-0.828	-0.599
V10	-0.8804	0.070	-12.543	0.000	-1.018	-0.743

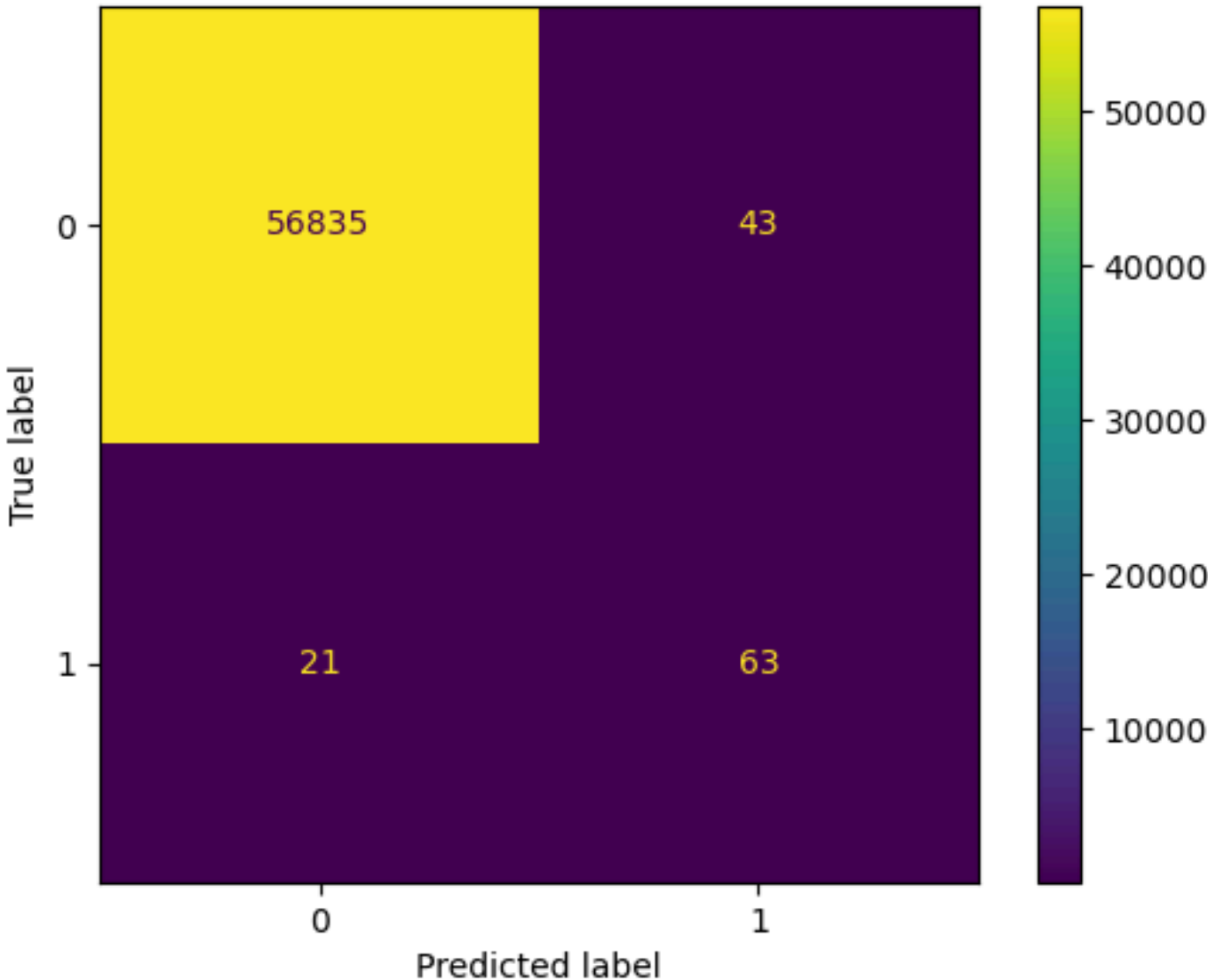
# Modélisation prédictive

## Rapport de classification et Matrice de confusion

**TN = 56835      FN = 21**

**FP = 43      TP = 63**

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56878
1	0.59	0.75	0.66	84
accuracy			1.00	56962
macro avg	0.80	0.87	0.83	56962
weighted avg	1.00	1.00	1.00	56962



# Modélisation prédictive

## Valeurs prédites en fonction des seuils de classification

Le seuil de classification détermine le seuil de valeur prédite qui, une fois dépassé, est considéré comme succès. Dans notre cas, pour une probabilité prédite ( $P=1$ ) supérieure à un seuil ( $S=0.84$ ), la réponse prédite est 1.

