Prédire les faillites

Détecter les entreprises en difficultés

MeJo & Co

Plan de la présentation

01.

Introduction

Contexte du projet, équipe et objectif

02.

Modèle retenu

Aperçu rapide des résultats obtenus par notre meilleur modèle

03.

Preprocessing

Etapes suivies pour obtenir le dataframe utilisé pour nos différentes itérations

04.

Itérations

Utilisation de modèles différents pour améliorer les résultats des précédentes itérations

05.

Suite du projet

Pistes à suivre, axes d'améliorations

I / Introduction

Intro

Contexte

Choix de notre équipe

Sélection de notre société pour une mission de prédiction. Volonté d'assister la prise de décision pour l'investissement de notre client.

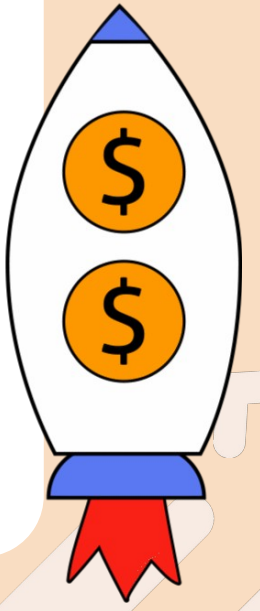
- Equipe de 2 personnes sur le projet (Mélody et Jonathan)
- Utilisation d'un jeu de données fourni par le client
- Indicateurs définis par les réglementations de la Bourse

Objectif

Détection des faillites

Trouver un modèle prédictif permettant de détecter les entreprises susceptibles de faire faillite. Le but étant d'investir sur les entreprises qui ne courent pas ce risque.

- Trouver un modèle de classification binaire.
- Utiliser un modèle ensembliste pour réaliser ce projet.
- Créer un protocole d'utilisation à destination du client.



II / Modèle retenu

Performances du modèle retenu

BaggingClassifier

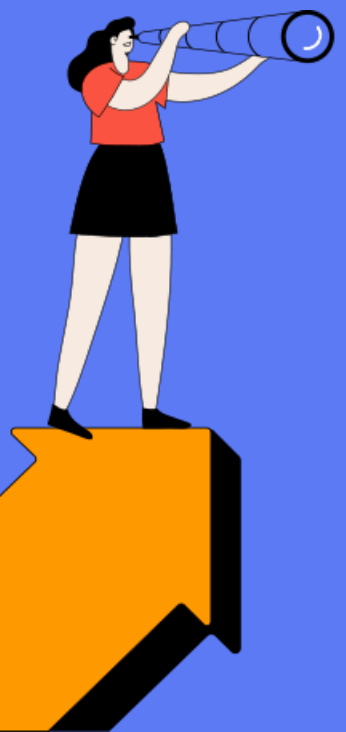
BaggingClassifier

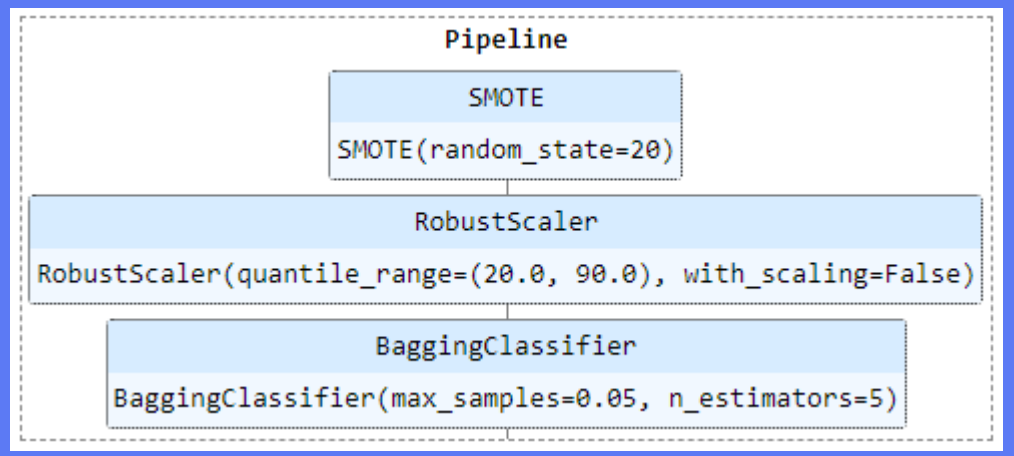
Utilisation de ce modèle fourni par la librairie Scikit Learn, avec optimisation de ces hyperpramètres.

