



Université Mohammed V - Rabat  
École Nationale Supérieure d'Informatique  
et d'Analyse des Systèmes



# DATA DRIVEN DECISION MAKING RAPPORT

FILIÈRE : GÉNIE LOGICIEL

---

## Prédiction de la Faillite des Entreprises par Analyse des Données Financières

---

*Réalisé par :*  
YOUSFI WIAME  
BOURAOU YOUSSEF

*Encadré par :*  
M.TABII YOUNESS

Année Universitaire 2024-2025

# Table des matières

<b>Remerciements</b>	<b>5</b>
<b>Résumé</b>	<b>6</b>
<b>Abstract</b>	<b>7</b>
<b>Introduction générale</b>	<b>8</b>
<b>1 Contexte Général du Projet</b>	<b>9</b>
1.1 Contexte du Projet	9
1.1.1 Contexte Général	9
1.1.2 Importance de la Prédiction de Faillite d'Entreprise	9
1.1.3 Enjeux Actuels et Perspectives	10
1.2 Objectifs du Projet	10
1.3 Contexte Général	10
1.4 Conclusion	11
<b>2 Collecte et Préparation des Données</b>	<b>12</b>
2.1 Source de Données	12
2.2 Description des Données	13
2.3 Prétraitement des Données	15
2.4 Analyse Exploratoire des Données	16
2.4.1 Taille du jeu de données	16
2.4.2 Aperçu du jeu de données	16
2.4.3 Vérification des données nulles	16
2.4.4 Répartition du nombre d'années par entreprise	18
2.4.5 Répartition des entreprises par statut	18
2.5 Conclusion	19
<b>3 Modélisation</b>	<b>20</b>
3.1 Choix des Algorithmes	20
3.1.1 Distinction entre les Modèles Classiques et les Modèles Deep Learning	21
3.2 Entraînement des modèles	22
3.2.1 Modèles du deep Learning	22
3.2.1.1 Premier Entraînement avec le jeu de données déséquilibré sur des séquences de deux années	22
3.2.2 Entraînement des modèles avec un jeu de données équilibré sur des séquences de deux années	24
3.2.2.1 Choix du modèle pour les séquences de deux années	28

3.2.3	L'effet de la longueur des séquences . . . . .	28
3.2.4	Choix du modèle pour les séquence de 5ans . . . . .	31
3.2.4.1	Optimisation du BiLSTM . . . . .	32
3.2.5	Modèles du machine learning . . . . .	34
3.2.6	Entraînement sur le jeu de données déséquilibré . . . . .	34
3.2.7	Entraînement sur le jeu de données équilibré . . . . .	36
3.2.8	Choix du modèle le plus équilibré : Random Forest . . . . .	38
3.3	Résultats . . . . .	38
3.4	Analyse des Résultats . . . . .	38
3.5	Élaboration des Stratégies . . . . .	39
3.5.1	Stratégies de Développement . . . . .	39
3.5.2	Actions Recommandées . . . . .	40
<b>Conclusion générale . . . . .</b>		<b>41</b>

# Table des figures

2.1	Taille du jeu de données	16
2.2	Données sur les entreprises	16
2.3	Pas de valeurs nulles	17
2.4	Fréquence du nombre d'observations (années) par entreprise	18
2.5	Répartition des entreprises par leur statut financier	19
3.1	LSTM	22
3.2	GRU	22
3.3	CNN 1D	22
3.4	BiLSTM	23
3.5	LSTM optimisé	23
3.6	GRU optimisé	23
3.7	CNN 1D optimisé	24
3.8	BiLSTM optimisé	24
3.9	LSTM sur un jeu de données équilibré	25
3.10	GRU sur un jeu de données équilibré	25
3.11	CNN 1 sur un jeu de données équilibré	25
3.12	BiLSTM sur un jeu de données équilibré	26
3.13	LSTM	26
3.14	GRU	27
3.15	CNN 1D	27
3.16	BiLSTM	27
3.17	LSTM	28
3.18	GRU	29
3.19	CNN 1D	29
3.20	BiLSTM	30
3.21	LSTM	30
3.22	GRU	31
3.23	CNN 1D	31
3.24	BiLSTM	31
3.25	BiLSTM	32
3.26	Courbes de loss et accuracy	33
3.27	Random Forest Et Logistic Regression	34
3.28	MLP et xgboost	35
3.29	Random Forest et Logistic Regression	36
3.30	MLP et xgboost	37

# Liste des tableaux

2.1	Description des variables comptables du dataset . . . . .	14
3.1	Comparaison entre les approches classiques (ML) et séquentielles (DL) pour la prédiction de faillite . . . . .	21

## Remerciements

Avant de détailler notre projet, il est primordial de commencer par exprimer notre profonde gratitude envers ceux qui nous ont tant appris cette année. Nous adressons d'abord nos remerciements sincères à notre encadrant académique, **M. Tabii Youness**, pour avoir accepté de guider notre projet de systèmes décisionnels et pour le soutien crucial qu'il nous a apporté tout au long de ce parcours.

Nous tenons également à remercier chaleureusement l'ensemble de l'équipe pédagogique de l'École Nationale Supérieure d'Informatique et d'Analyse des Systèmes (ENSIAS) pour la qualité remarquable de leur formation. Un merci tout particulier à **M. Guermah Hatim**, notre chef de filière, pour son assistance continue durant cette année. Nous leur adressons nos salutations les plus respectueuses.

À tous ceux qui ont contribué de près ou de loin à l'aboutissement de ce projet, nous exprimons notre gratitude la plus sincère.

## Résumé

Ce rapport présente une étude approfondie sur la prédiction de la faillite d'entreprise fondée sur l'exploitation de données financières historiques. Face à la complexité croissante des données et à l'instabilité économique mondiale, nous développons une approche méthodologique structurée visant à identifier les signaux avant-coureurs de défaillance financière.

Notre démarche repose sur quatre axes principaux : la préparation minutieuse des données comptables, comprenant leur nettoyage, structuration et enrichissement ; l'analyse des tendances financières pour identifier les schémas récurrents annonciateurs de difficultés ; l'évaluation comparative de différentes méthodes d'analyse prédictive tenant compte de la diversité des profils d'entreprises ; et la conception d'un cadre opérationnel d'exploitation des résultats pour guider les décisions stratégiques.

Cette recherche s'inscrit dans une logique d'aide à la décision et de prévention, avec des applications concrètes pour la sécurisation des investissements, la gestion proactive du risque de crédit et l'amélioration des pratiques d'audit. En permettant une détection précoce des entreprises à risque, notre travail contribue à limiter les effets systémiques des faillites en chaîne et à renforcer la stabilité économique globale. Les résultats obtenus démontrent l'efficacité des approches d'apprentissage automatique pour traiter cette problématique complexe et multifactorielle, ouvrant la voie à une nouvelle génération d'outils analytiques au service des décideurs économiques.

---

**Mots clés : prédiction de faillite, données financières, apprentissage automatique, gestion du risque, stabilité économique, indicateurs de défaillance**

---

## Abstract

This report presents a comprehensive study on corporate bankruptcy prediction based on historical financial data analysis. In a context of increasing economic instability and data complexity, anticipating business failures has become a strategic challenge for investors, financial institutions, and regulatory bodies.

Our research develops a predictive solution for identifying early warning signals of financial distress through a structured methodological approach. The study encompasses four main objectives : thorough preparation of accounting data, analysis of financial trends to identify recurrent patterns preceding bankruptcy, comparative evaluation of various predictive methods adapted to diverse company profiles, and design of an operational framework to guide strategic decision-making.

By leveraging machine learning techniques applied to financial indicators, our work provides a decision support tool capable of detecting vulnerable companies before critical situations arise. This approach contributes to investment security, proactive credit risk management, and improved audit practices. The findings demonstrate the effectiveness of data-driven methods in addressing this complex and multifactorial issue, with potential applications extending beyond individual risk assessment to broader economic stability preservation at both national and international levels.