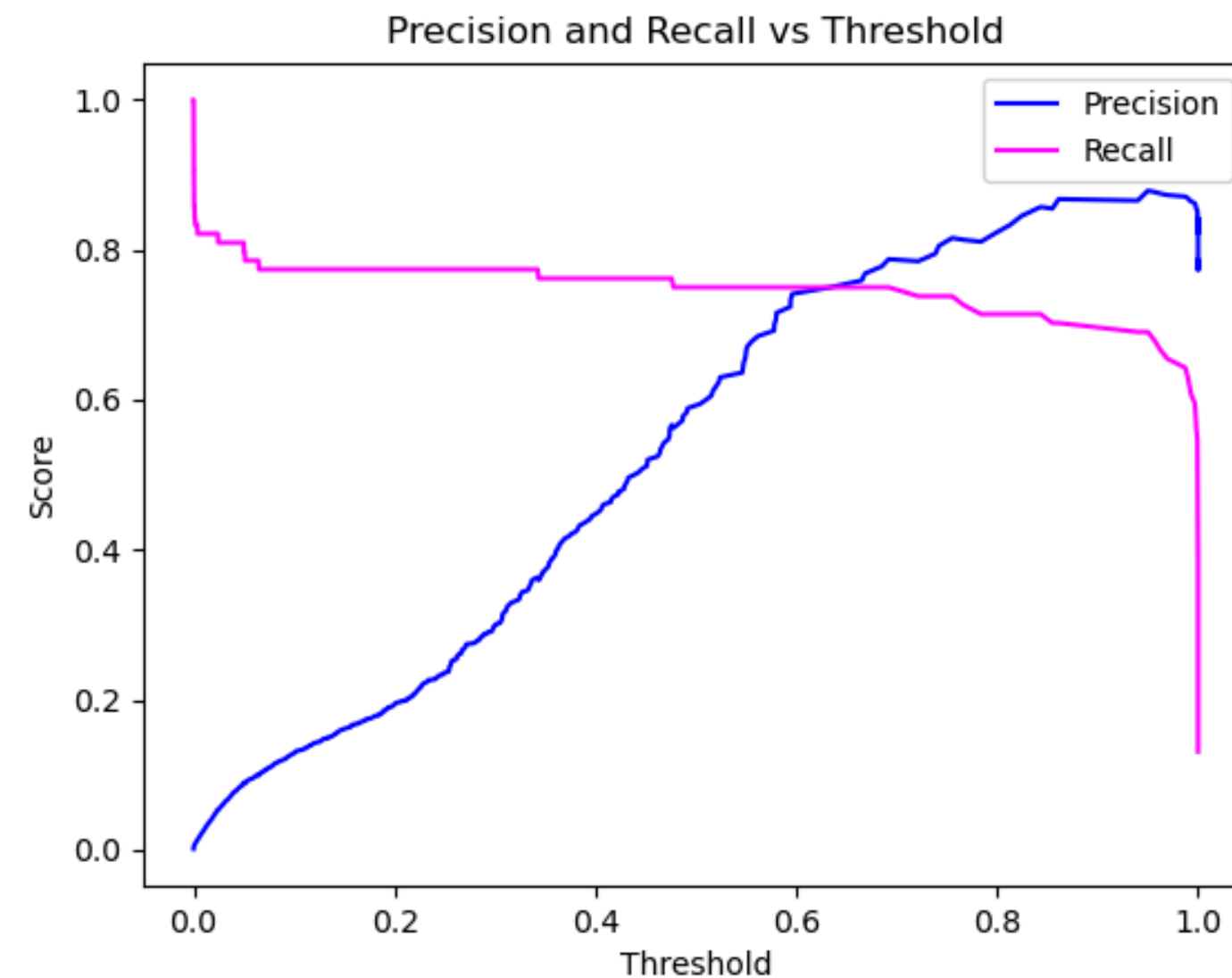
	True Value	Predicted Probability	S= 6e-30	S= 0.4	S= 0.6	S= 0.8	S= 0.84
10620	0	1.000000e+00	1	1	1	1	1
150601	1	1.000000e+00	1	1	1	1	1
9509	1	1.000000e+00	1	1	1	1	1
12156	0	1.000000e+00	1	1	1	1	1
42635	1	1.000000e+00	1	1	1	1	1
...	...	...	...	...	...	...	...
228525	0	5.948530e-27	1	0	0	0	0
91896	0	4.454920e-27	1	0	0	0	0
238412	0	3.773454e-27	1	0	0	0	0
240739	0	5.513707e-30	0	0	0	0	0
227921	0	1.332392e-40	0	0	0	0	0