68 – 73%

Score d'évaluation

Score obtenu avec ce modèle. Choix du recall pour évaluer nos prédictions, suite à la discussion avec notre client. Oscillation du résultat suivant les options de model tuning retenues





III / Preprocessing

Les étapes

Etape 1

Séparation des données en 2 dataframe (Colonnes avec valeur max de 1, puis colonnes avec valeur max supérieure à 1

Etape 2

Scaling des différentes colonnes (MinMax pour une partie, et Standard pour l'autre partie)

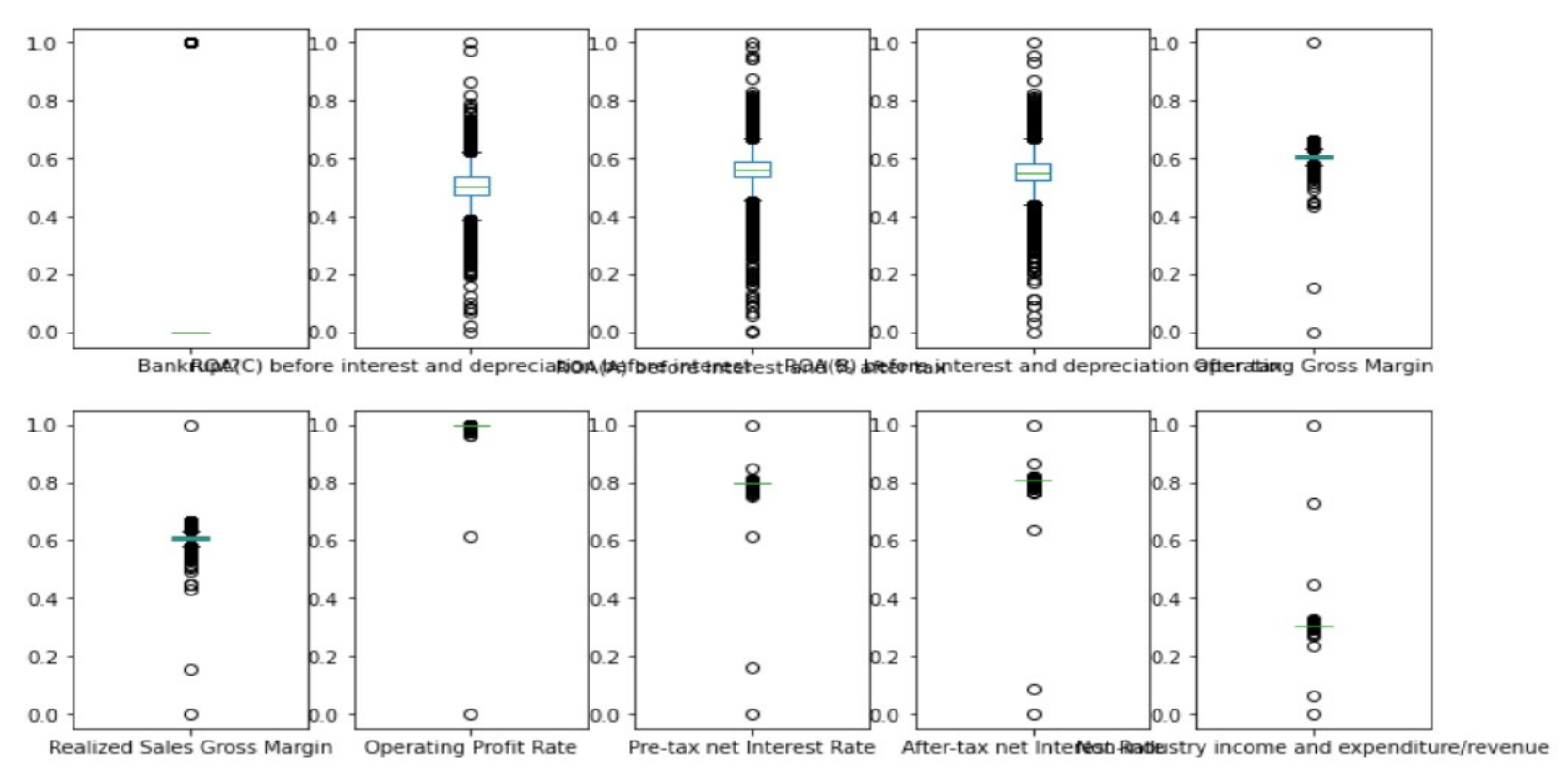
Etape 3

Réduction des dimensions de notre jeu de données et concaténation des deux différents dataframe

