SISSOKO Moussa

Rapport d'analyse

Défaut de paiements des clients taiwanais

2024

Contexte de l'étude

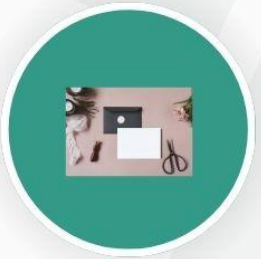
Dans cette section, nous définissons les objectifs de notre analyse des données financières des clients taiwanais sur la période de 2009 à 2016.

Identifier les variables significatives

Dans cette première partie, nous chercherons à déterminer quelles variables sont les plus influentes sur le risque de défaut de paiement des clients. Nous utiliserons des techniques d'analyse statistique pour identifier les variables qui présentent une corrélation significative avec le défaut de paiement.

Évaluer l'impact des caractéristiques démographiques

La deuxième partie de notre analyse vise à comprendre comment des facteurs démographiques tels que l'âge, le sexe, le niveau d'éducation et l'état matrimonial influent sur le risque de défaut de paiement. Nous examinerons comment ces caractéristiques sont associées au défaut de paiement et évaluerons leur importance relative.



Créer un modèle prédictif

Dans cette troisième partie, notre objectif est de développer un modèle prédictif capable d'estimer la probabilité de défaut de paiement des clients. Nous utiliserons les variables les plus significatives identifiées lors des étapes précédentes pour construire un modèle robuste qui puisse prédire avec précision le risque de défaut de paiement.

1. Introduction

Le présent rapport vise à analyser les facteurs associés au défaut de paiement des clients taiwanais dans le secteur financier. L'objectif est de comprendre les caractéristiques des clients qui présentent un risque accru de défaut de paiement afin d'aider les institutions financières à prendre des décisions plus éclairées en matière de prêt . Avec le secteur financier taiwanais confronté à des défis croissants en matière de prêts et de recouvrement, il est impératif de comprendre les caractéristiques des clients présentant un risque accru de défaut de paiement .

Dans ce contexte , il serait bien d'élaborer une équation de modèle pour évaluer le défaut de paiement , prenant en compte plusieurs variables explicatives telles que le sexe, le niveau d'éducation , le statut matrimonial, l'état de paiement pour plusieurs mois, la limite de crédit, l'âge, les montants des factures et des paiements .

Grace à cette analyse , j'espère offrir aux institutions financières le soutils nécessaires pour identifier et évaluer les clients à risque de défaut de paiement, réduisant ainsi les pertes et favorisant une prise de décision plus responsable et durable dans le secteur financier taiwanais .

L'équation du modèle r est la suivante :

Model defaultpayment(Y) = Sex Education Marriage PAY_0 PAY_2 PAY_3 PAY_4 PAY_5 PAY_6 Limit_Bal Age BILL_AMT1 BILL_AMT2 BILL_AMT3 BILL_AMT4 BILL_AMT5 BILL_AMT6 PAY_AMT1 PAY_AMT2 PAY_AMT3 PAY_AMT4 PAY_AMT5 PAY_AMT6 .

2. Methodologies

- Collecte , Nettoyage et Normalisation des Données :

  - Collecte de données financières des clients taiwanais sur la période de 2009 à 2016 .
  - Nettoyage des données pour traiter les valeurs manquantes et les erreurs .

3. Analyses statistiques

- Sélection des variables explicatives pertinentes

  - Identification des variables potentiellement influentes sur le défaut de paiement .

Tableau 1 : Statistiques descriptives

La procédure MEANS

Variable	Libellé	N	Moyenne	Ec-type	Minimum	Quartile inférieur	Médiane	Quartile supérieur	Maximum
AGE	AGE	30000	35.4855000	9.2179041	21.0000000	28.0000000	34.0000000	41.0000000	79.0000000
BILL_AMT1	BILL_AMT1	30000	51223.33	73635.86	-165580.00	3558.50	22381.50	67092.00	964511.00
BILL_AMT2	BILL_AMT2	30000	49179.08	71173.77	-69777.00	2984.50	21200.00	64008.50	983931.00
BILL_AMT3	BILL_AMT3	30000	47013.15	69349.39	-157264.00	2665.50	20088.50	60165.50	1664089.00
BILL_AMT4	BILL_AMT4	30000	43262.95	64332.86	-170000.00	2326.50	19052.00	54509.00	891586.00
BILL_AMT5	BILL_AMT5	30000	40311.40	60797.16	-81334.00	1763.00	18104.50	50196.00	927171.00
BILL_AMT6	BILL_AMT6	30000	38871.76	59554.11	-339603.00	1256.00	17071.00	49200.50	961664.00
LIMIT_BAL	LIMIT_BAL	30000	167484.32	129747.66	10000.00	50000.00	140000.00	240000.00	1000000.00