56962 rows × 7 columns

# Modélisation prédictive

## Courbes de précision et de rappel en fonction des thresholds

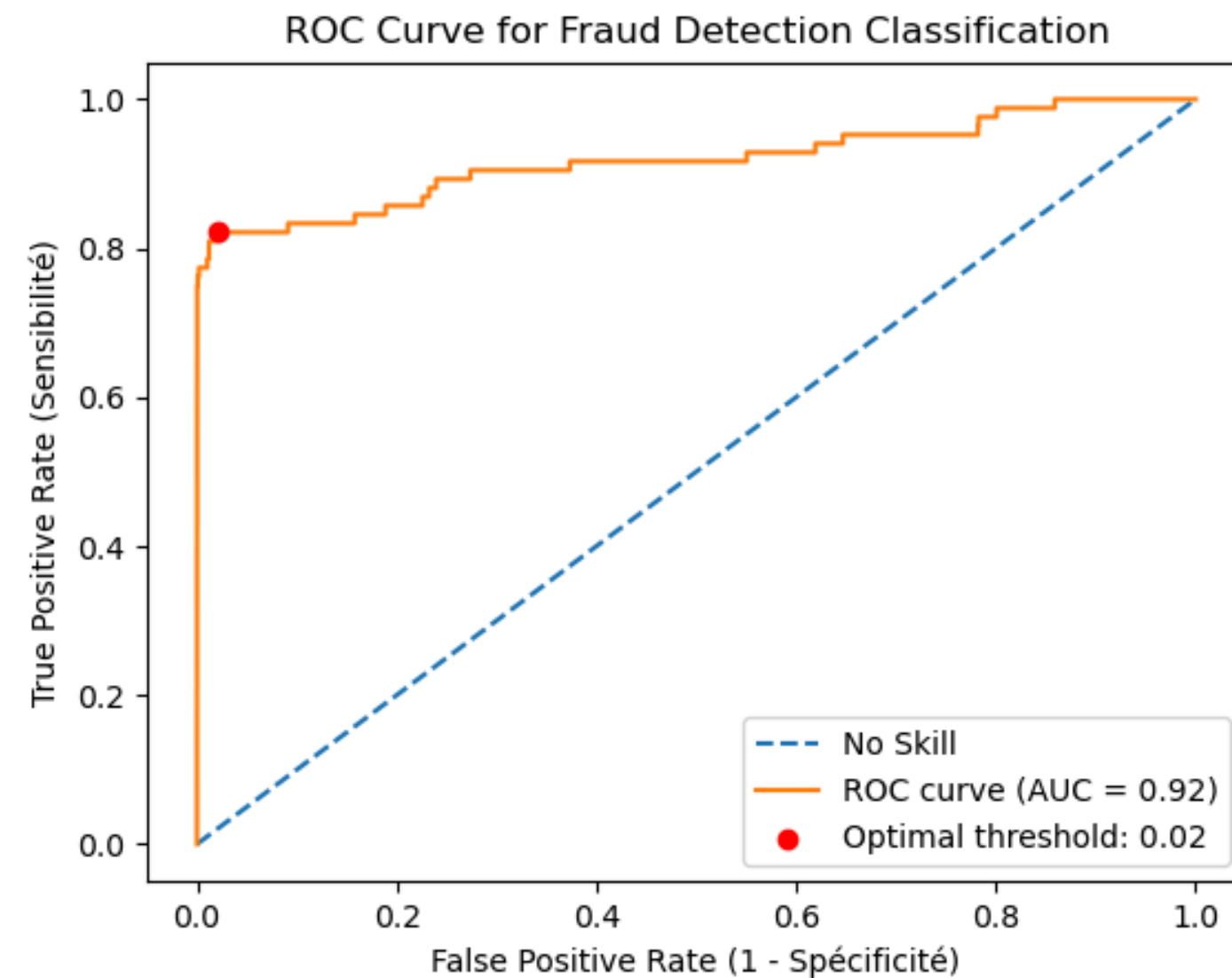
On constate un déphasage des courbes de précision et de rappel en fonction des seuils sauf à un point d'intersection où on a un seuil approximativement égal à 0.65



