PROJET

Prédiction des entreprises en faillites

Objectif

Client :

Société d’investissement

Problématique :

Créer un modèle de machine learning pour détecter la probabilité qu’une entreprise fasse faillite

EDA

Description du dataset d’origine

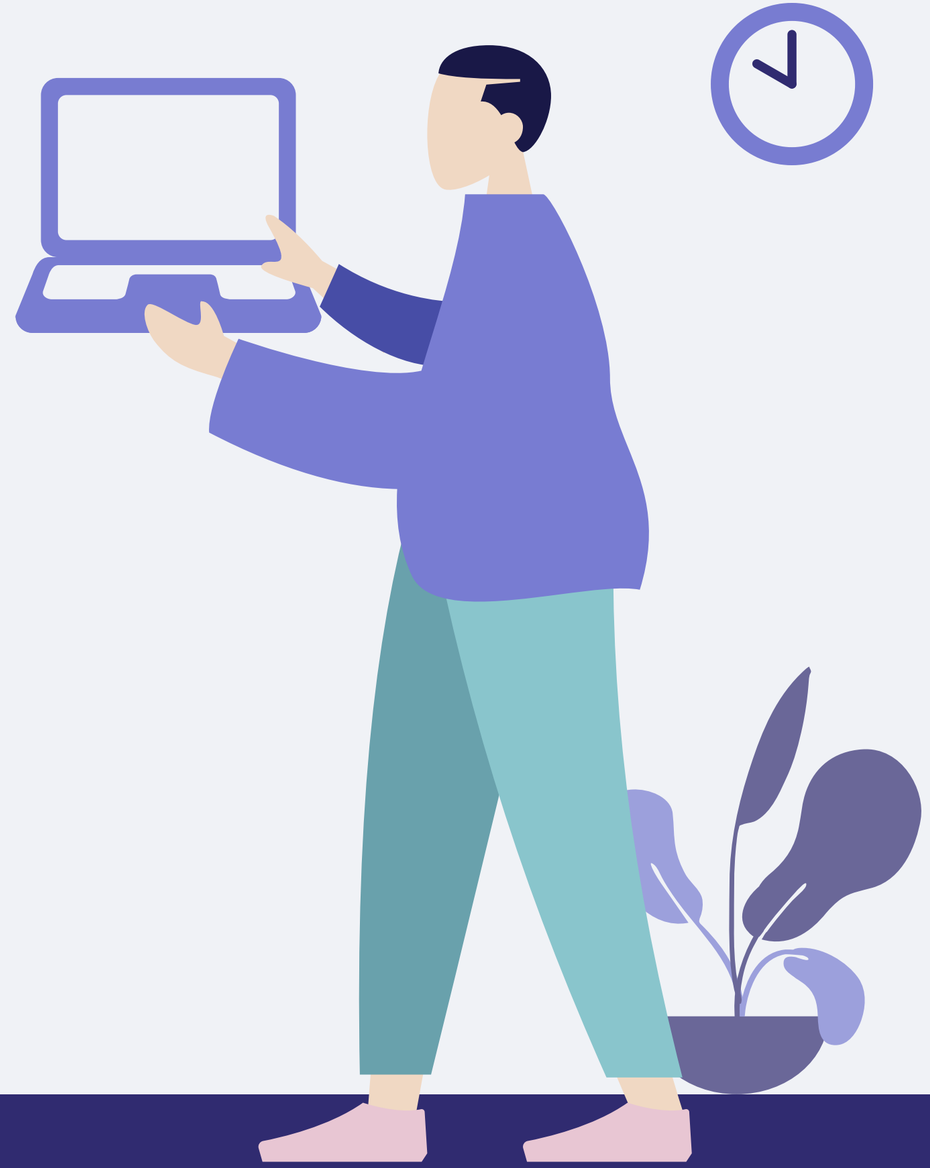
Target : 'Bankrupt?' (
→ 0 : non faillite
→ 1 : en faillite