Source : Defaut de paiement des clients Taiwanais

Figure 1 : Répartition selon défaut de paiement

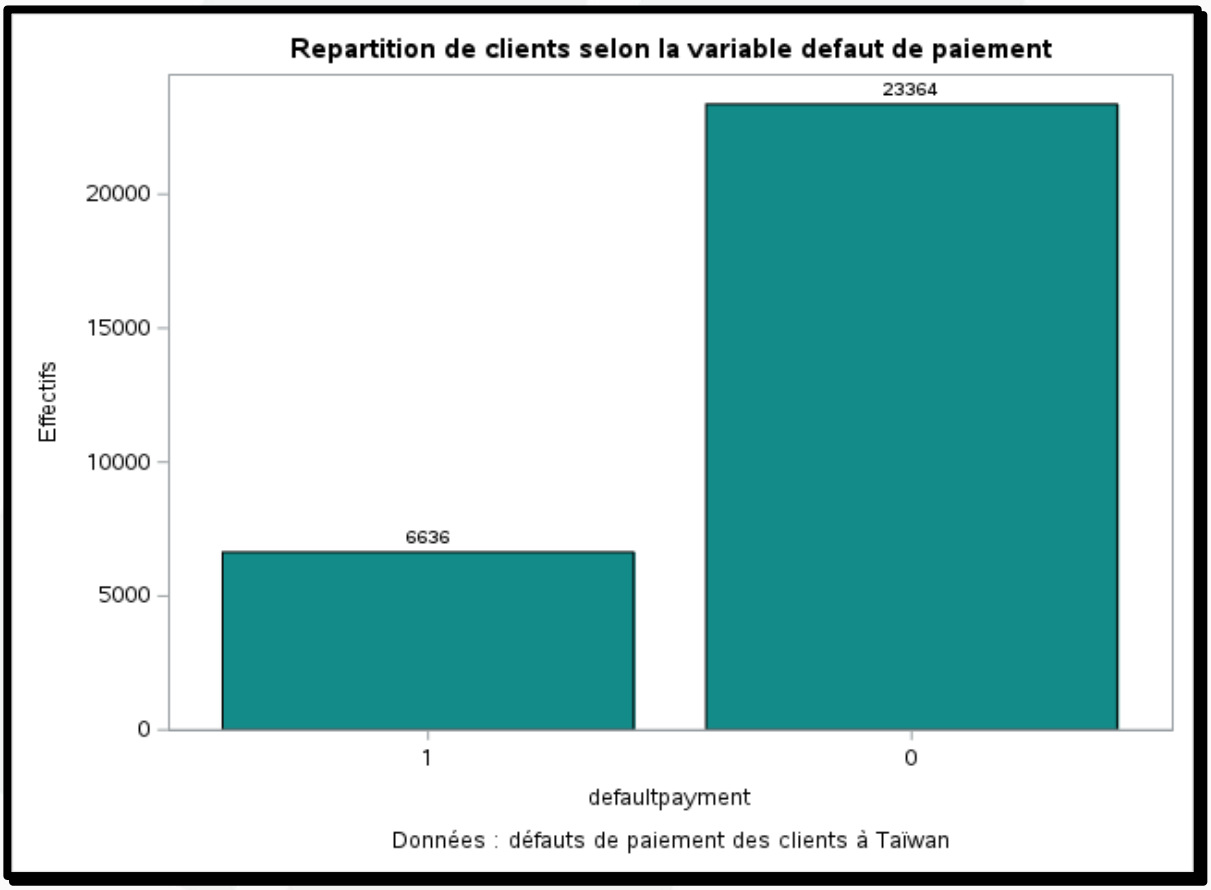


Figure 2 : Répartition selon la situation matrimonial

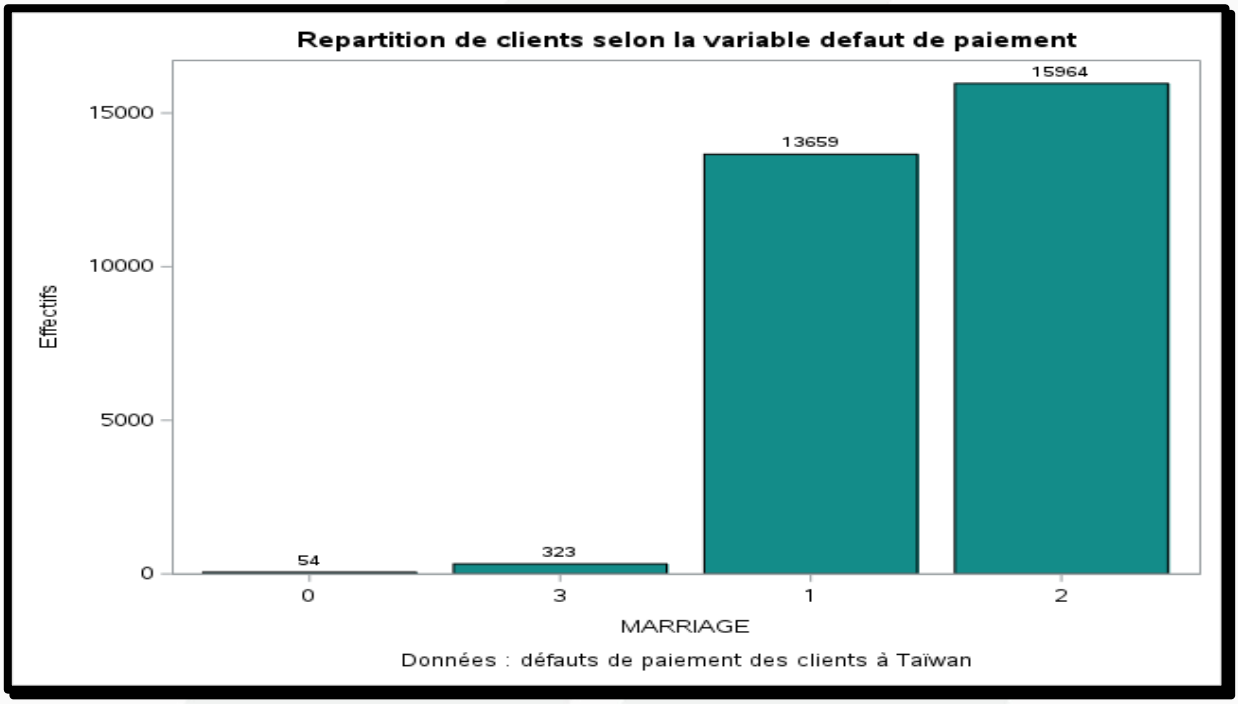