# Evaluation

## ROC Curve

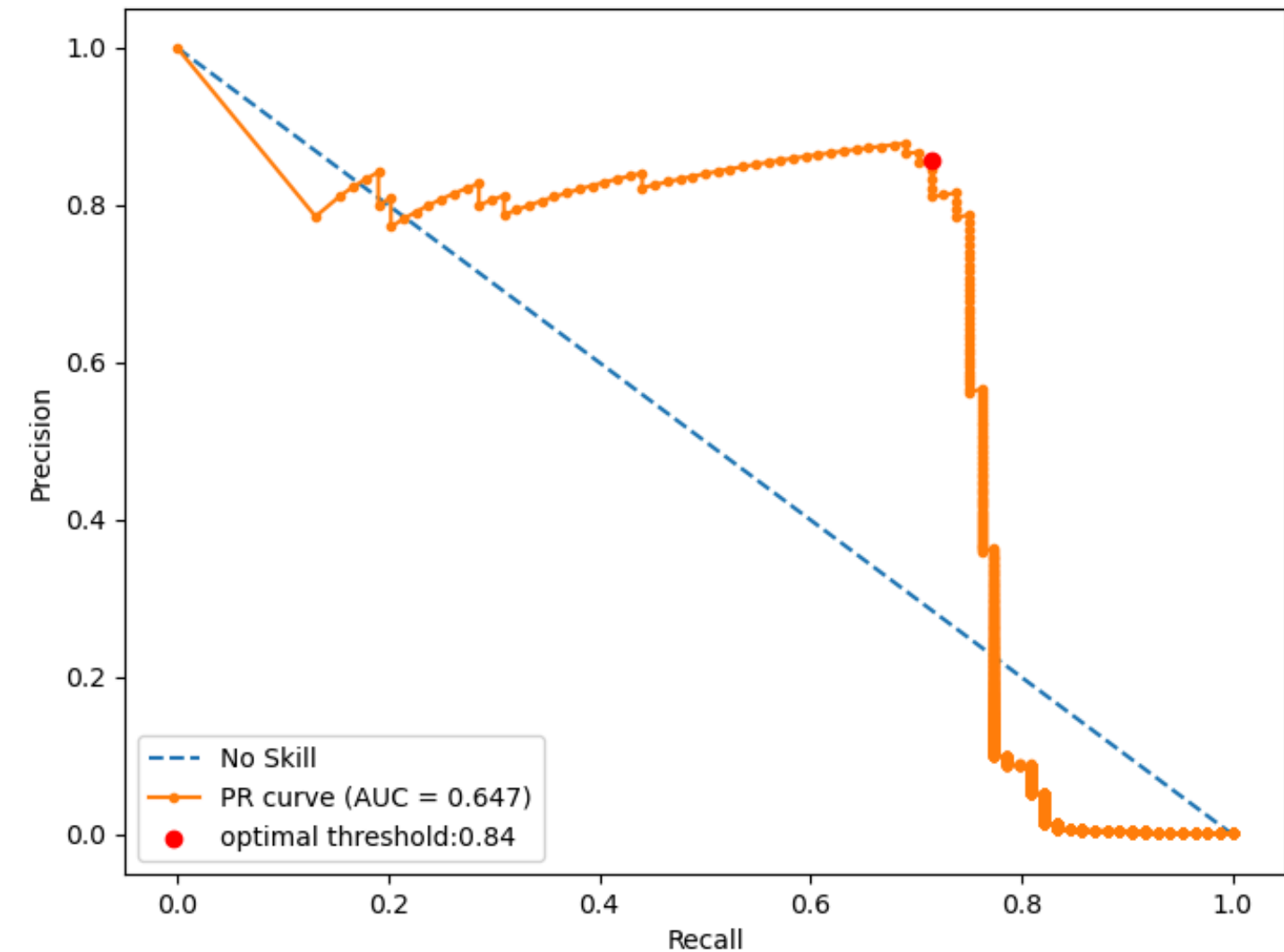
- **AUC=0.92**: cela indique que le modèle a une très bonne capacité de discrimination; 92% de chances de classer correctement une instance positive au-dessus d'une instance négative.
- **Threshold=0.02**: Un seuil aussi bas signifie que le modèle considère une probabilité très faible comme suffisante pour prédire la classe positive.