Our research ultimately aims to strengthen the analytical tools available to economic decision-makers, thereby enhancing market stability and potentially preserving thousands of jobs through early intervention in financially distressed companies.

---

**Keywords :** bankruptcy prediction, financial data analysis, machine learning, risk management, early warning systems, corporate financial distress

---



# Introduction générale

Dans un environnement économique caractérisé par une instabilité croissante, la santé financière des entreprises constitue un indicateur fondamental de la stabilité économique globale. La capacité à anticiper les défaillances financières des entreprises représente un enjeu majeur pour l'ensemble des acteurs économiques : investisseurs, institutions financières, organismes de régulation et gestionnaires de risques. À l'ère du numérique et de l'intelligence artificielle, la prédiction de la faillite d'entreprise s'est imposée comme un domaine de recherche stratégique à l'intersection de la finance et de la science des données.

Ce rapport s'attache à développer une solution prédictive robuste permettant d'identifier, à partir de données financières historiques, les signaux précurseurs d'une détérioration de la santé financière susceptible de conduire à une faillite. Cette démarche s'inscrit dans une volonté d'anticipation économique fondée sur une compréhension approfondie de l'évolution des indicateurs financiers. En repérant les schémas récurrents et les variables critiques, notre étude vise à établir un cadre opérationnel d'aide à la décision pour les différentes parties prenantes.

Au-delà de son utilité immédiate pour la gestion des risques, ce projet contribue à une meilleure stabilité des marchés financiers, à la sécurisation des investissements, et potentiellement à la préservation de nombreux emplois grâce à l'identification précoce des entreprises en difficulté. Dans un contexte où les crises économiques peuvent avoir des effets domino dévastateurs, la prédiction de faillite constitue un levier stratégique pour renforcer la résilience économique tant au niveau national qu'international.

# Chapitre 1

## Contexte Général du Projet

### 1.1 Contexte du Projet

#### 1.1.1 Contexte Général

La santé financière des entreprises est un indicateur fondamental de stabilité économique. La capacité de prédire la faillite d'une entreprise permet aux investisseurs, aux institutions financières et aux organismes de régulation de prendre des décisions éclairées et préventives. Avec la montée des données financières numériques et des technologies d'apprentissage automatique, la prédiction de la faillite est devenue un domaine de recherche clé dans la science des données appliquée à la finance.

#### 1.1.2 Importance de la Prédiction de Faillite d'Entreprise

**Économie Nationale** La prédiction de la faillite a un impact direct sur la stabilité de l'économie d'un pays :

- **Prévention des Crises Financières** : En identifiant précocement les entreprises à risque, les institutions financières peuvent éviter les effets domino provoqués par des faillites en chaîne.
- **Sécurisation des Investissements** : Les investisseurs peuvent utiliser ces prédictions pour orienter leurs décisions de placement et minimiser les pertes.
- **Stabilité de l'Emploi** : Anticiper les difficultés financières permet de mettre en place des stratégies de sauvetage, préservant ainsi des milliers d'emplois.

**Économie Internationale** Sur le plan global, la prédiction de la faillite est essentielle pour :

- **La Réduction du Risque Systémique** : Les grandes entreprises opérant à l'international peuvent avoir des répercussions globales en cas de faillite.
- **L'Optimisation des Marchés Financiers** : Des outils prédictifs fiables favorisent la transparence et l'efficacité des marchés.

- **Le Respect des Normes de Conformité** : La détection précoce des risques de faillite aide les institutions à se conformer aux normes réglementaires internationales.

### 1.1.3 Enjeux Actuels et Perspectives

Dans un monde économique instable, caractérisé par des cycles de crises et de reprises, les enjeux autour de la détection précoce de la faillite sont multiples :

- **Complexité des Données Financières** : L'explosion des données, leur diversité et leur granularité rendent l'analyse traditionnelle inefficace.
- **Besoin de Prédictions Fiables** : Il est impératif de mettre en place des outils capables de détecter des signaux faibles annonciateurs de défaillance financière.
- **Utilité Stratégique** : La détection précoce de la faillite permet une meilleure planification, gestion du risque, et allocation des ressources pour les entreprises et les investisseurs.

Cette problématique s'inscrit dans une volonté d'anticipation économique fondée sur une compréhension approfondie de l'évolution des indicateurs financiers. Elle permet de mieux appréhender les vulnérabilités et d'orienter les actions de prévention ou d'intervention.

## 1.2 Objectifs du Projet

Ce projet vise à construire une solution prédictive de détection de la faillite des entreprises en s'appuyant sur des données financières historiques. Les objectifs spécifiques sont :

1. **Préparation des Données** : Nettoyage, structuration et enrichissement des données comptables collectées auprès des entreprises.
2. **Analyse des Tendances Financières** : Identifier les schémas récurrents et les variables susceptibles d'indiquer une défaillance future.
3. **Évaluation Comparative des Approches** : Étudier plusieurs méthodes d'analyse, en tenant compte de la diversité des profils d'entreprises et de la disponibilité des données.
4. **Mise en Oeuvre d'un Cadre Opérationnel** : Concevoir un cadre d'exploitation des résultats permettant de guider les décisions financières et stratégiques.

## 1.3 Contexte Général

L'objectif principal de ce projet est de mieux anticiper les situations de faillite d'entreprises à partir de leurs données financières historiques. Dans un contexte économique instable, cette capacité de prédiction représente un levier stratégique pour de nombreux acteurs économiques : investisseurs, institutions financières, régulateurs ou encore gestionnaires de risques.

L'analyse des signaux annonciateurs d'une détérioration financière repose sur l'exploitation de tendances historiques, de ratios comptables et de comportements économiques passés. L'enjeu est de repérer, à travers l'évolution des indicateurs, les signes précurseurs d'un déséquilibre pouvant mener à une défaillance.

Ce projet s'inscrit dans une logique d'aide à la décision et de prévention. Il ambitionne de contribuer à :

- la sécurisation des investissements à travers une meilleure visibilité sur la solvabilité des entreprises ;
- la gestion proactive du risque de crédit par l'identification anticipée des entreprises les plus vulnérables ;
- l'amélioration des pratiques d'audit et de contrôle en orientant les efforts vers les entreprises les plus à risque.

Enfin, cette démarche s'inscrit dans un processus global de renforcement des outils analytiques mis à disposition des décideurs économiques, avec pour finalité une meilleure stabilité des marchés.

## 1.4 Conclusion

Ce premier chapitre a permis de poser les bases générales du projet centré sur la prédiction de la faillite des entreprises à partir de leurs données financières. En replaçant cette problématique dans un cadre économique global, nous avons souligné son importance stratégique pour les différents acteurs concernés, aussi bien au niveau national qu'international.

Les objectifs du projet s'orientent ainsi vers l'analyse des trajectoires financières des entreprises afin d'anticiper les situations de risque et d'améliorer la prise de décision. Le chapitre suivant présentera en détail les sources de données mobilisées, ainsi que les étapes de collecte, de préparation et de structuration indispensables pour rendre possible une telle analyse.

# Chapitre 2

## Collecte et Préparation des Données

### 2.1 Source de Données

Les données utilisées dans ce projet proviennent d'un jeu de données novateur intitulé "*A Novel Dataset for Bankruptcy Prediction*", publié sur la plateforme **Kaggle**. Ce dataset porte sur des entreprises américaines cotées à la **Bourse de New York (NYSE)** et au **NASDAQ**. Il contient des données comptables précises relatives à **8 262** entreprises distinctes, couvrant une période allant de **1999 à 2018**.