Features : 95
→ indicateurs économiques des entreprises

Nombre d’observations : 6 819

Caractéristiques du dataset d’origine

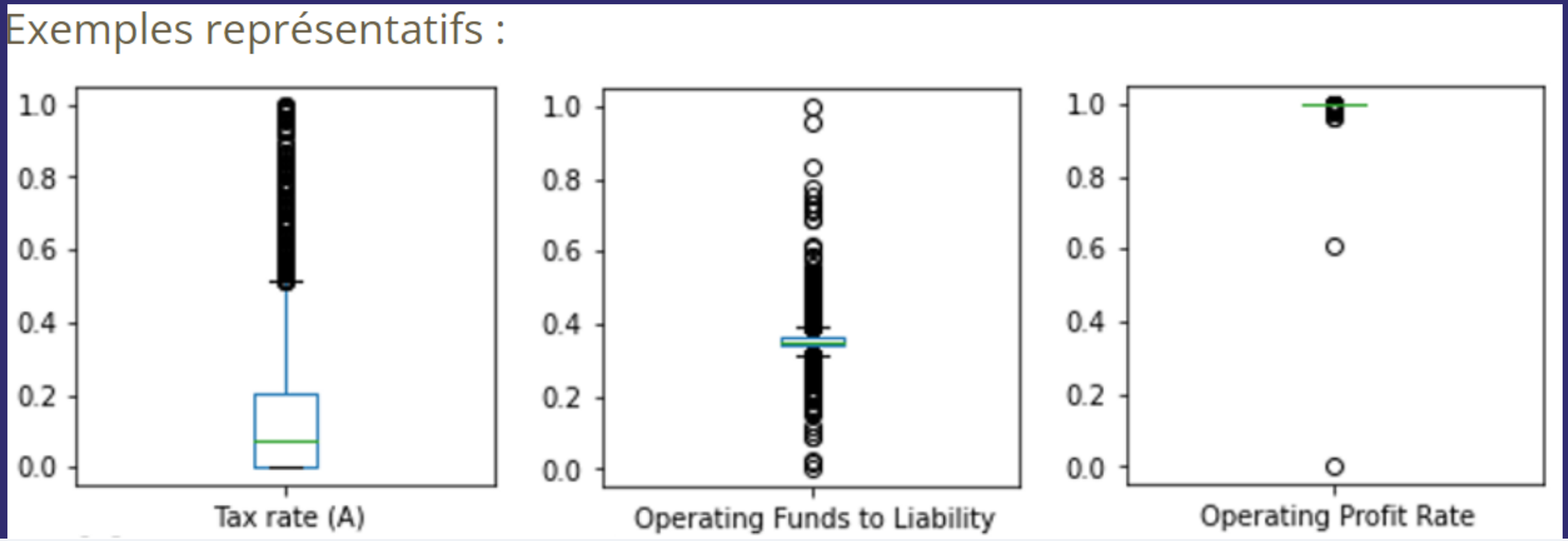
- Pas de valeurs manquantes
- Pas de doublons
- Feature ' Net Income Flag' (colonne de 1)