# Evaluation

## Precision-Recall Curve

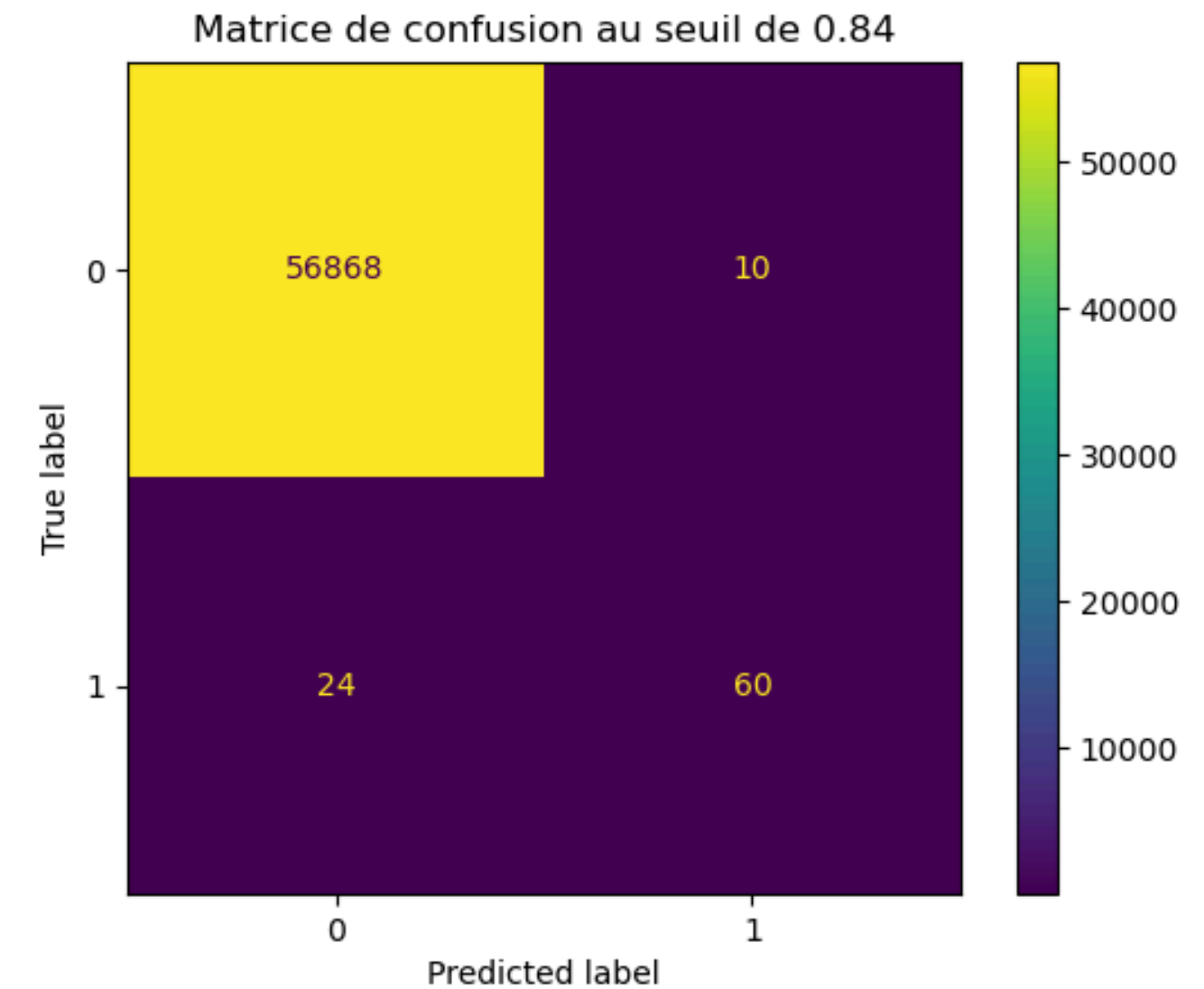
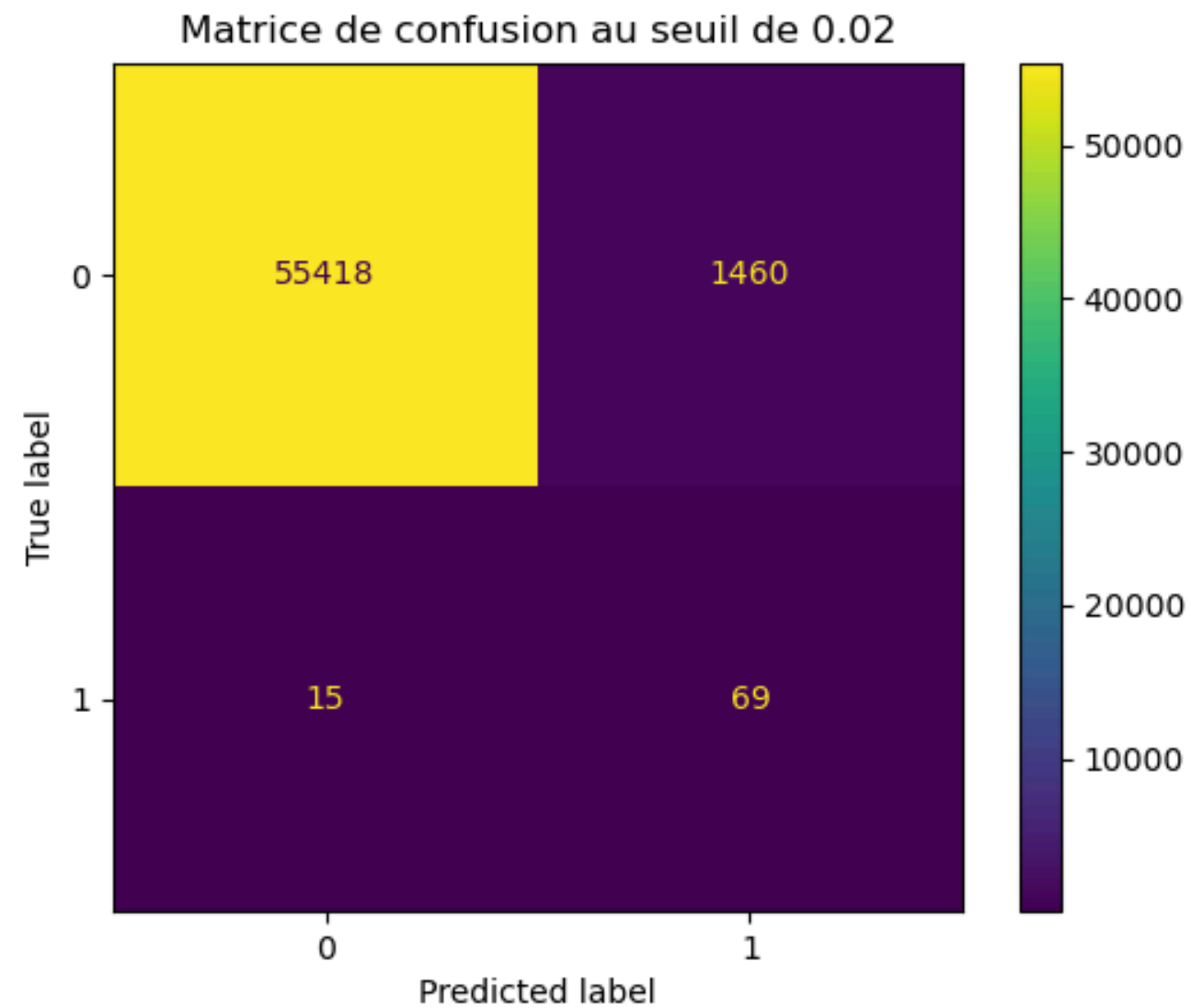
- **AUC=0.64** : cela indique que le modèle a une performance moyenne de détecter les fraudes.
- **Threshold=0.84** : Un seuil élevé signifie que le modèle considère une probabilité supérieure ou égale à 84% pour prédire la classe positive.





