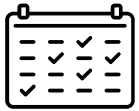
IA pour la prédiction

faillites d'entreprises

DATASET



Données

96 colonnes
6819 lignes

Target

Bankrupt

Classification binaire
Score recall en priorité

Aucunes données manquantes

Les colonnes catégoriques

	Bankrupt?	_Liability_Assets_Flag	_Net_Income_Flag
0	1	0	1
1	1	0	1
2	1	0	1
3	1	0	1
4	1	0	1
...	...	...	...
6814	0	0	1
6815	0	0	1
6816	0	0	1
6817	0	0	1
6818	0	0	1

[6819 rows x 3 columns]

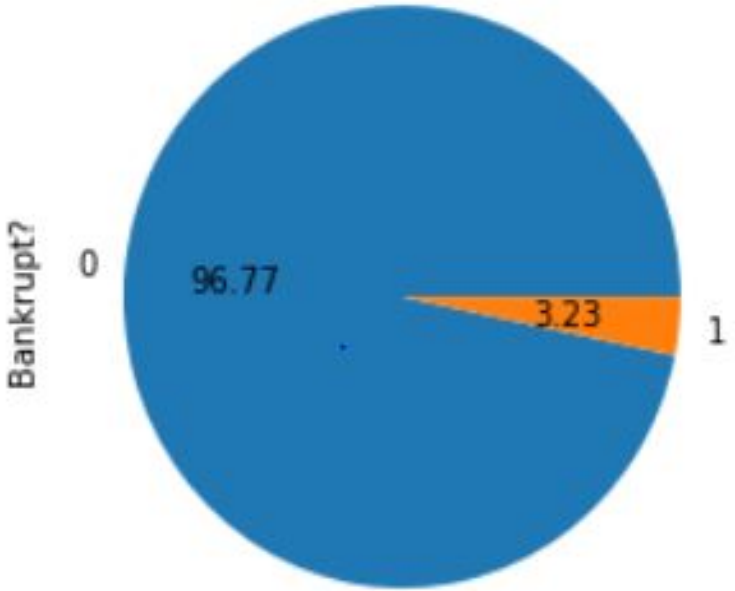
Techniques:

1.Suppression les colonnes:

- Liability-Assets Flag
- Net Income Flag

2.Bankrupt?: Le Target

Notre target



Techniques:

1.SMOTE()
2.Combinaison Under Simpling
Et over Simpling
3.Under Simpling

Matrice de corrélation / Feature permutation

	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)
Bankrupt?	1.000000	-0.260807	-0.282941	-0.273051	-0.100043	-0.099445	-0.000230	-0.008517	-0.008857	-0.016593	-0.008395
ROA(C) before interest and depreciation before interest	-0.260807	1.000000	0.940124	0.986849	0.334719	0.332755	0.035725	0.053419	0.049222	0.020501	0.051328
ROA(A) before interest and % after tax	-0.282941	0.940124	1.000000	0.955741	0.326969	0.324956	0.032053	0.053518	0.049474	0.029649	0.049909
ROA(B) before interest and depreciation after tax	-0.273051	0.986849	0.955741	1.000000	0.333749	0.331755	0.035212	0.053726	0.049952	0.022366	0.052261
Operating Gross Margin	-0.100043	0.334719	0.326969	0.333749	1.000000	0.999518	0.005745	0.032493	0.027175	0.051438	0.029430
Realized Sales Gross Margin	-0.099445	0.332755	0.324956	0.331755	0.999518	1.000000	0.005610	0.032232	0.026851	0.051242	0.029166
Operating Profit Rate	-0.000230	0.035725	0.032053	0.035212	0.005745	0.005610	1.000000	0.916448	0.862191	-0.592006	0.915544
Pre-tax net Interest Rate	-0.008517	0.053419	0.053518	0.053726	0.032493	0.032232	0.916448	1.000000	0.986379	-0.220045	0.993617
After-tax net Interest Rate	-0.008857	0.049222	0.049474	0.049952	0.027175	0.026851	0.862191	0.986379	1.000000	-0.115211	0.984452

Suppression des colonnes avec plus de 0.90 de corrélation (positive ou négative)

