Projet Machine Learning : Prédiction de Faillite d'Entreprise

Auteurs : Matthieu Hanna Gerguis, Renaud de l'Epine, Ilian Segoin
Classe : IF3, ESILV
Date : Novembre 2025

I. Introduction

Prédire la faillite d'une entreprise est un enjeu majeur en finance, car une défaillance non anticipée peut entraîner des pertes importantes pour les créanciers, les investisseurs et l'économie. Notre objectif dans ce projet est de concevoir un pipeline de machine learning capable d'identifier les entreprises à risque à partir d'indicateurs financiers historiques, en mettant l'accent sur la détection précoce des signaux de détresse.

Le jeu de données présente une difficulté importante : la classe *faillite* représente seulement environ 3% des observations, créant un fort déséquilibre et faisant du *recall* la métrique la plus critique.

Ce pré-projet introduit le contexte, la source des données, la méthodologie et les résultats attendus.

II. Contexte Opérationnel

Les institutions financières s'appuient sur des systèmes d'alerte pour anticiper les situations de détresse financière. Ne pas détecter une faillite peut entraîner des défauts de paiement, des réactions en chaîne dans le tissu économique et des évaluations erronées du risque. Les modèles de machine learning permettent d'analyser de grands volumes de données financières et d'identifier des schémas difficiles à percevoir avec des méthodes traditionnelles.

Notre projet vise donc à développer un modèle capable d'identifier les signaux précurseurs de la défaillance d'entreprise.

III. Description du Jeu de Données

Le jeu de données provient du Taiwan Economic Journal (TEJ) et est disponible sur Kaggle sous le nom *Company Bankruptcy Prediction*, accessible à l'adresse : `https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction`.

Il contient des indicateurs financiers enregistrés entre 1999 et 2009 et comprend :

- 6 819 entreprises,
- 95 variables financières numériques,
- 1 variable cible binaire « Bankrupt? ».

Toutes les variables sont numériques et aucune valeur manquante n'est présente, ce qui simplifie le prétraitement. Cependant, le jeu de données est difficile en raison de sa forte dimension, de la corrélation entre les variables et de son déséquilibre extrême.

Ces caractéristiques justifient l'utilisation de filtrage par variance, d'analyse de corrélation, de réduction dimensionnelle et de techniques de rééquilibrage.

IV. Objectifs du Projet

L'objectif principal est de construire un modèle de prédiction de faillite fiable et interprétable. Pour cela, le projet inclut :

- l'exploration et l'analyse du jeu de données,
- le nettoyage et la réduction dimensionnelle via variance et corrélations,
- la standardisation de toutes les variables et l'application d'une PCA,
- l'entraînement de modèles de base (Logistic Regression, Naive Bayes, Random Forest),
- la gestion du déséquilibre grâce à l'oversampling et à SMOTE,
- l'évaluation centrée sur le recall de la classe minoritaire,
- l'optimisation d'hyperparamètres pour les meilleurs modèles.

Le modèle final doit fournir des indications interprétables sur les variables financières les plus liées à la faillite.

V. Aperçu Méthodologique

A. Analyse Exploratoire

Analyse de la variance, des distributions, des zéros et des corrélations afin d'identifier les variables inutiles ou redondantes.

B. Prétraitement

Standardisation des variables et application d'une PCA conservant 95% de la variance.

C. Modélisation

Trois modèles de base sont testés avec et sans PCA :

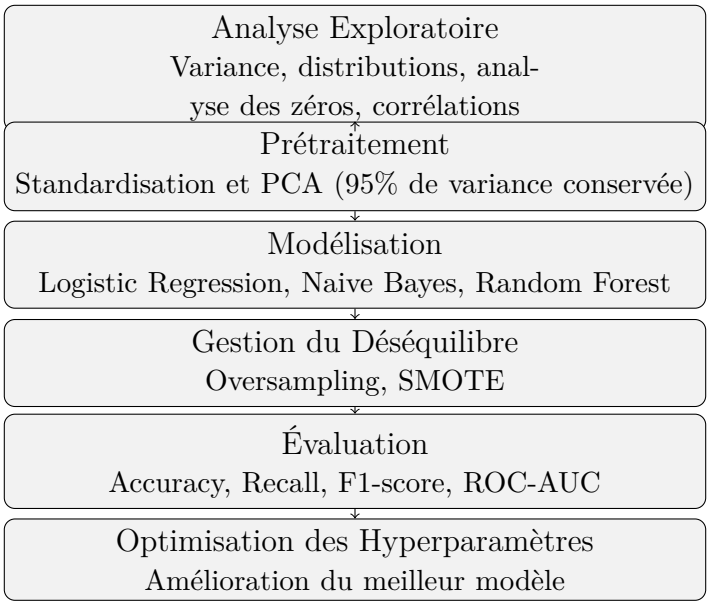
- Logistic Regression,
- Naive Bayes,
- Random Forest.

D. Gestion du Déséquilibre

Techniques telles que Random Oversampling et SMOTE pour améliorer le recall de la classe minoritaire.

E. Évaluation

Évaluation via accuracy, recall, précision, F1-score et ROC-AUC, avec une attention particulière portée à la détection des faillites.



VI. Apports Attendus

Ce projet vise à fournir :

- un pipeline complet, clair et reproductible,
- des indications sur les indicateurs financiers les plus pertinents,
- une analyse de l'impact du prétraitement et du rééquilibrage sur les performances.

VII. Conclusion

Ce pré-projet résume les motivations, les données et la méthodologie employées pour la prédiction de faillite. Le projet complet évaluera les limites des modèles de base, l'impact des techniques de rééquilibrage sur le recall et explorera des modèles plus avancés tels que XGBoost ou les modèles calibrés.

L'objectif final est de construire un système robuste, interprétable et utile pour l'aide à la décision financière, tout en soulignant ses limites et pistes d'amélioration.