# Evaluation

Vérification des métriques avec la matrice de confusion aux seuils de 0.02 et 0.84





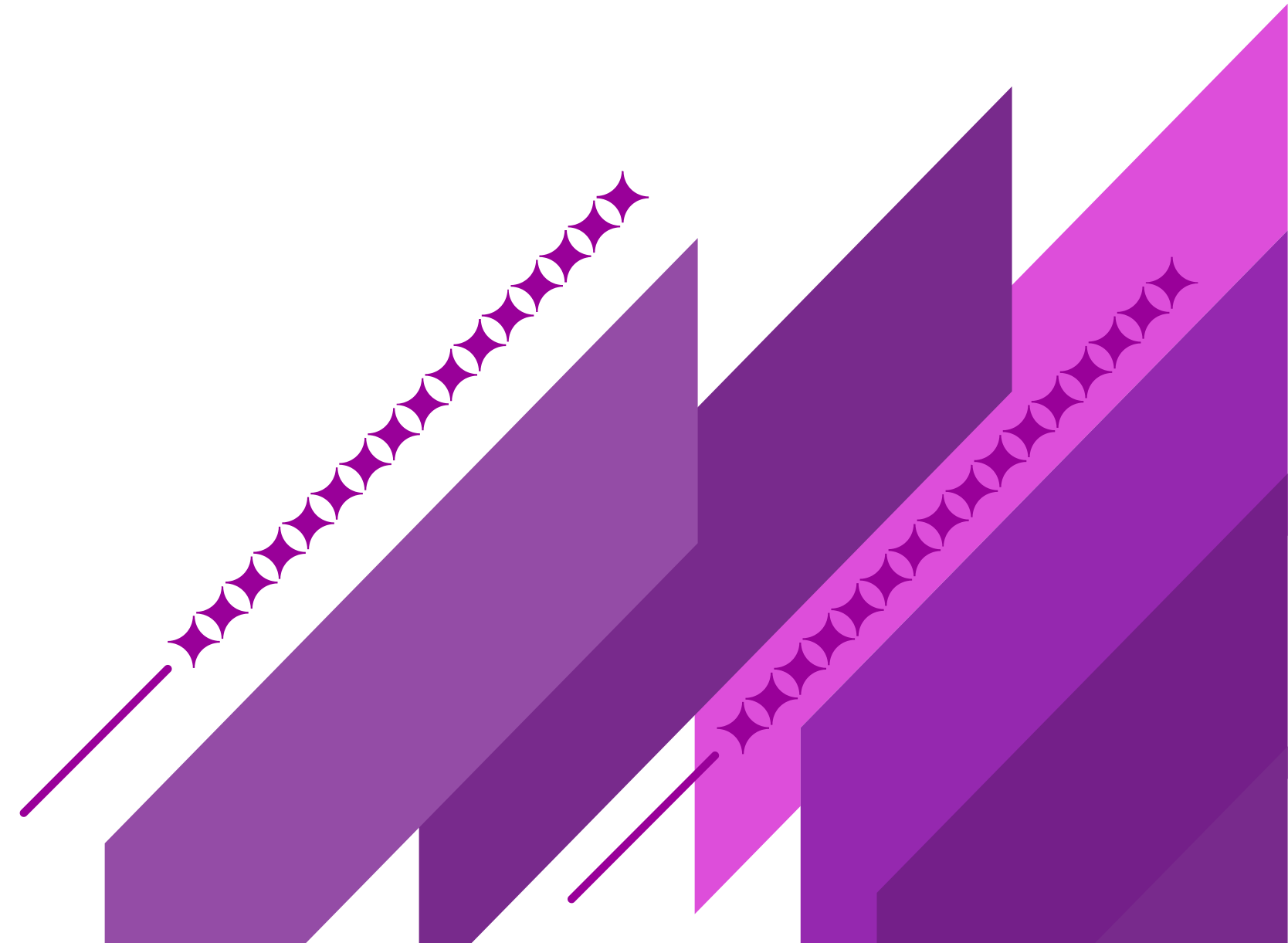


# Evaluation

Analyse des métriques en fonction des seuils

	Metrique	Threshold=0.02	Threshold=0.84
0	precision	0.0451	0.8551
1	recall	0.8214	0.7143
2	fpr	0.0269	0.0002
3	F1_score	0.0856	0.7792

Discussion



# Discussion

## En fonction des seuils

- **Threshold=0.02** : un seuil optimal bas signifie que notre modèle est configuré pour être très sensible à tout signal de fraude, même si ce signal est faible.
  1. Ce seuil bas favorise une baisse de la précision(0.0451) et une augmentation des transactions légitimes déclarées à tort comme frauduleuses(FPR=0.0269).
  2. Des coûts opérationnelles supplémentaires et nuisance de l'expérience utilisateur pour les clients légitimes.
- **Threshold=0.84** : un seuil optimal élevé signifie que notre modèle ne classera une transaction comme frauduleuse que s'il est très confiant.
  1. Ce seuil favorise une hausse de la précision(0.85) et une baisse de FPR(0.0002).
  2. Cependant, le nombre de fraudes non détectées(FN=24) a augmenté.
  3. Conséquences graves comme perte financière ou risque réputationnel important, dans le cas où une fraude passe inaperçue.

# Discussion

## En fonction de l'AUC

- **AUC=0.92** : signifie que le modèle a une bonne performance pour détecter les fraudes, alors 92% de chances de classer correctement une instance positive au-dessus d'une instance négative. Donc le modèle est bon pour discriminer entre transactions frauduleuses et légitimes.