+

Suppression de 2 features suite aux résultats de la feature permutation

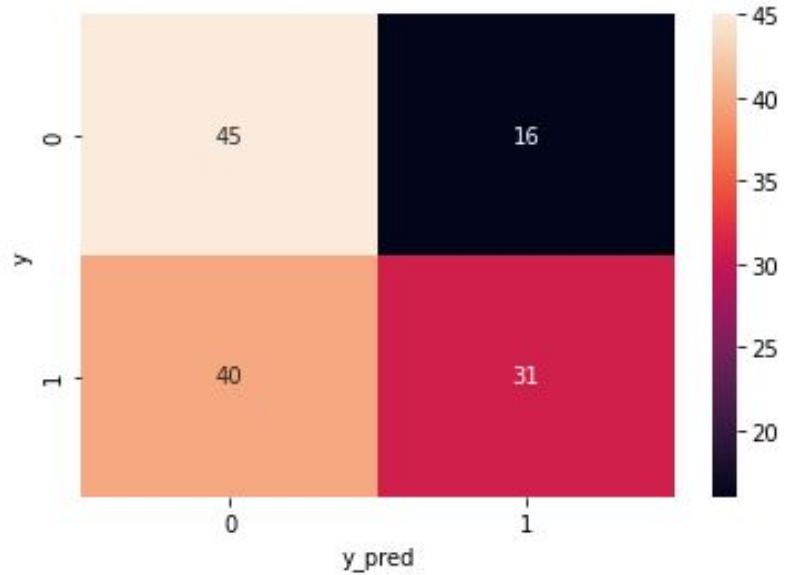
95 features -> 74 features

Nos itérations

Baseline model

Modèle

Regression logistique
Sur le dataset non-modifié
undersampling



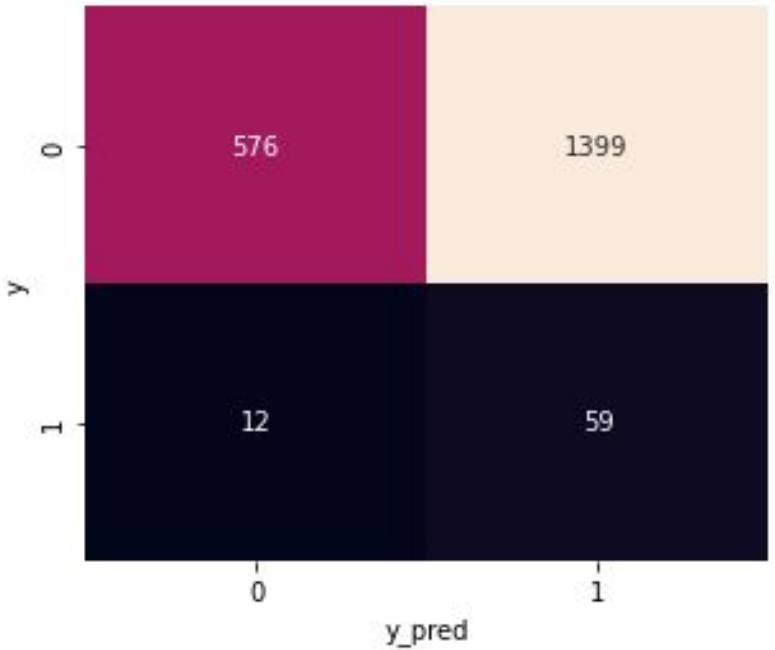
Résultats

Recall = 0.44

Modèle 1

Modèle

Regression logistique
SMOTE + SCALING (robust)