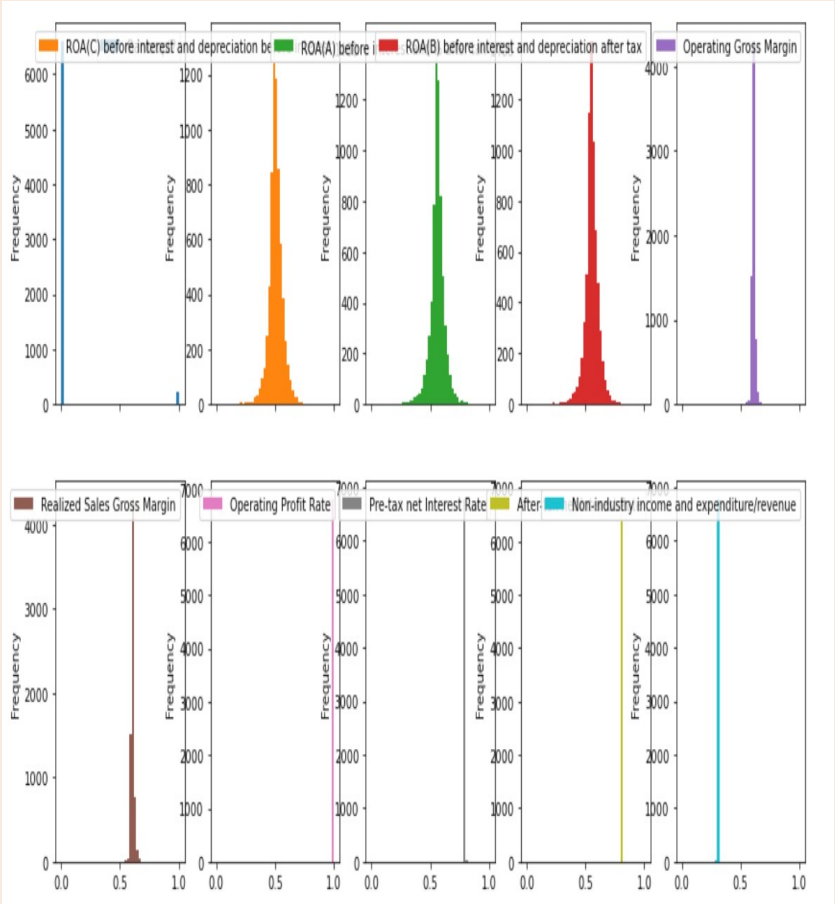
Vue d'ensemble

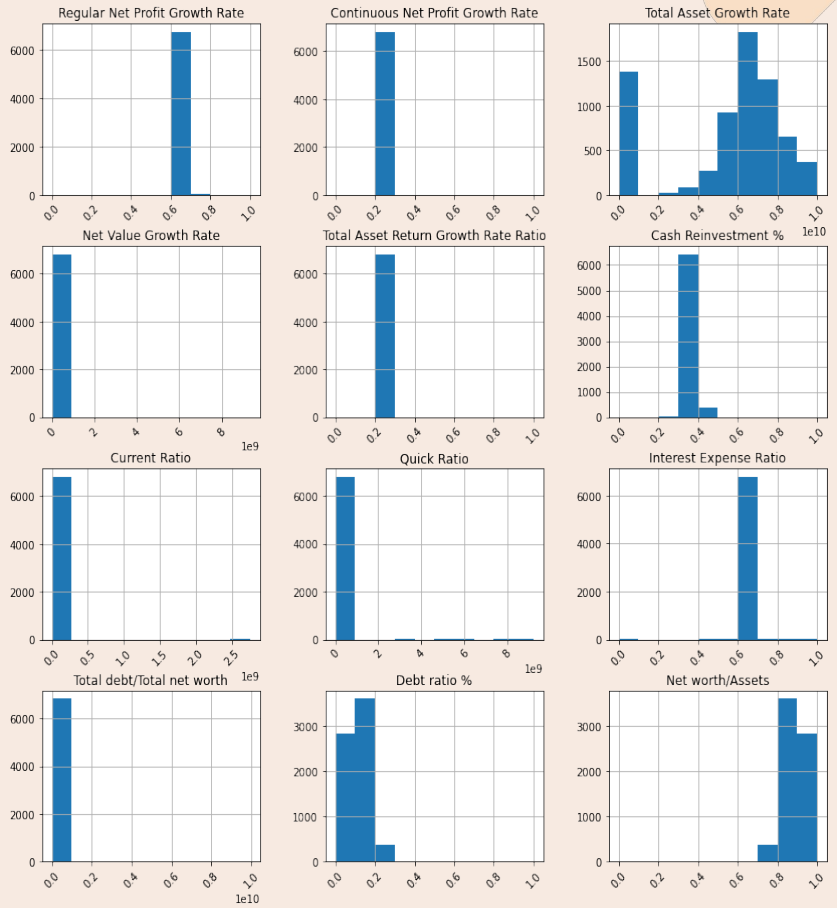
	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue	Continuous interest rate (after tax)	Operating Expense Rate	Research and development expense rate	Cash flow rate	Interest-bearing debt interest rate	Tax rate (A)	Net Value Per Share (B)	Net Value Per Share (A)	Net Value Per Share (C)
count	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6.819000e+03	6.819000e+03	6819.000000	6.819000e+03	6819.000000	6819.000000	6819.000000	6819.000000
mean	0.032263	0.505180	0.558625	0.553589	0.607948	0.607929	0.998755	0.797190	0.809084	0.303623	0.781381	1.995347e+09	1.950427e+09	0.467431	1.644801e+07	0.115001	0.190661	0.190633	0.190672
std	0.176710	0.060686	0.065620	0.061595	0.016934	0.016916	0.013010	0.012869	0.013601	0.011163	0.012679	3.237684e+09	2.598292e+09	0.017036	1.082750e+08	0.138667	0.033390	0.033474	0.033480
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000e+00	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.476527	0.535543	0.527277	0.600445	0.600434	0.998969	0.797386	0.809312	0.303466	0.781567	1.566874e-04	1.281880e-04	0.461558	2.030203e-04	0.000000	0.173613	0.173613	0.173676
50%	0.000000	0.502706	0.559802	0.552278	0.605997	0.605976	0.999022	0.797464	0.809375	0.303525	0.781635	2.777589e-04	5.090000e+08	0.465080	3.210321e-04	0.073489	0.184400	0.184400	0.184400
75%	0.000000	0.535563	0.589157	0.584105	0.613914	0.613842	0.999095	0.797579	0.809469	0.303585	0.781735	4.145000e+09	3.450000e+09	0.471004	5.325533e-04	0.205841	0.199570	0.199570	0.199612
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	9.990000e+09	9.980000e+09	1.000000	9.900000e+08	1.000000	1.000000	1.000000	1.000000

Outliers


1.0
0.8
0.6
0.4
0.2
0.0
Operating Gross Margin
Realized Sales Gross Margin
Operating Profit Rate
Pre-tax net Interest Rate