Selon la *Security Exchange Commission* (SEC), une entreprise est déclarée en faillite dans l'un des deux cas suivants :

- Dépôt de bilan sous le **Chapitre 11 du Code de la Faillite** : il s'agit d'une réorganisation, où la direction de l'entreprise conserve la gestion opérationnelle, mais toutes les décisions importantes nécessitent l'approbation d'un tribunal.
- Dépôt de bilan sous le **Chapitre 7 du Code de la Faillite** : cela implique une cessation complète des activités et une liquidation totale de l'entreprise.

Dans le dataset, toute entreprise ayant effectué une déclaration de faillite (sous le chapitre 7 ou 11) est étiquetée comme **Bankruptcy = 1** pour l'année précédant l'événement. Les autres entreprises sont considérées comme en activité normale (**Bankruptcy = 0**).

#### Caractéristiques principales du dataset :

- **Total des observations** : 78 682 combinaisons entreprise-année.
- **Variables financières** : 18 indicateurs comptables nommés **X1** à **X18**, mesurés annuellement.
- **Nettoyage** : le jeu de données est entièrement propre, sans valeurs manquantes, sans données synthétiques ni valeurs imputées.
- **Découpage temporel** :

Les données sont de 1999 jusqu'à 2018

Ce découpage temporel rigoureux vise à simuler des scénarios réels de prévision, permettant de tester les performances des modèles sur des périodes futures non observées au moment de l'entraînement. Il garantit ainsi une évaluation plus robuste et fidèle à l'utilisation pratique des modèles de prédiction dans un contexte réel.

## 2.2 Description des Données

Le jeu de données utilisé pour la prédiction de la faillite contient 18 variables financières principales, notées de **X1** à **X18**. Chaque variable représente un indicateur comptable clé permettant de capturer l'état financier d'une entreprise. Voici une description détaillée de ces variables :

Nom de Variable	Description
X1	<b>Actifs courants</b> – Ensemble des actifs d’une entreprise destinés à être utilisés ou vendus au cours de l’année suivante dans le cadre de ses opérations normales.
X2	<b>Coût des biens vendus</b> – Montant total payé par une entreprise directement lié à la vente de produits.
X3	<b>Amortissement et dépréciation</b> – Réduction de la valeur des actifs tangibles (dépréciation) ou intangibles (amortissement) sur la durée.
X4	<b>EBITDA</b> – Résultat avant intérêts, impôts, dépréciation et amortissement ; indicateur de performance globale d’une entreprise.
X5	<b>Stocks</b> – Évaluation comptable des matières premières, composants, et produits finis détenus pour la production ou la vente.
X6	<b>Résultat net</b> – Bénéfice net d’une entreprise après déduction de toutes les charges et impôts.
X7	<b>Créances totales</b> – Montant dû à l’entreprise pour des biens ou services fournis mais non encore payés.
X8	<b>Valeur de marché</b> – Capitalisation boursière d’une entreprise cotée, c’est-à-dire la valeur totale de ses actions.
X9	<b>Ventes nettes</b> – Total des ventes brutes moins les retours, remises et réductions.
X10	<b>Actifs totaux</b> – Ensemble des biens possédés par une entreprise ayant une valeur économique.
X11	<b>Dettes à long terme</b> – Ensemble des emprunts et passifs ayant une échéance supérieure à un an.
X12	<b>EBIT</b> – Résultat avant intérêts et impôts.
X13	<b>Marge brute</b> – Bénéfice restant après déduction des coûts de production des biens ou services vendus.
X14	<b>Passifs courants</b> – Ensemble des dettes exigibles à court terme (moins d’un an), incluant salaires dus, taxes, dettes fournisseurs, etc.
X15	<b>Résultats non distribués</b> – Bénéfices conservés par l’entreprise après paiement des dividendes et impôts.
X16	<b>Revenus totaux</b> – Somme de tous les revenus générés avant déduction des charges.
X17	<b>Passifs totaux</b> – Total des dettes et obligations financières d’une entreprise envers des tiers.
X18	<b>Dépenses opérationnelles</b> – Dépenses liées aux activités normales de l’entreprise.

TABLE 2.1 – Description des variables comptables du dataset

En résumé, les données fournissent une vue détaillée et complète de la situation financière des entreprises cotées aux États-Unis, ce qui permet d'effectuer une analyse approfondie du risque de faillite et de développer des modèles prédictifs efficaces basés sur le machine learning et le deep learning.

## 2.3 Prétraitement des Données

Le prétraitement des données est une étape essentielle dans le processus d'analyse des données, visant à assurer la qualité, la cohérence et la fiabilité des informations utilisées dans les modèles prédictifs. Dans le cadre de ce projet sur la prédiction de la faillite des entreprises américaines, les étapes suivantes ont été appliquées :

1. **Création de la variable cible** : La variable `bankrupt_next_year` a été générée en décalant la variable `status_label` d'une année. Une entreprise est considérée comme en faillite l'année suivante si elle a déposé le bilan selon le Chapitre 11 ou le Chapitre 7 du Code de la faillite américain.
2. **Tri des données temporelles** : Les données ont été triées chronologiquement par entreprise et par année afin de préserver la structure temporelle nécessaire à la création de séquences pour les modèles séquentiels.
3. **Création de nouvelles caractéristiques temporelles** :
  - *Variation en pourcentage* : Calculée pour chaque variable comptable afin d'observer les évolutions d'une année à l'autre.
  - *Croissance cumulée* : Utilisée pour capturer la tendance globale de croissance ou de déclin des entreprises.
  - *Écarts-types glissants* : Calculés avec une fenêtre glissante pour estimer la volatilité annuelle de chaque indicateur financier.
4. **Nettoyage des données** : Les valeurs infinies ont été remplacées par des NaN, et toutes les valeurs manquantes ont été remplies par des zéros après vérification. Les lignes comportant des valeurs non définies ont été supprimées si nécessaire.
5. **Normalisation** : Les variables financières dérivées ont été standardisées (z-score) à l'aide de la méthode `StandardScaler` de `sklearn`, afin d'assurer que toutes les caractéristiques soient sur une échelle comparable et d'améliorer la convergence des modèles de deep learning.
6. **Création des séquences** : Les observations ont été organisées sous forme de séquences temporelles de longueur 2 pour chaque entreprise. Chaque séquence représente l'évolution des caractéristiques financières sur deux années consécutives.
7. **Équilibrage du jeu de données** : Un sur-échantillonnage de la classe minoritaire (entreprises en faillite) a été réalisé à l'aide de la fonction `resample`, afin de corriger le déséquilibre de classes et améliorer la capacité des modèles à détecter les cas de faillite.
8. **Découpage en jeux d'entraînement et de test** : Le jeu de données équilibré a été divisé en un ensemble d'entraînement (80%) et un ensemble de test (20%) de façon stratifiée.



En appliquant ces étapes, nous avons veillé à ce que les données utilisées dans notre analyse soient de haute qualité, cohérentes dans le temps, et prêtes à être intégrées dans les modèles d'apprentissage automatique et de deep learning pour la prédiction de la faillite.