Caractéristiques du dataset d’origine : Nombreux outliers

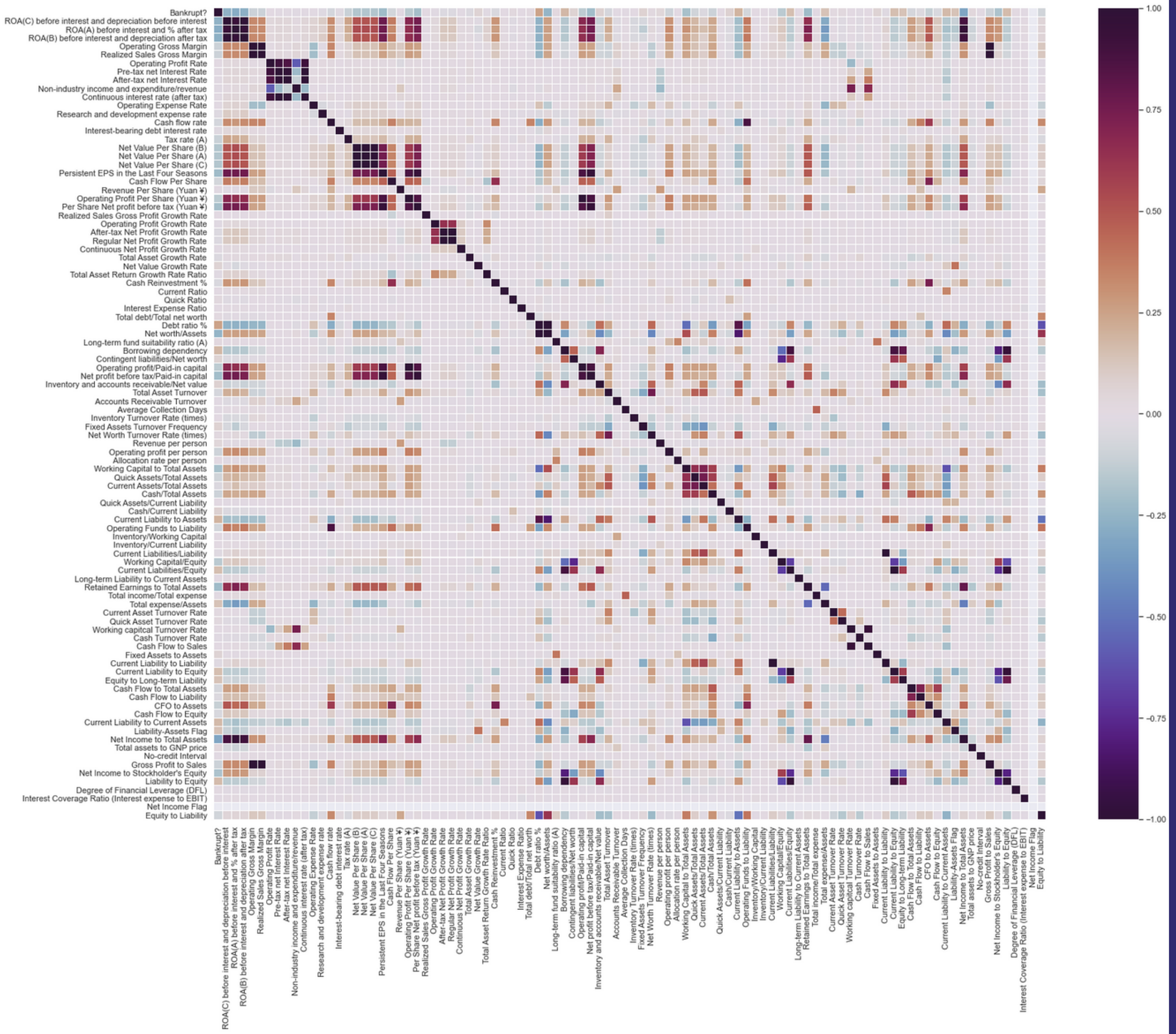
Beaucoup d’outliers (liés à des distributions non normales)

Exemples représentatifs :



Caractéristiques du dataset d’origine : " features corrélées "

Plusieurs variables corrélées entre elles : 43 couples de corrélation >0.8 et <-0.8
→ beaucoup de calculs à partir de d’autres du dataset



Exemples :