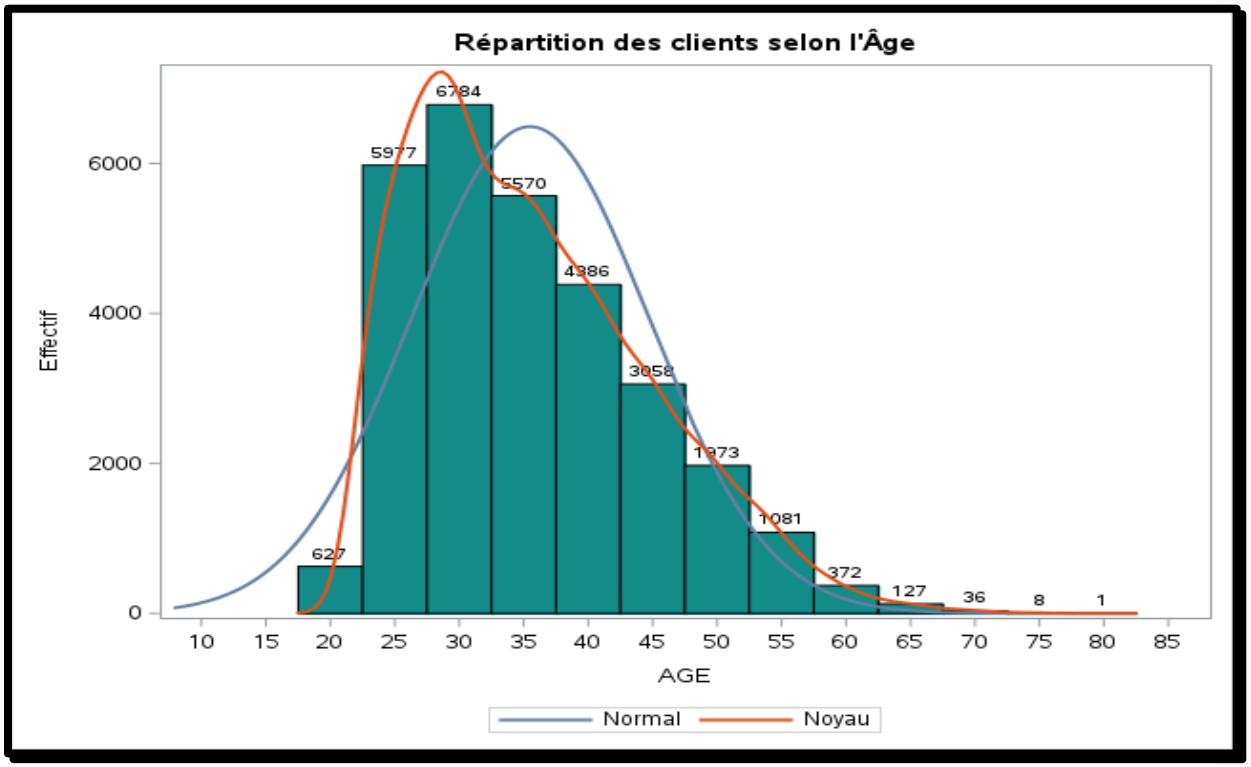


Figure 3 : Répartition selon l'Age



Dans le graphique ci-dessus on peut observer que la majorité des clients ont environ 30 ans ce qui nous permet de comprendre la composition démographique de la clientèle .