- **AUC=0.647** : signifie que le modèle a une performance moyenne pour détecter les fraudes, donc une valeur de 0,647 suggère que le modèle a une capacité modérée à identifier correctement les fraudes tout en minimisant les fausses alertes.
  - Les caractéristiques (features) utilisées ne sont pas assez discriminantes.
  - Le modèle peut être amélioré (par exemple, par un meilleur équilibrage des classes ou des ajustements d'hyperparamètres).

Conclusion

Perspectives



# Conclusion

- Après avoir fait des analyses exploratoire et multivariée sur nos données, j'ai utilisé la régression logistique pour modéliser notre problème.
- L'évaluation de la performance du modèle avec la courbe ROC donne un seuil bas augmentant les transactions légitimes déclarées à tort comme frauduleuses et cela peut causer des coûts opérationnelles supplémentaires et nuire à l'expérience utilisateur pour les clients légitimes.
- Celle de la courbe de precision-recall donne un seuil élevé augmentant les cas de fraudes non détectés et cela peut causer des pertes financières ou risque réputationnel important.
- Notre modèle est peu performant et peut être amélioré.

# Perspectives

- Amélioration des données :
  - Enrichir nos données avec des caractéristiques (features) plus discriminantes.
  - Utiliser des techniques de génération de données synthétiques pour équilibrer les classes (par exemple, SMOTE).
- Optimisation du modèle :
  - Envisager des algorithmes tels que les forêts aléatoires, arbres de décision, les modèles d'ensemble comme XGBoost, les machines à vecteurs de support, ou les réseaux de neurones.
  - Ajuster les hyperparamètres pour maximiser l'AUC-PR.
- Stratégie de seuil :
  - Tester différents seuils autour de 0.84 pour trouver un meilleur compromis entre précision et rappel.

Thank  
you