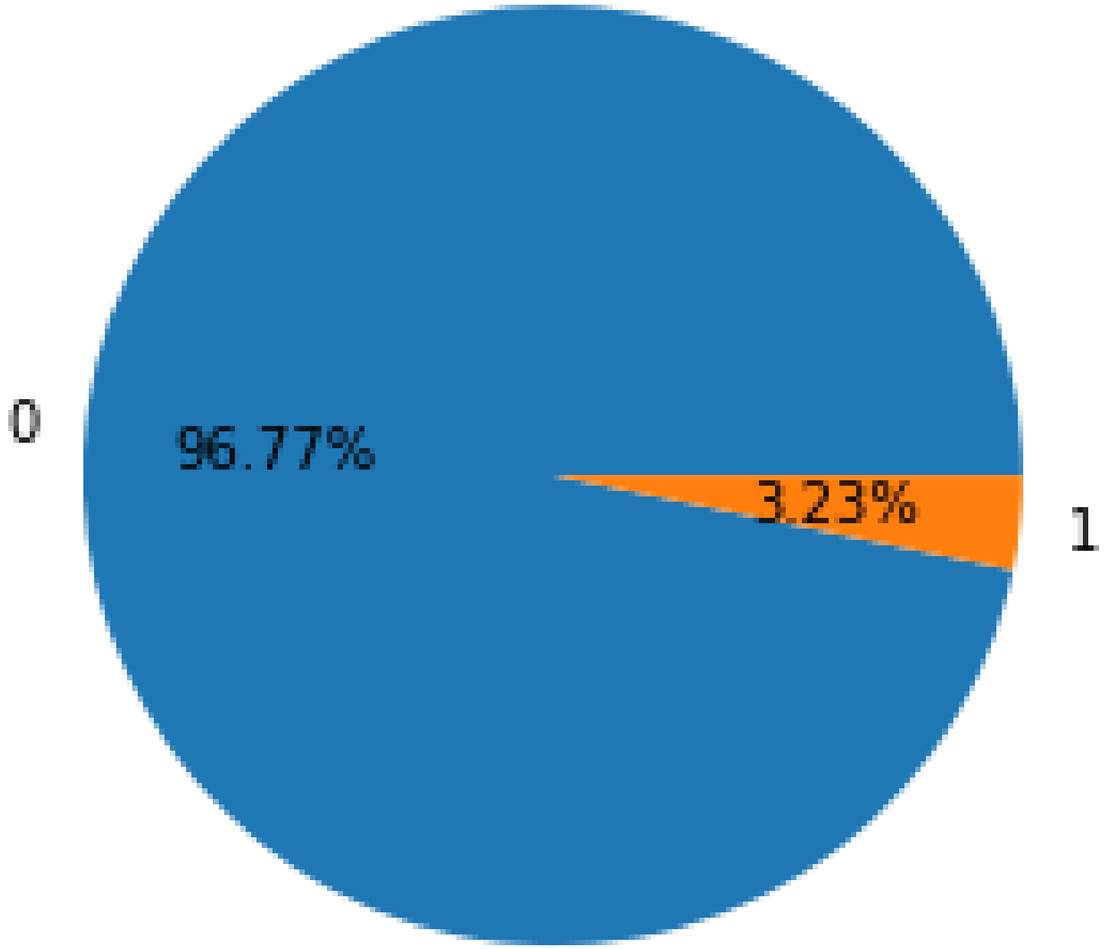
feature_1	feature_2	correlation
Current Liability to Liability	Current Liabilities/Liability	1.000000
Current Liability to Equity	Current Liabilities/Equity	1.000000
Gross Profit to Sales	Operating Gross Margin	1.000000
Net Value Per Share (C)	Net Value Per Share (A)	0.999837
Operating Gross Margin	Realized Sales Gross Margin	0.999518
Realized Sales Gross Margin	Gross Profit to Sales	0.999518
Net Value Per Share (B)	Net Value Per Share (A)	0.999342
Net Value Per Share (B)	Net Value Per Share (C)	0.999179
Operating Profit Per Share (Yuan ¥)	Operating profit/Paid-in capital	0.998696

Caractéristiques du dataset d’origine : "non balancé"

6599 Entreprises qui n’ont pas fait faillite (valeur 0)

220 Entreprise qui ont fait faillite (valeur 1)

Ce graphe montre que l’ensemble du dataset n’est pas équilibré.