-En ce qui concerne l'Utilisation d'analyses statistiques pour identifier les facteurs significatifs :

- Analyse statistique approfondie pour évaluer l'impact des variables sur le risque de défaut .
- En examinant le tableau des coefficients de Pearson fourni, voici quelques observations importantes :
  - On constate que les variables telles que : PAY_0 jusqu'à PAY_6 sont corrélées de manière positive au défaut de paiement .
  - D'autres variables explicatives ne présentent pas de corrélation significative avec le défaut de paiement .
  - les variables telles que le montant de crédit accordé (Limit-Bal) , PAY_0 jusqu'à PAY_6 , le sexe , l'état matrimonial, et les montants payés (PAY_AMT1 jusqu'à PAY_AMT6) sont significatives pour expliquer le défaut de paiement des clients .

4. Les tests statistiques

```
%macro table_freq_chi2(var1, var2);
   proc freq data=work.taiwan;
      tables &var1 * &var2 / chisq;
   run;
%mend;

%table_freq_chi2(Sex, Defaultpayment);
%table_freq_chi2(Age, Defaultpayment);
%table_freq_chi2(education, Defaultpayment);
%table_freq_chi2(Marriage, Defaultpayment);
```

En utilisant le test de khi-deux, nous pouvons constate que les variables comme Age, Sexe , éducation et mariage sont significatives au seuil de 5% . Ce qui signifie qu'il existe une association significative entre ces variables et le défaut de paiement des clients taiwanais .

```
%macro shapirowilks(var_a_tester);
proc univariate data =work.taiwan NORMAL;
var &Var_a_tester;
run;
%mend ;

%shapirowilks(Age);
```

```
%shapirowilks(limit_bal);
```

Les variables comme l'Age et le montant de crédit accordé ne suivent pas une distribution normale .

```
%macro Student_test(var_interet, var_explicative);
proc ttest data = work.taiwan;
var &var_explicative;
run;
%mend;

%student_test(Defaultpayment, Age);
```

Il existe une association significative entre la variable d'ineteret (Défaut de paiement) et la variable explicative (Age) .Alors on peut donc dire que l'age à un impact significatif sir leur propension à etre en défaut de paiement .

5. Modèle de machine learning

Pour l'analyse prédictive de defaut de paiement des clients , j'ai separé le modèle par Sélection aléatoire des données .

Tableau 1 : Echantillonnage

Méthode de sélection	Echantillonnage aléatoire simple

Table d'entrée	TAIWAN
Valeur initiale du nombre aléatoire	2
Taux d'échantillonnage	0.8
Taille d'échantillon	24000
Probabilité de sélection	0.8
Poids d'échantillonnage	0
Table de sortie	TAIWAN2

Données : défauts de paiement des clients à Taïwan

J'ai d'un côté crée une base d'apprentissage et d'un autre une base test .

- ## Régression logistique

Tableau 2: Effets supprimés

Récapitulatif sur l'élimination descendante						
Etape	Effet supprimé	DDL	Nombre dans	Khi-2 de Wald	Pr > khi-2	Libellé de la variable
1	BILL_AMT4	1	22	0.0386	0.8442	BILL_AMT4
2	PAY_AMT4	1	21	0.0991	0.7529	PAY_AMT4
3	BILL_AMT6	1	20	0.3620	0.5474	BILL_AMT6
4	PAY_AMT3	1	19	0.4638	0.4958	PAY_AMT3
5	BILL_AMT1	1	18	0.9001	0.3428	BILL_AMT1
6	BILL_AMT2	1	17	1.9533	0.1622	BILL_AMT2
7	BILL_AMT5	1	16	1.9631	0.1612	BILL_AMT5
8	PAY_AMT6	1	15	3.3187	0.0685	PAY_AMT6
9	PAY_2	10	14	17.7184	0.0599	PAY_2
10	AGE	1	13	3.7912	0.0515	AGE

Toutes les variables mentionnées ont été exclues du modèle, car elles n'ont pas démontré de signiifcativité statistique, ayant des probabilités critiques supérieures au seuil de 5%.

Cette sélection a été réalisée en utilisant la méthode de pas à pas descendante (backward selection) qui supprime en premier tous les variables ayant des coefficients les moins significatifs . Ce processus à été répété jusqu'à ce que toutes les variables restantes aient des