## 2.4 Analyse Exploratoire des Données

### 2.4.1 Taille du jeu de données

On va voir la taille de notre jeu de données

```
[1] import pandas as pd
    df = pd.read_csv('american_bankruptcy.csv')

[2] df.shape

(78682, 21)
```

FIGURE 2.1 – Taille du jeu de données

### 2.4.2 Aperçu du jeu de données

Aperçu des premières lignes du jeu de données brut contenant les états financiers annuels des entreprises américaines.

	company_name	status_label	year	X1	X2	X3	X4	X5	X6	X7	...	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18
0	C_1	alive	1999	511.267	833.107	18.373	89.031	336.018	35.163	128.348	...	1024.333	740.998	180.447	70.658	191.226	163.816	201.026	1024.333	401.483	935.302
1	C_1	alive	2000	485.856	713.811	18.577	64.367	320.590	18.531	115.187	...	874.255	701.854	179.987	45.790	160.444	125.392	204.065	874.255	361.642	809.888
2	C_1	alive	2001	436.656	526.477	22.496	27.207	286.588	-58.939	77.528	...	638.721	710.199	217.699	4.711	112.244	150.464	139.603	638.721	399.964	611.514
3	C_1	alive	2002	396.412	496.747	27.172	30.745	259.954	-12.410	66.322	...	606.337	686.621	164.658	3.573	109.590	203.575	124.106	606.337	391.633	575.592
4	C_1	alive	2003	432.204	523.302	26.680	47.491	247.245	3.504	104.661	...	651.958	709.292	248.666	20.811	128.656	131.261	131.884	651.958	407.608	604.467



5 rows x 21 columns

FIGURE 2.2 – Données sur les entreprises

Ce jeu de données constitue la base de notre étude. Il contient des informations comptables détaillées sur plus de 8 000 entreprises cotées entre 1999 et 2018. Avant toute modélisation, une étape de prétraitement est nécessaire pour construire des séquences temporelles cohérentes à partir de ces données.

### 2.4.3 Vérification des données nulles

Comme on peut le voir nous avons de la chance d'avoir une base de données qui n'a pas de valeurs nulles.

	0
company_name	0
status_label	0
year	0
X1	0
X2	0
X3	0
X4	0
X5	0
X6	0
X7	0
X8	0
X9	0
X10	0
X11	0
X12	0
X13	0
X14	0
X15	0
X16	0
X17	0
X18	0

dtype: int64

FIGURE 2.3 – Pas de valeurs nulles

### 2.4.4 Répartition du nombre d'années par entreprise

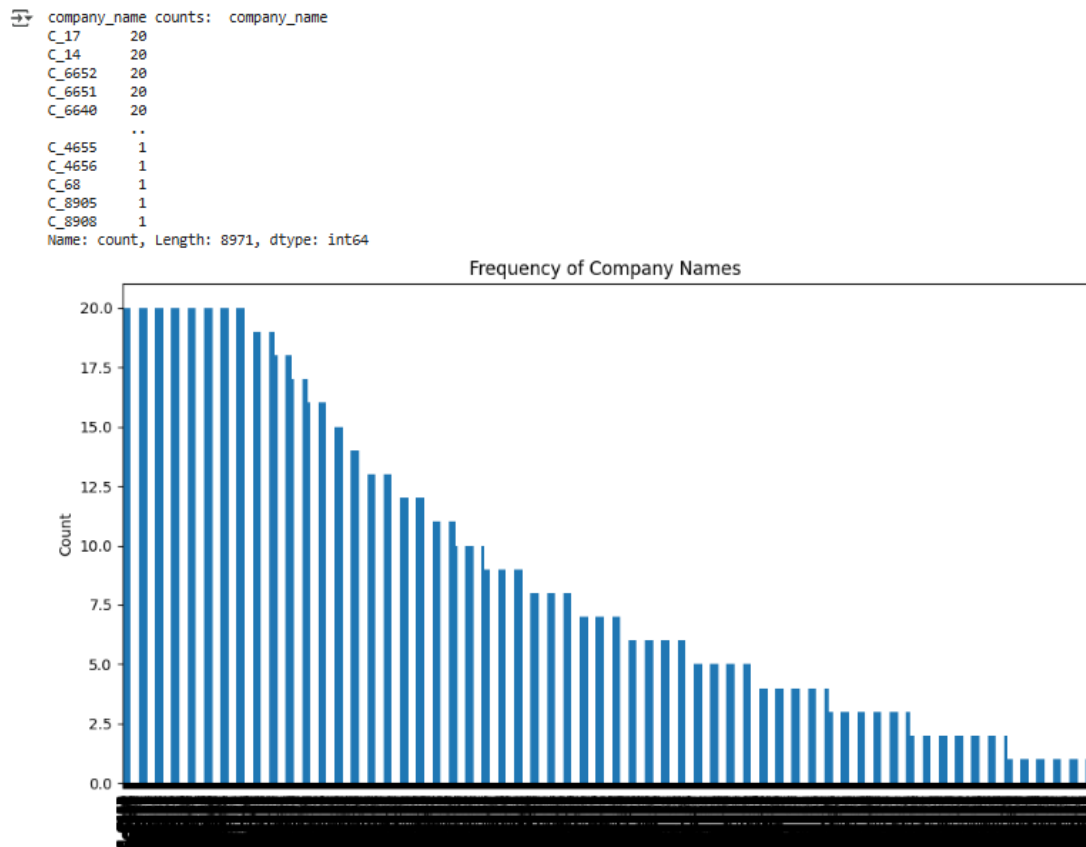


FIGURE 2.4 – Fréquence du nombre d'observations (années) par entreprise

Ce graphique montre que la majorité des entreprises ont un historique très court, souvent inférieur à 5 années. En revanche, un nombre limité d'entreprises disposent de 20 années consécutives d'historique financier. Cette hétérogénéité justifie le choix méthodologique suivant :

- **Pour les entreprises avec peu d'années** : des modèles classiques de Machine Learning sont appliqués, sans prise en compte de la dimension temporelle..
- **Pour les entreprises avec un historique suffisant ( $\geq 5$  années)** : des modèles de Deep Learning séquentiels (LSTM, BiLSTM, RNN...) seront utilisés, permettant d'exploiter la dynamique temporelle des données.

### 2.4.5 Répartition des entreprises par statut

Ce graphique met en évidence un déséquilibre important entre les deux classes : la majorité des entreprises sont actives, tandis qu'une faible proportion a connu une faillite. Ce déséquilibre justifie l'utilisation de techniques de rééchantillonnage, comme le suréchantillonnage de la classe minoritaire (SMOTE ou duplication), afin d'améliorer l'apprentissage des modèles de classification.

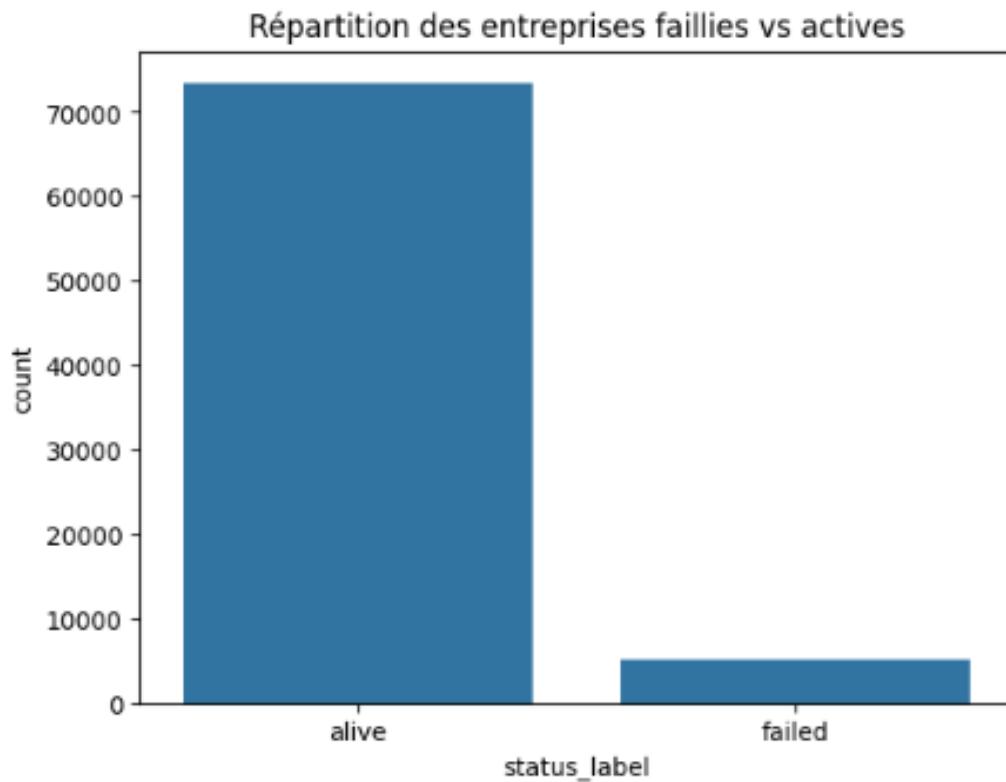


FIGURE 2.5 – Répartition des entreprises par leur statut financier

## 2.5 Conclusion

En résumé, les données fournissent une vue détaillée et complète de la situation financière des entreprises cotées aux États-Unis, ce qui permet d'effectuer une analyse approfondie du risque de faillite et de développer des modèles prédictifs efficaces.