Preprocessing

Etape préliminaire : "suppression des données erronées"

Suppression de la colonne : ' Net Income Flag' (colonne de 1)

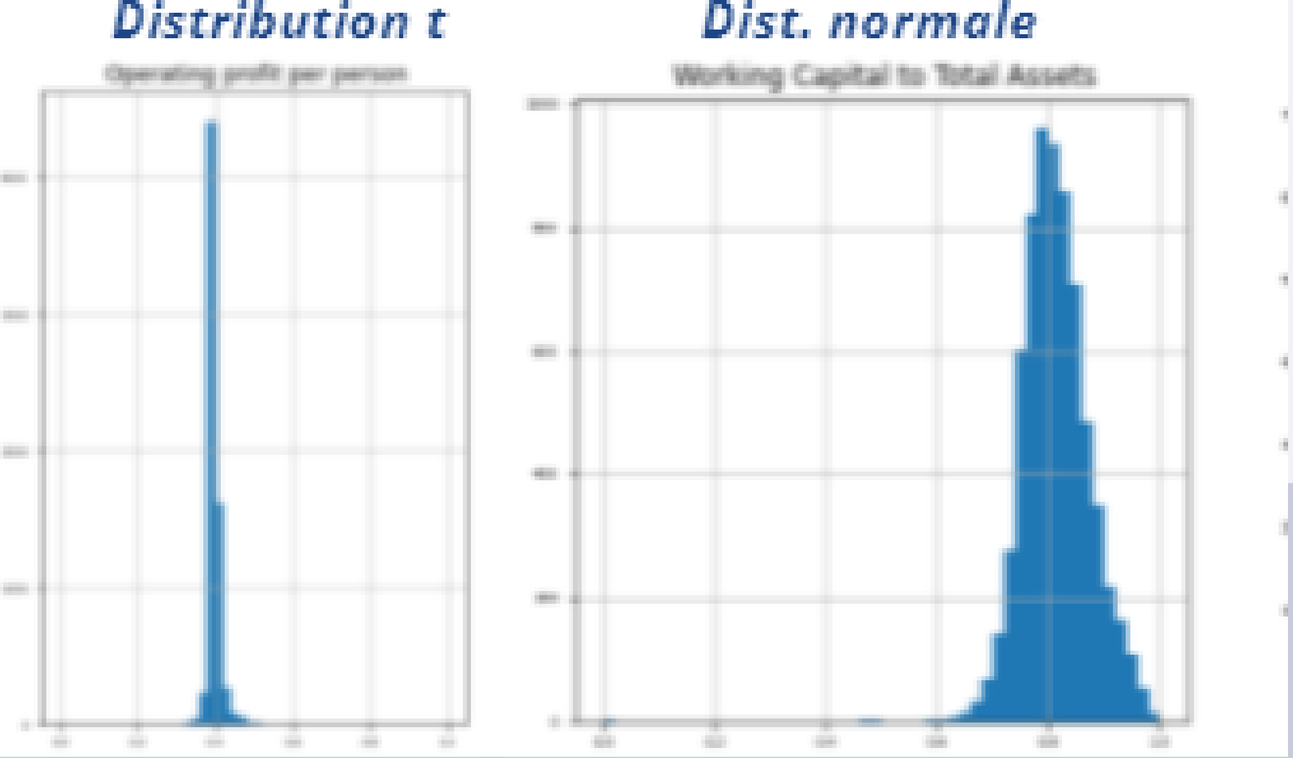
Mise à l’échelle (scaling)