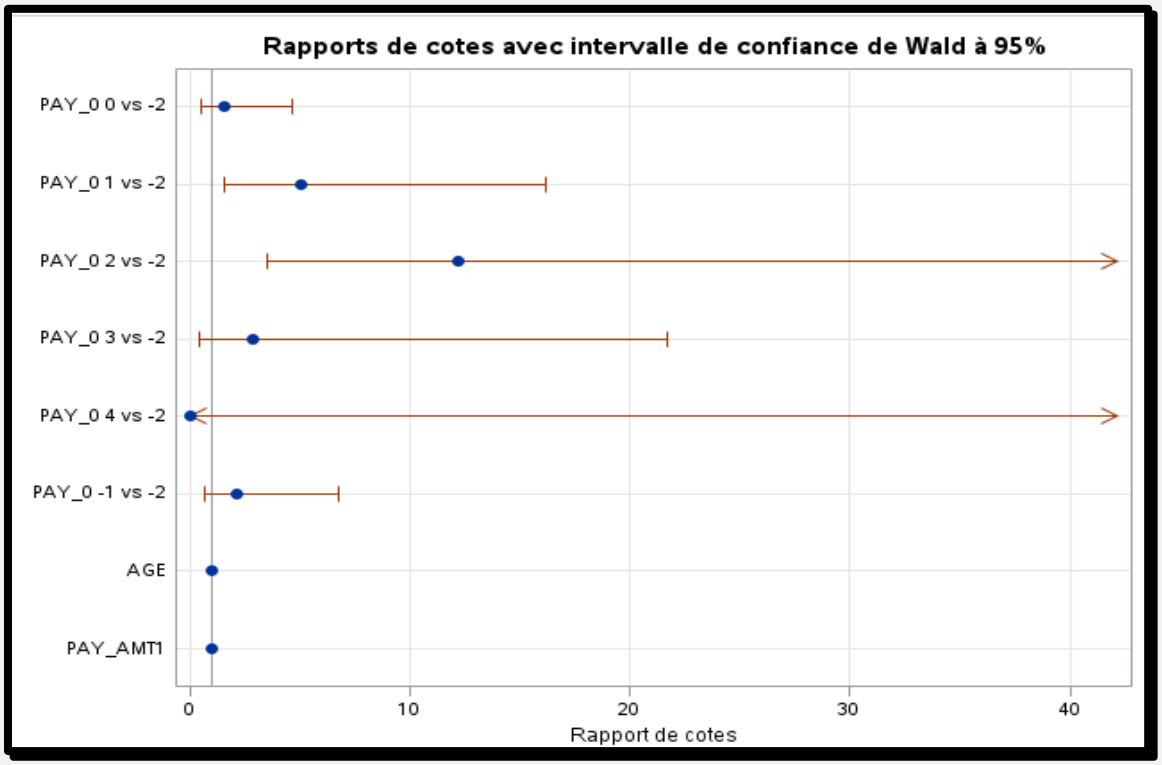
coefficients significatifs, assurant ainsi la robustesse et la pertinence du modèle .

Tableau 3 : Variables significatives

Analyse des effets Type 3			
Effet	DDL	Khi-2 de Wald	Pr > khi-2
SEX	1	23.9998	<.0001
EDUCATION	6	27.8385	0.0001
MARRIAGE	3	45.6645	<.0001
PAY_0	10	1556.1728	<.0001
PAY_3	10	37.3968	<.0001
PAY_4	8	24.6515	0.0018
PAY_5	8	23.2169	0.0031
PAY_6	8	50.9264	<.0001
LIMIT_BAL	1	107.0583	<.0001
BILL_AMT3	1	42.5790	<.0001
PAY_AMT1	1	19.4201	<.0001
PAY_AMT2	1	20.7592	<.0001
PAY_AMT5	1	6.3180	0.0120

Les variables ci-dessus sont tous significatives au seuil de 5% alors ont un impact significatif sur le défaut de paiement des clients .