# Chapitre 3

## Modélisation

### 3.1 Choix des Algorithmes

Dans le cadre de ce projet, notre objectif est de prédire la probabilité de faillite des entreprises américaines cotées en bourse. Pour cela, nous avons expérimenté une combinaison de modèles d'apprentissage automatique et de deep learning, en mettant l'accent sur la capacité des modèles séquentiels à capturer la dynamique temporelle des données financières.

Les modèles sélectionnés sont les suivants :

#### 1. Modèles Deep Learning :

- **LSTM (Long Short-Term Memory)** : adapté au traitement des séquences temporelles, ce modèle permet de capturer les dépendances à long terme dans les séries financières.
- **GRU (Gated Recurrent Unit)** : une alternative plus légère à LSTM, également efficace pour les données séquentielles.
- **CNN 1D (Convolutional Neural Network)** : utile pour capturer des motifs locaux dans les séries temporelles.
- **BiLSTM (Bidirectional LSTM)** : exploite l'information passée et future dans une séquence pour améliorer les performances prédictives.

#### 2. Modèles Classiques de Machine Learning (expérimentés pour comparaison dans une phase exploratoire du projet) :

- **Régression logistique** : modèle de base pour les problèmes de classification binaire.
- **Random Forest et XGBoost** : utilisés pour évaluer l'importance des variables et leur pouvoir prédictif hors séquence.
- **Perceptron Multi-Couches (MLP)** : testé pour sa capacité à modéliser des relations non linéaires dans des données tabulaires sans dépendance temporelle.

Ces modèles ont été entraînés sur des séquences temporelles de deux années consécutives, à partir de données financières normalisées et enrichies à l'aide d'indicateurs statistiques (croissance cumulée, variation annuelle, écart-type glissant).

### 3.1.1 Distinction entre les Modèles Classiques et les Modèles Deep Learning

Dans ce projet, deux approches complémentaires ont été adoptées pour la prédiction de la faillite des entreprises américaines :

- **Les modèles classiques de machine learning**, tels que la régression logistique, les forêts aléatoires ou XGBoost, ont été utilisés pour prédire si une entreprise allait faire faillite l'année suivante. Ces modèles exploitent des données financières *agrégées par année*, sans prendre en compte l'évolution temporelle d'une entreprise sur plusieurs années.
- **Les modèles de deep learning**, notamment les réseaux de neurones récurrents (LSTM, GRU, BiLSTM) et les réseaux convolutifs 1D (CNN), ont été utilisés pour prédire si une entreprise allait faire faillite *dans deux ans*, en se basant sur les données des **deux années précédentes**. Ces modèles permettent de capturer la dynamique temporelle de la santé financière d'une entreprise grâce à la modélisation de séquences.

Caractéristique	Modèles Classiques (ML)	Modèles Deep Learning (DL)
<b>Horizon de prédiction</b>	Prédiction de la faillite pour l'année suivante	Prédiction de la faillite pour l'année suivante
<b>Type de données utilisées</b>	États financiers annuels (séquence de longueur 1)	Séquences temporelles (longueur $\geq 2$ )
<b>Modèles utilisés</b>	Régression logistique, Random Forest, XGBoost, MLP	LSTM, GRU, BiLSTM, 1D CNN
<b>Avantages</b>	Simple, rapides, interprétables (ex. SHAP)	Capturent les dynamiques temporelles complexes
<b>Limites</b>	Ignorent la dépendance temporelle	Moins interprétables, nécessitent des séquences

TABLE 3.1 – Comparaison entre les approches classiques (ML) et séquentielles (DL) pour la prédiction de faillite

**Résumé comparatif :** Cette double approche permet de comparer la performance de méthodes classiques basées sur des caractéristiques agrégées avec des modèles plus complexes capables d'exploiter les dépendances temporelles dans les données.

## 3.2 Entraînement des modèles

### 3.2.1 Modèles du deep Learning

#### 3.2.1.1 Premier Entraînement avec le jeu de données déséquilibré sur des séquences de deux années

```
\n--- lstm ---
Classification Report:\n
              0          0.94          1.00          0.97          11612
              1          0.00          0.00          0.00           687

    accuracy              0.94          12299
   macro avg              0.47          0.50          0.49          12299
  weighted avg              0.89          0.94          0.92          12299

Confusion Matrix:\n [[11612    0]
 [ 687    0]]
ROC AUC Score: 0.6742376380203985
```

FIGURE 3.1 – LSTM

```
\n--- GRU ---
Classification Report:\n
              0          0.94          1.00          0.97          11612
              1          0.00          0.00          0.00           687

    accuracy              0.94          12299
   macro avg              0.47          0.50          0.49          12299
  weighted avg              0.89          0.94          0.92          12299

Confusion Matrix:\n [[11611    1]
 [ 687    0]]
ROC AUC Score: 0.6529797012677243
```

FIGURE 3.2 – GRU

```
\n--- 1D CNN ---
Classification Report:\n
              0          0.94          1.00          0.97          11612
              1          0.10          0.01          0.01           687

    accuracy              0.94          12299
   macro avg              0.52          0.50          0.49          12299
  weighted avg              0.90          0.94          0.92          12299

Confusion Matrix:\n [[11575   37]
 [ 683    4]]
ROC AUC Score: 0.5491219368008099
```

FIGURE 3.3 – CNN 1D

```

\n--- BiLSTM ---
Classification Report:\n
              0      0.94      1.00      0.97      11612
              1      0.00      0.00      0.00       687

    accuracy              0.94      12299
    macro avg              0.47      0.50      0.49      12299
    weighted avg           0.89      0.94      0.92      12299

Confusion Matrix:\n [[11612    0]
 [ 687    0]]
ROC AUC Score: 0.6800267103097183

```

FIGURE 3.4 – BiLSTM

Même après l’optimisation des modèles et l’augmentation du nombre d’**epochs** de 9 à 50, les performances sont restées pauvre pour la classe minoritaire 1 qui représente la faillite.

```

\n--- LSTM Optimisé ---
Classification Report:\n
              0      0.94      1.00      0.97      11612
              1      0.00      0.00      0.00       687

    accuracy              0.94      12299
    macro avg              0.47      0.50      0.49      12299
    weighted avg           0.89      0.94      0.92      12299

Confusion Matrix:\n [[11612    0]
 [ 687    0]]
ROC AUC Score: 0.7078653262874676

```

FIGURE 3.5 – LSTM optimisé

```

\n--- GRU Optimisé ---
Classification Report:\n
              0      0.94      1.00      0.97      11612
              1      0.00      0.00      0.00       687

    accuracy              0.94      12299
    macro avg              0.47      0.50      0.49      12299
    weighted avg           0.89      0.94      0.92      12299

Confusion Matrix:\n [[11612    0]
 [ 687    0]]
ROC AUC Score: 0.6953183876941035

```

FIGURE 3.6 – GRU optimisé



```

\n--- 1D CNN Optimisé ---
Classification Report:\n
              0          0.94          1.00          0.97          11612
              1          0.00          0.00          0.00           687

    accuracy          0.94          12299
   macro avg          0.47          0.50          0.49          12299
weighted avg          0.89          0.94          0.92          12299

Confusion Matrix:\n [[11612    0]
 [   687    0]]
ROC AUC Score: 0.5006078388015008

```

FIGURE 3.7 – CNN 1D optimisé

```

\n--- BiLSTM Optimisé ---
Classification Report:\n
              0          0.94          1.00          0.97          11612
              1          0.00          0.00          0.00           687

    accuracy          0.94          12299
   macro avg          0.47          0.50          0.49          12299
weighted avg          0.89          0.94          0.92          12299

Confusion Matrix:\n [[11612    0]
 [   687    0]]
ROC AUC Score: 0.6882290242338273

```

FIGURE 3.8 – BiLSTM optimisé

Comme on peut le voir la performance des modèles est très pauvre. Maintenant on va équilibrer le jeu de données et on va essayer d'entraîner les modèles à nouveau.