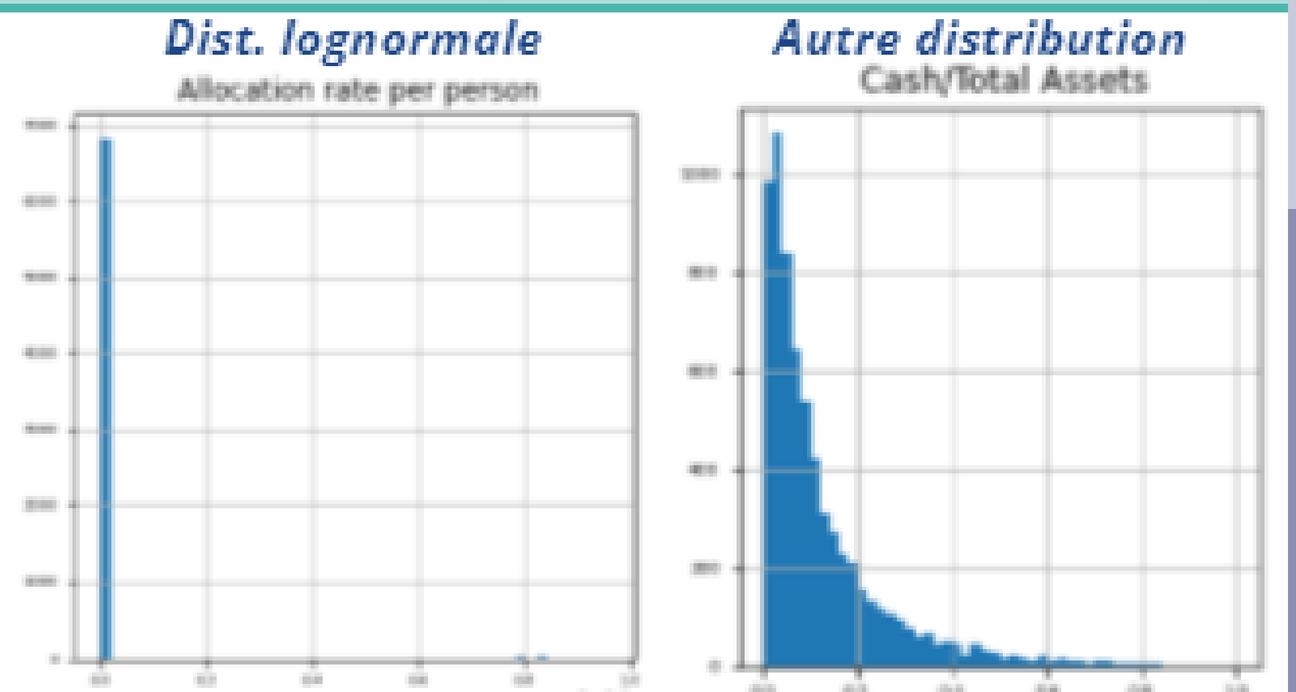
1. Détermination de la loi de distribution pour chaque features (librairie disfit)

2. Regroupement selon 4 catégories : distribution t, distribution normale, distribution lognormale et autre distribution

3. Scaling selon la loi de distribution suivie :
- distribution t : RobustScaler()
- distribution normale : StandardScaler()
- distribution lognormale : transformation au log + 1
- autre distribution : MinMaxScaler()

Mise à l'échelle (scaling)





1 Détermination de la loi de distribution pour chaque features (librairie disfit)

2 Regroupement selon 4 catégories : distribution t, distribution normale, distribution lognormale et autre distribution

3. scaling selon la loi de distribution suivie :
- distribution t : RobustScaler()
- distribution normale : StandardScaler()
- distribution lognormale : transformation au log + 1
- autre distribution : MinMaxScaler()

Balancing : SMOTE

1

Technique d’over-sampling

2

Réglé en 'auto' ou 'not majority' : rééchantillonne toutes les classes sauf la classe majoritaire (ici classe 0)

Modélisation

Choix du score **Recall**

Permet mieux prédire les entreprises en faillite, quitte à avoir un peu plus de faux positifs (entreprise non en faillite prédite comme en faillite)

Remarque :
la sélection est basée sur le recall mais pas uniquement pour éviter que le modèle prédise toujours une faillite