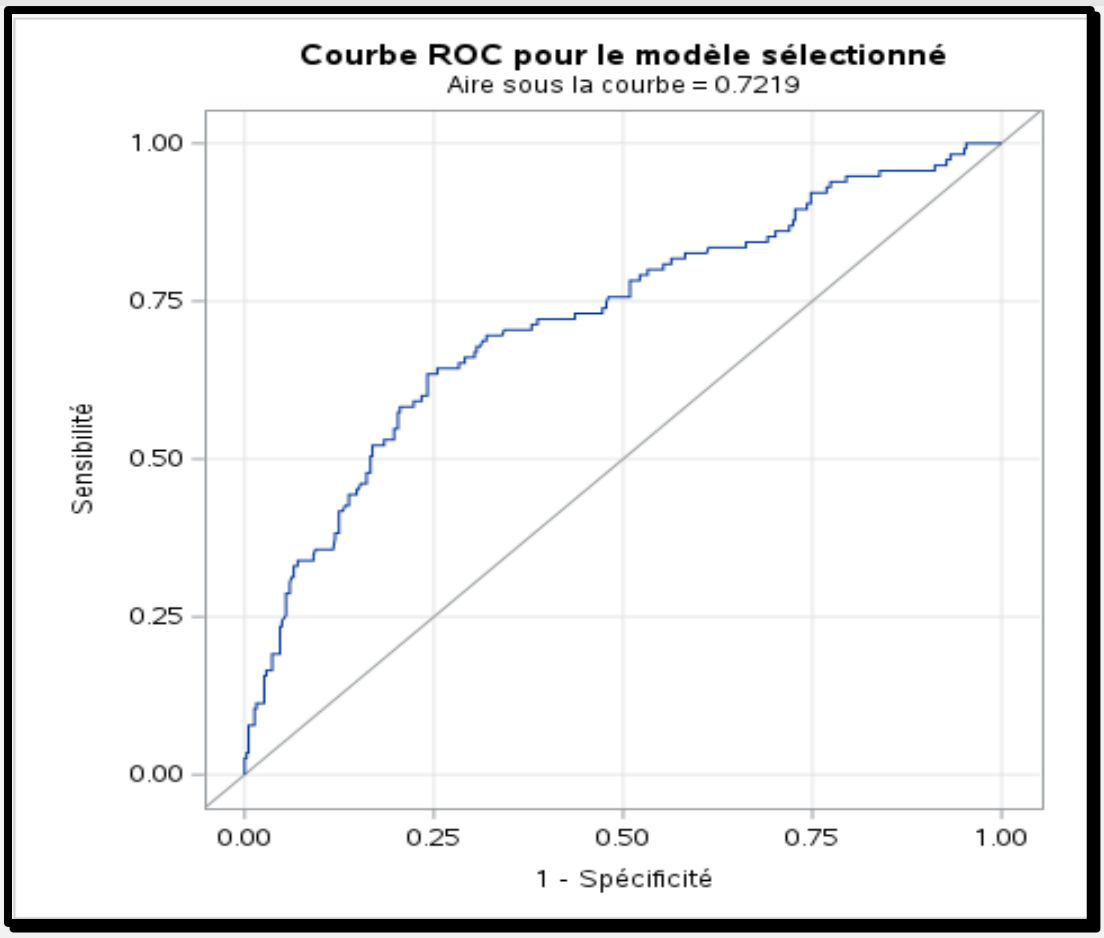
Graphique 1 : Rapport de cotes



Pour les clients qui ont un retard de paiement d'un mois (PAY_0=1 versus ceux qui ont un retard de paiement de deux mois (PAY_0=-2) , le rapport de cotes est estimer à 5,024 , ce qui signifie que les clients ayant un retard de paiement de deux mois ont environ 5,024 fois plus chances d'etre en défaut de paiement par rapport ceux ayant un retard d'un mois .

Pour l'Age , les odds-ratios est égale à 1,025 ce qui signifie que chaque année supplémentaire augmente les chances d'etre en défaut de paiement de 2,5 % .

Figure 2 : La courbe ROC



Concernant la qualité de l'ajustement, la statistique de Wald confirme la signification commune des variables incluses dans le modèle, en réjetant l'hypothèse nulle selon laquelle les coefficients sont égaux à zéro .

Comme on peut le voir dans la figure 2, la courbe ROC (Receiver Operating Characetristic) du modèle n' approche pas très bien le coin supérieur gauche du graphique, ce qui peut dire que le modèle n'est très performant mais nous avons quand même un pourcentage de 72,19% ce qui est acceptable pour un modèle de machine learning .

Tableau 4 : Tests de Hosmer et Lemeshow

Test d'adéquation de Hosmer et de Lemeshow		
khi-2	DDL	Pr > khi-2
11.1134	8	0.1954

On peut voir que ci-dessus que la probabilité critique est supérieur au seuil XIIde 5 % alors de ce fait on accepte l'hypothèse nulle (H0) selon laquelle le modèle à un bon ajustement . En d'autres termes le modèle s'ajuste raisonnablement bien aux données, conformément à l'hypothèse nulle .

Tableau 5 : Matrice de confusion

	Prediction	
	Non	Oui
	N	N
defaultpayment		
0	71	6
1	12	6

Afin de savoir si le modèle predit correctement , on va calculer certains indicateurs :

- **Indicateur de sensibilité** : 0,33 , le modèle explique alors le défaut de paiement des clients à hauteur de 33% . Cet indicateur mésure la capacité du modèle à identifier correctement les vrais positifs .