### 3.2.2 Entraînement des modèles avec un jeu de données équilibré sur des séquences de deux années

Les résultats sont très contents puisque maintenant la performance des modèles est équilibrée et n'est pas biaisée vers la classe majoritaire 0. Pour le choix du modèle on va faire une optimisation de ces modèles et on va choisir le modèle le plus performant sur les séquences de deux années.

```

\n--- LSTM ---
Classification Report:\n
              0          0.69      0.67      0.68      11612
              1          0.68      0.69      0.68      11612

    accuracy              0.68      23224
   macro avg              0.68      23224
  weighted avg              0.68      23224

Confusion Matrix:\n [[7752 3860]
 [3554 8058]]
ROC AUC Score: 0.7406889902341277

```

FIGURE 3.9 – LSTM sur un jeu de données équilibré

```

\n--- GRU ---
Classification Report:\n
              0          0.67      0.66      0.66      11612
              1          0.66      0.67      0.67      11612

    accuracy              0.66      23224
   macro avg              0.66      23224
  weighted avg              0.66      23224

Confusion Matrix:\n [[7633 3979]
 [3804 7808]]
ROC AUC Score: 0.7240418029135645

```

FIGURE 3.10 – GRU sur un jeu de données équilibré

```

\n--- 1D CNN ---
Classification Report:\n
              0          0.67      0.35      0.46      11612
              1          0.56      0.83      0.67      11612

    accuracy              0.59      23224
   macro avg              0.61      23224
  weighted avg              0.61      23224

Confusion Matrix:\n [[4015 7597]
 [1989 9623]]
ROC AUC Score: 0.6538812485249025

```

FIGURE 3.11 – CNN 1 sur un jeu de données équilibré

```

\n--- BiLSTM ---
Classification Report:\n
              0          0.70      0.66      0.68      11612
              1          0.68      0.72      0.70      11612

    accuracy              0.69      23224
  macro avg              0.69      0.69      0.69      23224
weighted avg              0.69      0.69      0.69      23224

Confusion Matrix:\n [[7653 3959]
 [3254 8358]]
ROC AUC Score: 0.7563443469101833

```

FIGURE 3.12 – BiLSTM sur un jeu de données équilibré

Chaque modèle de base a été ensuite optimisé. Les versions optimisées se distinguent par l'ajout de plusieurs éléments : d'abord, l'utilisation de couches empilées (**stacked layers**) avec le paramètre `return sequences=True` pour permettre aux couches **LSTM** ou **GRU** de transmettre toute la séquence à la couche suivante. Ensuite, nous avons introduit des couches de **Dropout** pour réduire le surapprentissage, et des couches Dense supplémentaires pour enrichir la représentation non linéaire. Enfin, nous avons intégré des **callbacks** comme l'arrêt anticipé (**EarlyStopping**) et la réduction du taux d'apprentissage (**ReduceLROnPlateau**) afin d'assurer une convergence plus stable et efficace du modèle. Ces optimisations visent à renforcer la robustesse et la généralisation des modèles sur des séquences complexes.

```

\n--- LSTM Optimisé ---
Classification Report:\n
              0          0.83      0.76      0.79      11612
              1          0.78      0.84      0.81      11612

    accuracy              0.80      23224
  macro avg              0.80      0.80      0.80      23224
weighted avg              0.80      0.80      0.80      23224

Confusion Matrix:\n [[8834 2778]
 [1866 9746]]
ROC AUC Score: 0.8796751654334091

```

FIGURE 3.13 – LSTM

```

\n--- GRU Optimisé ---
Classification Report:\n
              0          0.77      0.73      0.75      11612
              1          0.75      0.78      0.76      11612

    accuracy              0.76      23224
    macro avg          0.76      0.76      0.76      23224
    weighted avg       0.76      0.76      0.76      23224

Confusion Matrix:\n [[8529 3083]
 [2549 9063]]
ROC AUC Score: 0.8344207981065118

```

FIGURE 3.14 – GRU

```

\n--- 1D CNN Optimisé ---
Classification Report:\n
              0          0.60      0.00      0.00      11612
              1          0.50      1.00      0.67      11612

    accuracy              0.50      23224
    macro avg          0.55      0.50      0.34      23224
    weighted avg       0.55      0.50      0.34      23224

Confusion Matrix:\n [[  28 11584]
 [  19 11593]]
ROC AUC Score: 0.5067016594305557
Epoch 1/50

```

FIGURE 3.15 – CNN 1D

```

\n--- BiLSTM Optimisé ---
Classification Report:\n
              0          0.86      0.79      0.82      11612
              1          0.80      0.87      0.84      11612

    accuracy              0.83      23224
    macro avg          0.83      0.83      0.83      23224
    weighted avg       0.83      0.83      0.83      23224

Confusion Matrix:\n [[ 9117 2495]
 [1474 10138]]
ROC AUC Score: 0.9120840254697498

```

FIGURE 3.16 – BiLSTM

Après optimisation, tous les modèles ont vu leurs performances s'améliorer, sauf le **1D CNN** dont la performance a chuté. Cette baisse peut s'expliquer par une **sur-régularisation** due aux Dropout, une complexité inutile avec l'ajout de couches sup-

plémentaires, et une perte d'information causée par la **GlobalMaxPooling1D**. Cela montre que les techniques d'optimisation doivent être adaptées à chaque type de modèle.

### 3.2.2.1 Choix du modèle pour les séquences de deux années

Pour comparer les performances des modèles, nous avons analysé trois éléments clés : le **rapport de classification**, la **matrice de confusion**, et le **ROC AUC Score**. Comme illustré dans les figures ci-dessous, le modèle BiLSTM Optimisé surpasse le LSTM Optimisé sur tous les plans. Il affiche une précision moyenne de **83 %**, un **F1-score** plus équilibré (**0.82** pour la **classe 0** et **0.84** pour la **classe 1**), et une **accuracy** globale de **83 %**, contre **80 %** pour le LSTM. Sa matrice de confusion montre également moins de faux positifs et de faux négatifs. Enfin, son **ROC AUC Score** de **0.91** témoigne d'une meilleure capacité à distinguer les classes comparé au LSTM, dont le score est de **0.87**. Ces résultats confirment que le BiLSTM Optimisé est le modèle le plus performant et le mieux adapté à notre problématique.