Tests de plusieurs modèles

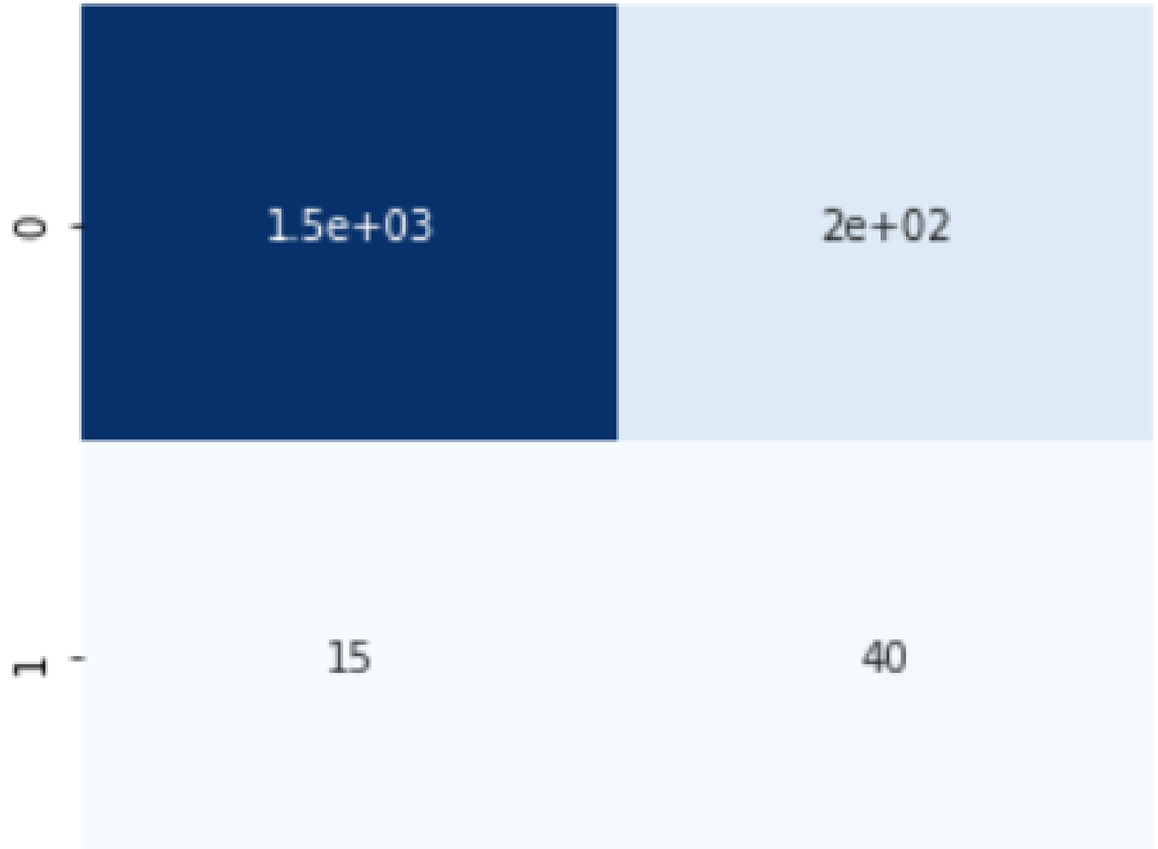
- Régression logistique
- KNN
- SVC
- Decision Tree
- Random Forest
- AdaBoost
- Stacking

Sélection du modèle de KNN

40 bonnes prédictions de faillite contre 15

Recall : 0.73 et f1-score : 0.27

	precision	recall	f1-score	support
0	0.99	0.88	0.93	1650
1	0.17	0.73	0.27	55
accuracy			0.88	1705
macro avg	0.58	0.80	0.60	1705
weighted avg	0.96	0.88	0.91	1705



Conclusion

Sélection du modèle KNN avec valeur plutôt satisfaisante de recall

Il reste à le perfectionner, pistes possibles :

- **Sélection de features (gestion des variables corrélées notamment)**
- **Traitement des outliers (notamment features normales, lognormales et t)**
- **Mieux ré-attribuer les distributions de features**
- **RandomizedSearchCV et GridSearchCV pour mieux régler les hyperparamètres**
- **Tester d’autres paramètres de SMOTE**