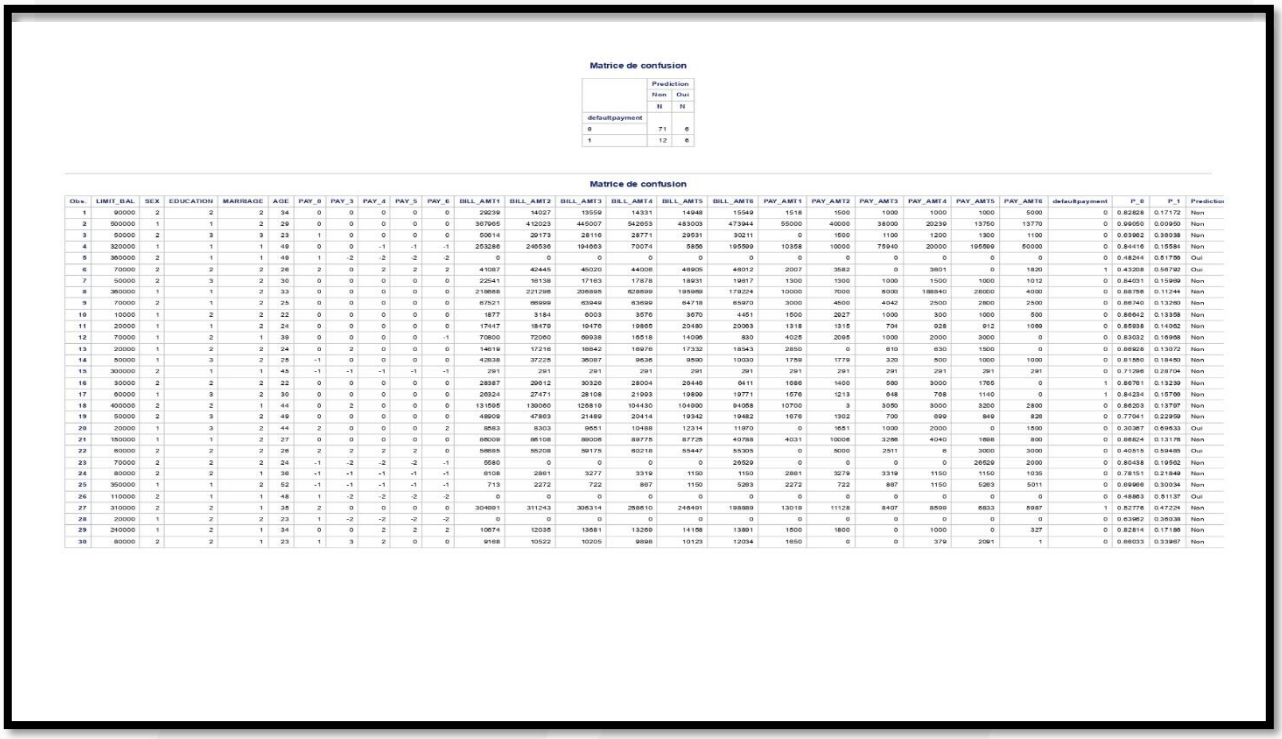
- **Indicateur de spécificité** : 0,92 , il existe 92% de chance que les clients soient pas en défaut de paiement .Celui-ci mesure la capacité du modèle à identifier correctement les vrais négatifs c'est-à-dire le non évènement .

- **Indicateur de précision** : 0,5 , le modèle explique correctement à hauteur de 50% .

- **Faux positifs** : 1-spécifité : 0,08 , cet indicateur correspond aux cas ou le modèle prédit à tort un défaut de paiement des clients .

- **Taux d'erreur** : 0,19 , ce taux est la proportion d'erreurs de prédiction globales du modèle .

Tableau 6 : Modèle prédictif



Matrice de confusion

	Prédiction Non	Prédiction Oui
defaultpayment	N	N
0	71	6
1	12	6

Matrice de confusion

Obs.	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	defaultpayment	P_0	P_1	Prediction
1	90000	2	2	2	34	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0	0.82628	0.17172	Non
2	500000	1	1	2	29	0	0	0	0	0	367905	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770	0	0.99050	0.00950	Non
3	50000	2	3	3	23	1	0	0	0	0	50614	29173	28116	28771	29531	30211	0	1500	1100	1200	1300	1100	0	0.63962	0.36038	Non
4	320000	1	1	1	49	0	0	-1	-1	-1	253286	246536	194663	70074	5856	195599	10358	10000	75940	20000	195599	50000	0	0.84416	0.15584	Non
5	360000	2	1	1	49	1	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0.48244	0.51756	Oui
6	70000	2	2	2	26	2	0	2	2	2	41087	42445	45020	44006	46905	48012	2007	3582	0	3601	0	1820	1	0.43268	0.56732	Oui
7	50000	2	3	2	30	0	0	0	0	0	22541	19138	17163	17878	18931	19817	1300	1300	1000	1500	1000	1012	0	0.84031	0.15969	Non
8	360000	1	1	2	33	0	0	0	0	0	219868	221298	206895	628899	195989	179224	10000	7000	8000	188840	28000	4000	0	0.88756	0.11244	Non
9	70000	2	1	2	25	0	0	0	0	0	67521	68999	63949	63699	64718	65970	3000	4500	4042	2500	2800	2500	0	0.86740	0.13260	Non
10	10000	1	2	2	22	0	0	0	0	0	1877	3184	6003	3576	3670	4451	1500	2027	1000	300	1000	500	0	0.86642	0.13358	Non
11	20000	1	1	2	24	0	0	0	0	0	17447	18479	19476	19855	20480	20963	1318	1315	704	928	912	1089	0	0.85898	0.14102	Non
12	75000	1	2	1	39	0	0	0	0	-1	70800	72060	59938	16518	14096	830	4025	2065	1000	2000	3000	0	0	0.83032	0.16968	Non
13	20000	1	2	2	24	0	2	0	0	0	14619	17218	18842	18976	17332	18543	2850	0	610	630	1500	0	0	0.86928	0.13072	Non
14	50000	1	3	2	28	-1	0	0	0	0	42838	37228	36087	9536	9590	10030	1759	1779	320	500	1000	1000	0	0.81850	0.18150	Non
15	300000	2	1	1	45	-1	-1	-1	-1	-1	291	291	291	291	291	291	291	291	291	291	291	291	0	0.71296	0.28704	Non
16	30000	2	2	2	22	0	0	0	0	0	28387	29612	30326	28004	29446	9411	1686	1400	590	3000	1765	0	1	0.86761	0.13239	Non
17	60000	1	3	2	30	0	0	0	0	0	26824	27471	28108	21993	19890	19771	1576	1213	648	768	1140	0	1	0.84234	0.15766	Non
18	400000	2	2	1	44	0	2	0	0	0	131595	130060	126819	104430	104090	94058	10700	3	3050	3000	3200	2800	0	0.86203	0.13797	Non
19	50000	2	3	2	49	0	0	0	0	0	48909	47803	21489	20414	19342	19482	1676	1302	700	699	849	826	0	0.77041	0.22959	Non
20	20000	1	3	2	44	2	0	0	0	2	8583	8303	9651	10488	12314	11970	0	1651	1000	2000	0	1500	0	0.30367	0.69633	Oui
21	150000	1	1	2	27	0	0	0	0	0	88009	88108	88006	89775	87725	40788	4031	10006	3286	4040	1888	800	0	0.86624	0.13376	Non
22	60000	2	2	2	26	2	2	2	2	0	58665	59208	59175	60218	59447	55305	0	5000	2811	6	3000	3000	0	0.40515	0.59485	Oui
23	70000	2	2	2	24	-1	-2	-2	-2	-1	5590	0	0	0	0	29529	0	0	0	0	26529	2000	0	0.80438	0.19562	Non
24	80000	2	2	1	36	-1	-1	-1	-1	-1	6108	2861	3277	3319	1150	1150	2861	3279	3319	1150	1150	1035	0	0.78151	0.21849	Non
25	350000	1	1	2	52	-1	-1	-1	-1	-1	713	2272	722	887	1150	5293	2272	722	887	1150	5293	5011	0	0.69966	0.30034	Non
26	110000	2	1	1	48	1	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0.48863	0.51137	Oui
27	310000	2	2	1	35	2	0	0	0	0	304991	311243	306314	258510	246491	199889	13018	11128	8407	8599	6833	5987	1	0.52776	0.47224	Non
28	20000	1	2	2	23	1	-2	-2	-2	-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0.63962	0.36038	Non
29	240000	1	2	1	34	0	0	2	2	2	10674	12035	13581	13260	14158	13891	1500	1600	0	1000	0	327	0	0.82814	0.17186	Non
30	80000	2	2	1	23	1	3	2	0	0	9168	10522	10205	9898	10123	12034	1650	0	0	379	2091	1	0	0.86033	0.13967	Non