### 3.2.3 L'effet de la longueur des séquences

Dans ce qui suit, on va entraîner nos modèles sur des séquences de 5 ans c'est pour voir l'effet de la longueur des séquences. On peut faire cela pour tous les autres longueur comprises entre 2 et 5. Mais, on va se concentrer sur la longueur 5.

```
\n--- LSTM ---
Classification Report:\n
              0          0.75          0.71          0.73          7965
              1          0.73          0.76          0.74          7964

    accuracy                    0.74          15929
   macro avg                    0.74          0.74          0.74          15929
weighted avg                    0.74          0.74          0.74          15929

Confusion Matrix:\n [[5676 2289]
 [1925 6039]]
ROC AUC Score: 0.8007089892589472
```

FIGURE 3.17 – LSTM

```

\n--- GRU ---
Classification Report:\n
              0          0.73      0.61      0.67      7965
              1          0.67      0.77      0.72      7964

    accuracy              0.69      15929
   macro avg              0.70      0.69      0.69      15929
  weighted avg              0.70      0.69      0.69      15929

Confusion Matrix:\n [[4882 3083]
 [1805 6159]]
ROC AUC Score: 0.7639258332300752

```

FIGURE 3.18 – GRU

```

\n--- 1D CNN ---
Classification Report:\n
              0          0.64      0.65      0.65      7965
              1          0.65      0.64      0.64      7964

    accuracy              0.65      15929
   macro avg              0.65      0.65      0.65      15929
  weighted avg              0.65      0.65      0.65      15929

Confusion Matrix:\n [[5208 2757]
 [2885 5079]]
ROC AUC Score: 0.7213475548947035

```

FIGURE 3.19 – CNN 1D

```

\n--- BiLSTM ---
Classification Report:\n
              0          0.76      0.76      0.76      7965
              1          0.76      0.76      0.76      7964

    accuracy              0.76      15929
   macro avg              0.76      15929
  weighted avg              0.76      15929

Confusion Matrix:\n [[6051 1914]
 [1889 6075]]
ROC AUC Score: 0.8387985813751335

```

FIGURE 3.20 – BiLSTM

Après, nous avons empilé deux couches récurrentes avec `return_sequences=True` pour permettre au réseau de mieux capturer les dépendances temporelles complexes. Ensuite, l'ajout de Dropout entre les couches permet de réduire le surapprentissage. Nous avons également introduit une couche Dense supplémentaire dans le CNN pour enrichir l'expressivité du modèle. Enfin, l'entraînement a été étendu à **50 epochs** avec les callbacks **EarlyStopping** et **ReduceLROnPlateau** pour un ajustement plus fin du modèle tout en évitant le surapprentissage. Et comme, on peut le voir la performance de tous les modèles a augmenté sauf **CNN 1D** une autre fois.

```

\n--- LSTM Optimisé ---
Classification Report:\n
              0          0.92      0.87      0.89      7965
              1          0.88      0.93      0.90      7964

    accuracy              0.90      15929
   macro avg              0.90      15929
  weighted avg              0.90      15929

Confusion Matrix:\n [[6914 1051]
 [ 575 7389]]
ROC AUC Score: 0.9624067090356069

```

FIGURE 3.21 – LSTM

```

\n--- GRU Optimisé ---
Classification Report:\n
              0          0.87      0.89      0.88      7965
              1          0.89      0.87      0.88      7964

    accuracy              0.88      15929
    macro avg            0.88      0.88      0.88      15929
    weighted avg         0.88      0.88      0.88      15929

Confusion Matrix:\n [[7119  846]
 [1049 6915]]
ROC AUC Score: 0.9588591379348941

```

FIGURE 3.22 – GRU

```

\n--- 1D CNN Optimisé ---
Classification Report:\n
              0          0.45      0.04      0.08      7965
              1          0.50      0.95      0.65      7964

    accuracy              0.50      15929
    macro avg            0.48      0.50      0.37      15929
    weighted avg         0.48      0.50      0.37      15929

Confusion Matrix:\n [[ 336 7629]
 [ 404 7560]]
ROC AUC Score: 0.4934667160413953

```

FIGURE 3.23 – CNN 1D

```

\n--- BiLSTM Optimisé ---
Classification Report:\n
              0          0.97      0.93      0.95      7965
              1          0.93      0.97      0.95      7964

    accuracy              0.95      15929
    macro avg            0.95      0.95      0.95      15929
    weighted avg         0.95      0.95      0.95      15929

Confusion Matrix:\n [[7409  556]
 [ 227 7737]]
ROC AUC Score: 0.9880136855649544

```

FIGURE 3.24 – BiLSTM

### 3.2.4 Choix du modèle pour les séquence de 5ans

Parmi tous les modèles testés, le **BiLSTM optimisé** a obtenu les **meilleurs résultats globaux**, avec une **accuracy de 95 %**, un **f1-score de 0,95** pour les deux classes, et un **ROC AUC Score exceptionnel de 0,988**, indiquant une excellente capacité



à distinguer les classes. Contrairement aux autres architectures, ce modèle combine la puissance des LSTM classiques avec une lecture *dans les deux directions de la séquence* (passé et futur), ce qui lui permet de **capturer plus efficacement les dépendances temporelles complexes**. En comparaison :

- Le **LSTM optimisé** atteint une accuracy de **90 %** avec un f1-score de **0,90**.
- Le **GRU optimisé** atteint **88 %** de précision avec un f1-score de **0,88**.
- Le **1D CNN optimisé** chute à une accuracy de **50 %**, montrant une incapacité à bien généraliser.

De plus, le BiLSTM a bien profité de l'*empilement de couches*, du *Dropout*, et de l'utilisation de *ReduceLROnPlateau*, ce qui a permis un **apprentissage stable et progressif**, comme le montrent les courbes de perte et de précision. Ces éléments justifient pleinement le choix du BiLSTM optimisé comme **meilleur modèle pour cette tâche de classification séquentielle**.

### 3.2.4.1 Optimisation du BiLSTM

Dans la dernière version optimisée du modèle **BiLSTM**, plusieurs améliorations ont été introduites pour renforcer la robustesse de l'apprentissage. Tout d'abord, nous avons ajouté une couche Dense intermédiaire avec **32 neurones** et une activation **ReLU** pour augmenter la capacité d'apprentissage non linéaire du modèle. Ensuite, une régularisation plus poussée a été appliquée via trois couches de Dropout successives, ce qui permet de réduire davantage le surapprentissage. De plus, nous avons pris en compte le déséquilibre entre les classes en utilisant la fonction `compute class weight` de Scikit learn afin de calculer des poids de classes équilibrés, passés à l'entraînement via le paramètre `class weight`. Enfin, le nombre d'épochs a été étendu à **70**, tout en maintenant les callbacks **EarlyStopping** et **ReduceLROnPlateau**, ce qui a permis une meilleure convergence sans surapprentissage. Ces ajustements ont significativement renforcé la performance du BiLSTM, qui s'est avéré être le modèle le plus performant de l'étude.

```
\n--- BiLSTM Optimisé ---
Classification Report:\n
              0          0.99      0.95      0.97      7965
              1          0.95      0.99      0.97      7964

    accuracy              0.97      0.97      0.97      15929
    macro avg              0.97      0.97      0.97      15929
    weighted avg           0.97      0.97      0.97      15929

Confusion Matrix:\n [[7578  387]
 [  72 7892]]
ROC AUC Score: 0.9937703185994223
```

FIGURE 3.25 – BiLSTM

Le modèle **BiLSTM optimisé**, entraîné sur **70 épochs** avec pondération des classes et régularisation par Dropout, a démontré des performances exceptionnelles. Il a atteint une **accuracy de 97 %** avec un **f1-score de 0,97** pour les deux classes. La **matrice de confusion** montre un taux d'erreur très faible, avec seulement **72 faux négatifs** et

**387 faux positifs** sur un total de **15 929 échantillons**. Le **ROC AUC Score** s'élève à **0,9937**, confirmant une capacité de discrimination quasi parfaite.