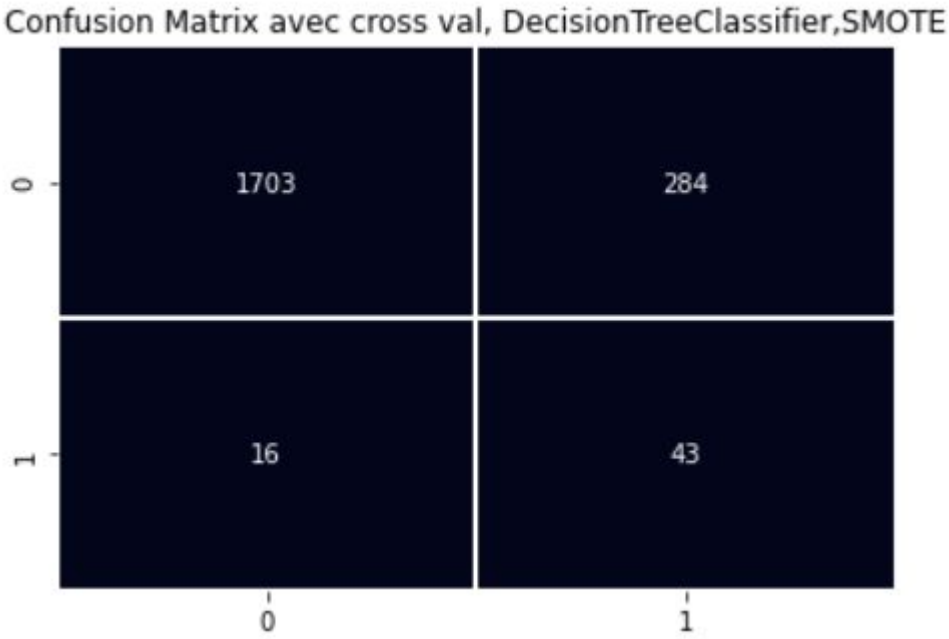
Résultats

Recall = 0.83

Modèle 2

Modèle

DecisionTreeClassifier
SMOTE + Scaling + cross validation



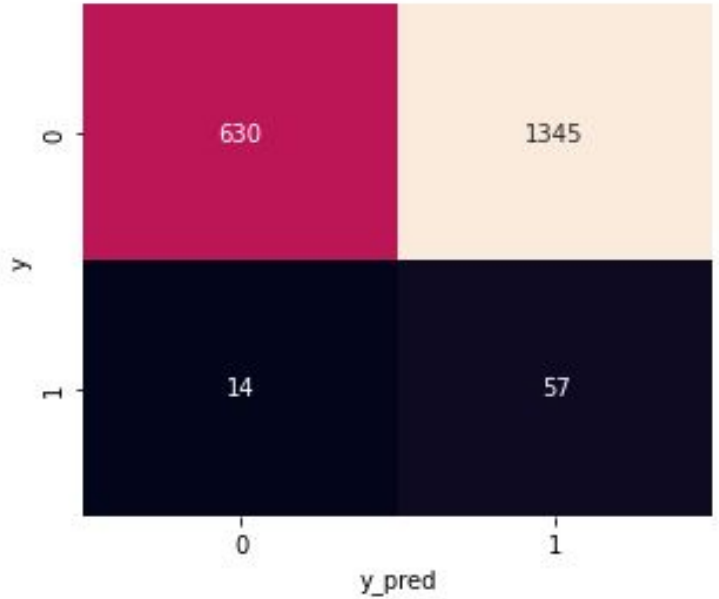
Résultats

Recall = 0.79

Modèle 3

Modèle

BaggingClassifier
Avec logistique regression



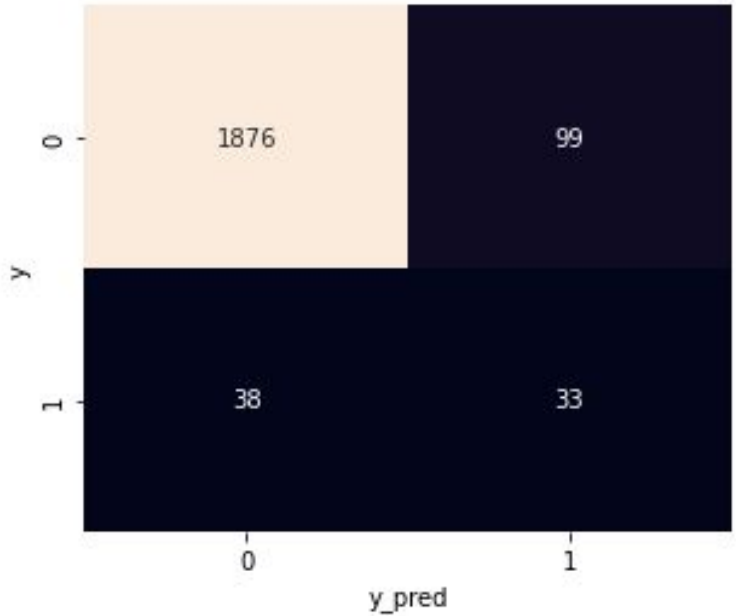
Résultats

Recall = 0.80

Modèle 4

Modèle

Boosting
Adaboost



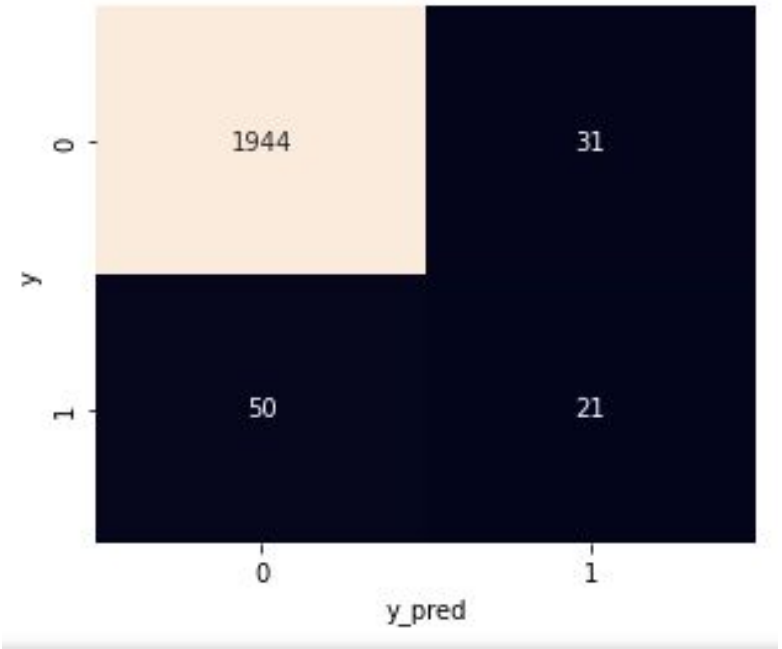
Résultats