En raison de contraintes d'affichage, seules les 10 premières observations ont été présentées ici, bien que le jeu de données contienne plus de 3000 observations. Malgré quelques prédictions incorrectes sur certaines observations, il est clair que le modèle prédictif est raisonnablement performant dans l'ensemble **(Pour voir le reste du modèle prédictif cliquer sur le lien en annexe**) .

6. Conclusions

En conclusion, notre analyse des données financières des clients taiwanais a révélé plusieurs point importants :

1. Certaines variables comme les retards de paiement, le montant de crédit accordé, le sexe, l'état matrimonial et les montants payés sont significativement liées au défaut de paiement.
2. Des tests ont montré une association significative entre des variables telles que l'age, le sexe, l'éducation et le mariage, et le défaut de paiement .
3. Notre modèle prédictif, basé sur une selection rigoureuse des variables, permet une prédiction précise du défaut de paiement .
4. L'interpretation des coefficients révèle l'impact significatif de certaines variables, comme le retard de paiement, sur le défaut de paiement .

5. Bien que la courbe ROC du modèle n'atteigne pas l'idéal, un taux de 72,19% reste acceptable pour un modèle de machine learning .

En somme, notre analyse fournit des informations essentielles pour la gestion des risques et les décisions financières .

Annexe

	Description de la Codification
Sex	1 = Masculin, 2 = Féminin
Education	1 = Graduate School, 2 = Université, 3 = High School, 4 = Autres
Mariage	1 = Marié, 2 = Célibataire, 3 = Autres
PAY_0 à PAY_6	État de paiement: -1 = Paiement complet, 1 = Retard de paiement de 1 mois, ..., 9 = Retard de paiement de 9 mois ou plus
Limit_Bal	Montant du crédit attribué
Age	Âge du client
BILL_AMT1 à BILL_AMT6	Montant de la facture pour les 6 derniers mois
PAY_AMT1 à PAY_AMT6	Montant du paiement pour les 6 derniers mois

Bibliographies :

Daniel Christophe , Cours d'économétrie 1 : Méthode d'évaluation .Université d'Angers

Navarro-Galera, Andrés, et al. « Quels sont les facteurs susceptibles d'accroître le risque de défaut de paiement des gouvernements locaux ? », *Revue Internationale des Sciences Administratives*, vol. 83, no. 2, 2017, pp. 403-426.

Lien vers les résultats du projet sous SAS

https://d.docs.live.net/3114a75d858e0a33/Bureau/Dossiers%20Perso/Projets/SAS/Résultats%20projet%20Taiwan.html