Distribution





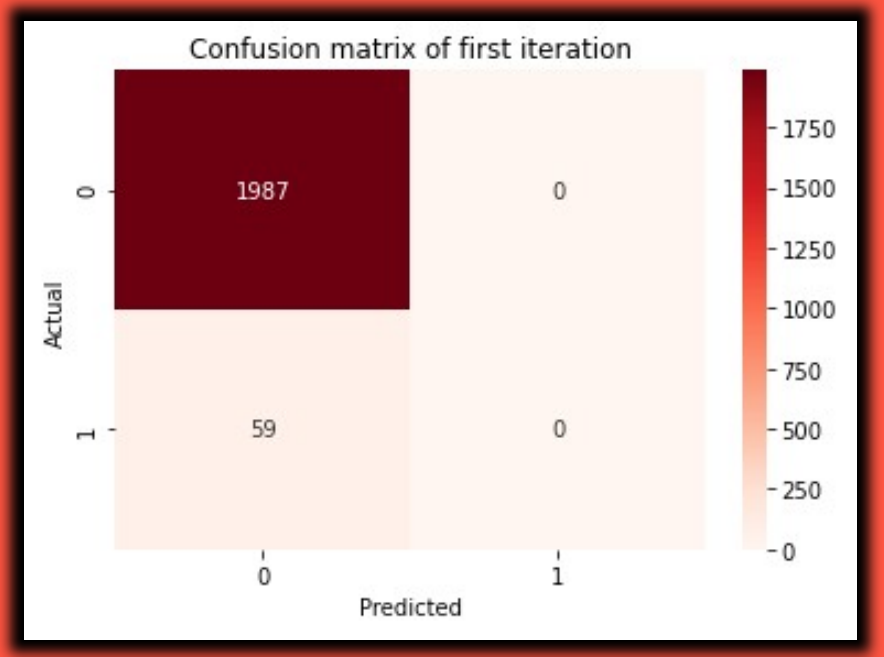
IV / Itérations

Première itération

01 Utilisation des données sans preprocessing

02 DummyRegressor() pour obenir notre baseline

03 Résultat médiocre pour détection des entreprises avec un risque de faillite



	precision	recall	f1-score	support
0	0.97	1.00	0.99	1987
1	0.00	0.00	0.00	59
accuracy			0.97	2046
macro avg	0.49	0.50	0.49	2046
weighted avg	0.94	0.97	0.96	2046

Recall du modèle:

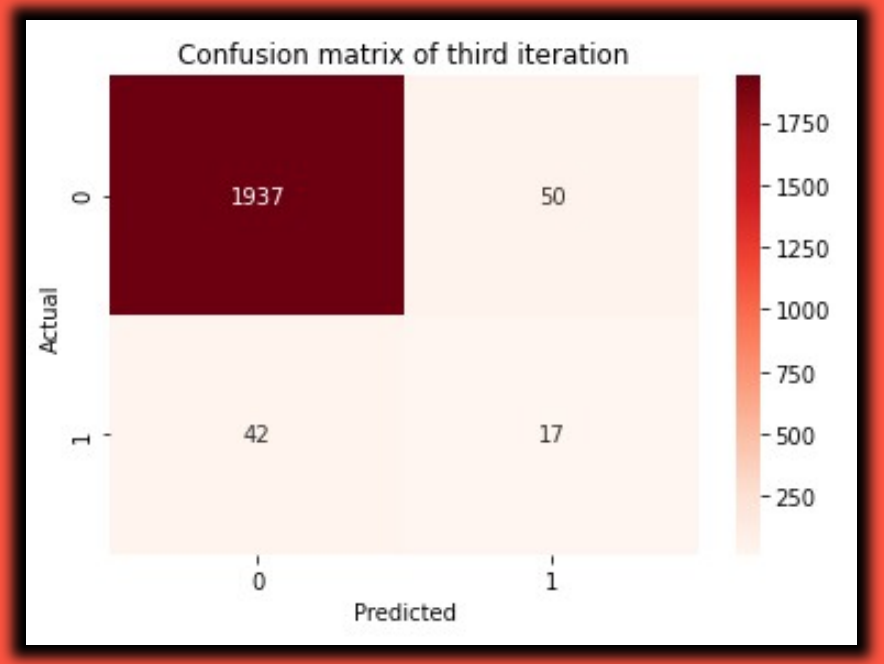
0%

Deuxième itération

01 Balancing (SMOTE) et scaling (RobustScaler()

02 BaggingClassifier() sans paramètrage

03 Résultat amélioré, mais toujours très décevant



	precision	recall	f1-score	support
0	0.98	0.97	0.98	1987
1	0.25	0.29	0.27	59
accuracy			0.96	2046
macro avg	0.62	0.63	0.62	2046
weighted avg	0.96	0.96	0.96	2046

Recall du modèle:

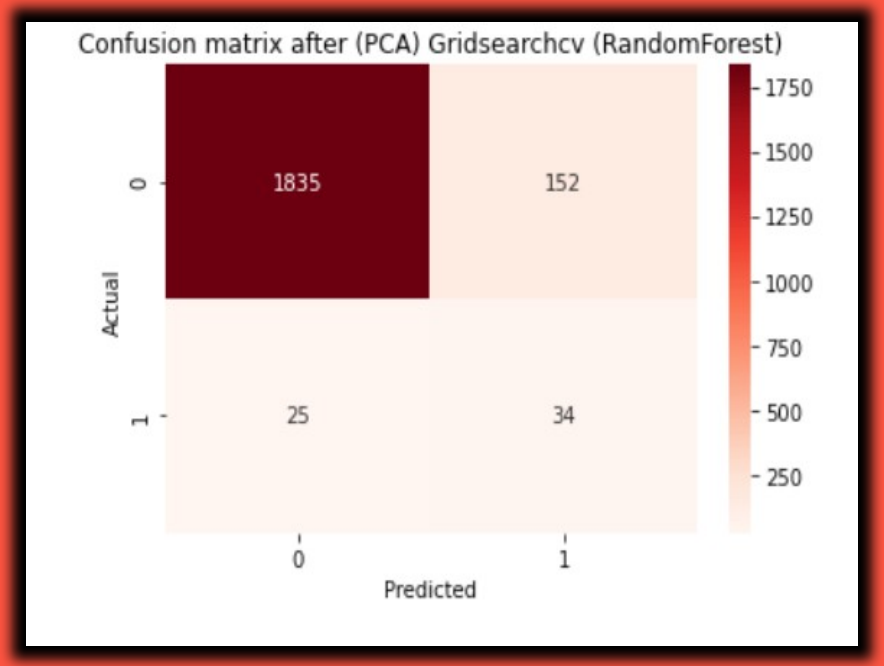
29%

Troisième itération

01 Balancing (SMOTE) et scaling (RobustScaler()

02 RandomForest() avec paramètrage GridSearchCV()

03 Résultat en net progrès sur le recall, mais une perte de précision sur les entreprises qui ne font pas faillites.



	precision	recall	f1-score	support
0	0.99	0.92	0.95	1987
1	0.18	0.58	0.28	59
accuracy			0.91	2046
macro avg	0.58	0.75	0.62	2046
weighted avg	0.96	0.91	0.93	2046

Recall du modèle:

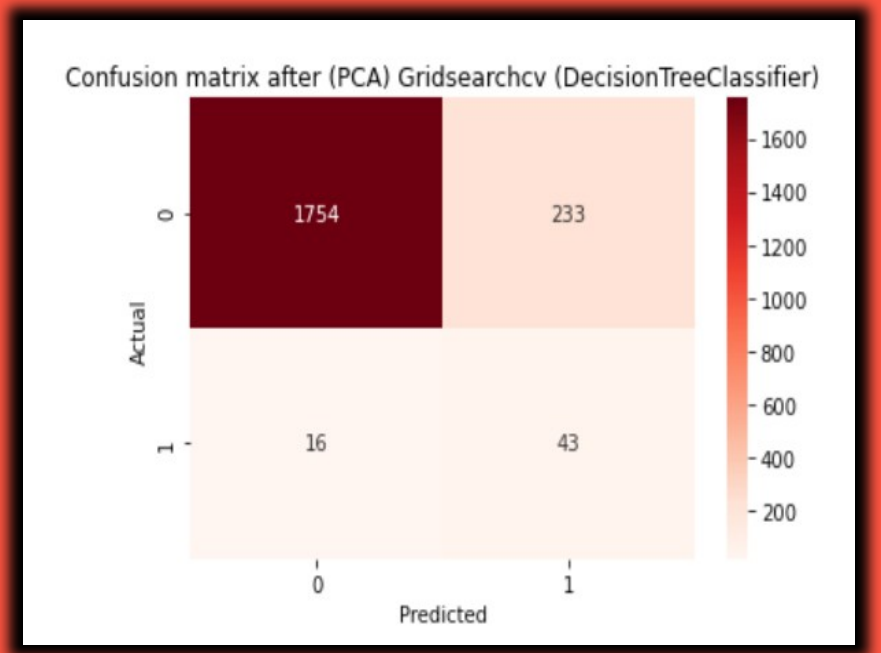
58%

Quatrième itération

01 Balancing (SMOTE) et scaling (RobustScaler()

02 BaggingClassifier() avec paramètrage GridSearchCV()

03 Résultat en nette amélioration sur le recall



	precision	recall	f1-score	support
0	0.99	0.88	0.93	1987
1	0.14	0.68	0.24	59
accuracy			0.87	2046
macro avg	0.57	0.78	0.58	2046
weighted avg	0.96	0.87	0.91	2046

Recall du modèle:

68%

V / Avenir du projet

Pistes à suivre

01

RandomForest

Meilleur paramè-trage avec Grid-SearchCV

02

Boosting

Tester le Boosting-Classifier pour voir si nos résultats s'améliorent

03

Preprocessing

Affinage avec ges-tion de la distribution de nos différentes valeurs

04

Stacking

Approfondir le stack-ing avec plus d'itéra-tions sur cette méth-ode.

Merci