Recall = 0.46

Modèle 5

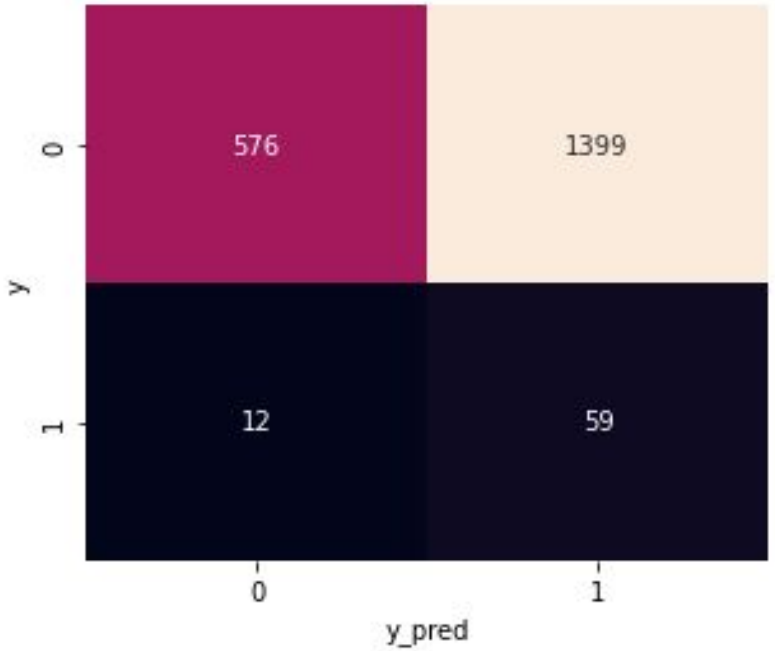
Modèle

Stacking
StackingClassifier
Estimator : LR / SVC / RF / KNC
Final estimator : KNeighborsClassifier()



Résultats

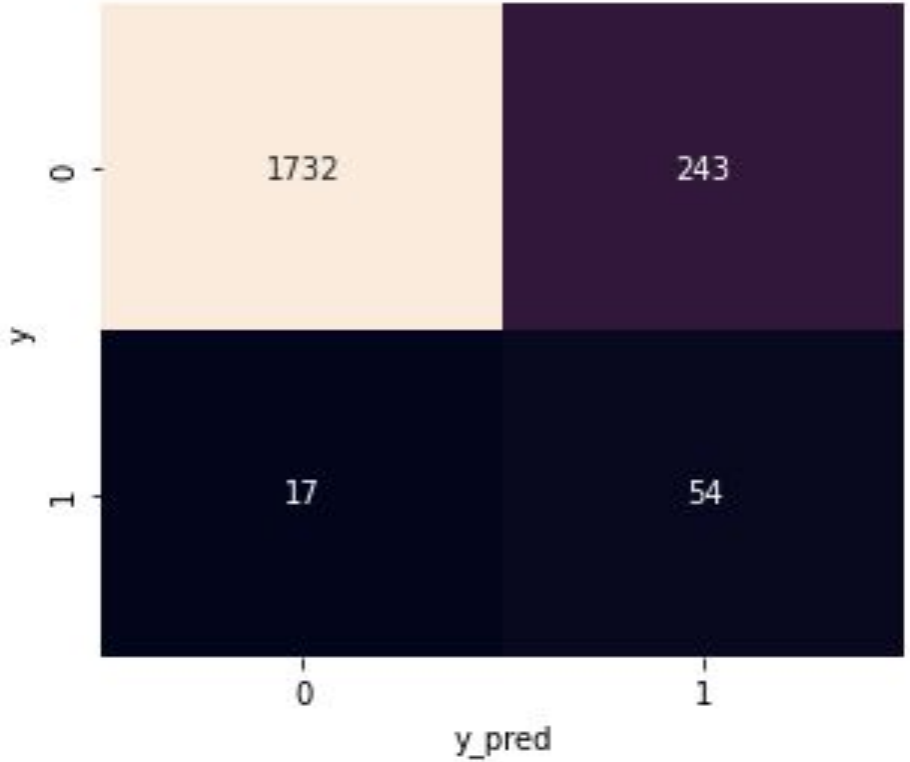
Recall = 0.2957



Conclusion

Regression logistique
SMOTE + SCALING (robust)
Recall = 0.83

2eme possibilité



Conclusion

2eme possibilité

Regression logistique
SMOTE + SCALING (MixMax)
Recall = 0.76

Merci !