Comme illustré dans la figure ci dessous, les courbes d'apprentissage montrent une **convergence nette** : la *val loss* continue de baisser jusqu'à la fin de l'entraînement, tandis que la *val accuracy* atteint plus de **95 %**. Ce comportement indique un modèle bien entraîné sans surapprentissage, validant le choix du **BiLSTM optimisé** comme **meilleur modèle** pour cette tâche de classification séquentielle.

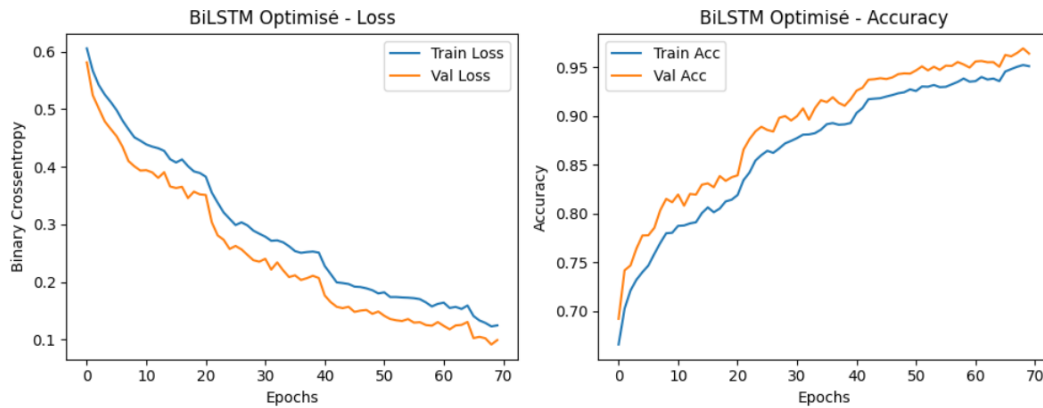


FIGURE 3.26 – Courbes de loss et accuracy

### 3.2.5 Modèles du machine learning

### 3.2.6 Entraînement sur le jeu de données déséquilibré

```

--- Random Forest ---
Classification Report:
              precision    recall  f1-score   support

         0       0.94      1.00      0.97    14815
         1       0.95      0.04      0.08      922

    accuracy      0.94    15737
   macro avg      0.95      0.52      0.53    15737
  weighted avg      0.94      0.94      0.92    15737

Confusion Matrix:
[[14813    2]
 [ 882   40]]
ROC AUC Score: 0.8686226292019507

--- Logistic Regression ---
Classification Report:
              precision    recall  f1-score   support

         0       0.97      0.31      0.47    14815
         1       0.07      0.86      0.13      922

    accuracy      0.34    15737
   macro avg      0.52      0.58      0.30    15737
  weighted avg      0.92      0.34      0.45    15737

Confusion Matrix:
[[ 4605 10210]
 [  131   791]]
ROC AUC Score: 0.6550634982572481

```

FIGURE 3.27 – Random Forest Et Logistic Regression

```

--- MLP ---
/usr/local/lib/python3.11/dist-packages/sklearn/neural_network/_mlp.py:158: UserWarning:
  warnings.warn(
Classification Report:
              precision    recall  f1-score   support

         0       0.95        1.00        0.97       14815
         1       0.64        0.07        0.13         922

 accuracy          0.94       15737
 macro avg          0.79       15737
weighted avg          0.93       15737

Confusion Matrix:
[[14779   36]
 [ 858   64]]
ROC AUC Score: 0.7478185400122845

--- XGBoost ---
/usr/local/lib/python3.11/dist-packages/xgboost/core.py:158: UserWarning:
  warnings.warn(smsg, UserWarning)
Parameters: { "use_label_encoder" } are not used.

Classification Report:
              precision    recall  f1-score   support

         0       0.95        1.00        0.97       14815
         1       0.65        0.08        0.15         922

 accuracy          0.94       15737
 macro avg          0.80       15737
weighted avg          0.93       15737

Confusion Matrix:
[[14774   41]
 [ 846   76]]
ROC AUC Score: 0.8243901099826273

```

FIGURE 3.28 – MLP et xgboost

En résumé la performance des modèles est médiocre en particulier pour la **classe minoritaire 1**. On va essayer d'équilibrer le jeu de données.

### 3.2.7 Entraînement sur le jeu de données équilibré

```

--- Random Forest ---
Classification Report:
              precision    recall  f1-score   support

         0       0.96      0.96      0.96     14815
         1       0.42      0.43      0.43        922

    accuracy      0.93      0.93      0.93     15737
   macro avg      0.69      0.70      0.69     15737
weighted avg      0.93      0.93      0.93     15737

Confusion Matrix:
[[14259  556]
 [ 522  400]]
ROC AUC Score: 0.8676942229653799

--- Logistic Regression ---
Classification Report:
              precision    recall  f1-score   support

         0       0.97      0.32      0.48     14815
         1       0.07      0.84      0.13        922

    accuracy      0.35      0.35      0.35     15737
   macro avg      0.52      0.58      0.31     15737
weighted avg      0.92      0.35      0.46     15737

Confusion Matrix:
[[ 4705 10110]
 [  144   778]]
ROC AUC Score: 0.6560633203581702

```

FIGURE 3.29 – Random Forest et Logistic Regression

```

--- MLP ---
/usr/local/lib/python3.11/dist-packages/sklearn/neural_network/_mult
warnings.warn(
Classification Report:
              precision    recall  f1-score   support

         0       0.97      0.75      0.84     14815
         1       0.13      0.61      0.21       922

    accuracy: 0.74
 macro avg:  0.55
weighted avg: 0.92      0.74      0.81     15737

Confusion Matrix:
[[11072  3743]
 [   362   560]]
ROC AUC Score: 0.7438035847762315

--- XGBoost ---
/usr/local/lib/python3.11/dist-packages/xgboost/core.py:158: UserWar
Parameters: { "use_label_encoder" } are not used.

warnings.warn(msg, UserWarning)
Classification Report:
              precision    recall  f1-score   support

         0       0.97      0.83      0.89     14815
         1       0.18      0.61      0.28       922

    accuracy: 0.81
 macro avg:  0.57
weighted avg: 0.92      0.81      0.86     15737

Confusion Matrix:
[[12235  2580]
 [   363   559]]
ROC AUC Score: 0.8008298296488214

```

FIGURE 3.30 – MLP et xgboost

Après optimisation, les performances des modèles ont nettement évolué, notamment pour **XGBoost** et **MLP (Multi-Layer Perceptron)**.

Le modèle **XGBoost** a montré une **progression significative**, avec un *ROC AUC Score* passant de **0,80 à 0,82**, et une légère amélioration du *f1-score* pour la **classe minoritaire (1)** et une chute dans la précision pour cette classe.

Le **MLP**, malgré une *accuracy* élevée, a continué à présenter une **très faible capacité à détecter la classe 1**, avec des *rappels autour de 7%*, même si le *f1-score* s'est légèrement amélioré par rapport à la version non optimisée.

D'autres modèles comme la **régression logistique** ou le **Random Forest** ont montré de bonnes performances sur la classe majoritaire, mais ont échoué à équilibrer correctement la **détection de la classe minoritaire**, ce qui les rend moins adaptés dans le contexte d'un problème de classification déséquilibrée.

### 3.2.8 Choix du modèle le plus équilibré : Random Forest

Parmi tous les modèles de machine learning classiques testés, le **Random Forest** s'est démarqué comme le plus équilibré. Il atteint une **accuracy globale de 93 %**, mais surtout un **f1-score de 0,43** sur la **classe minoritaire**, soit le meilleur score obtenu parmi les modèles non profonds. Contrairement à d'autres modèles comme la régression logistique ou le MLP, qui présentent soit un biais fort en faveur de la classe majoritaire soit un manque de précision, le Random Forest parvient à préserver un **bon compromis entre précision et rappel pour les deux classes**. Son *ROC AUC Score* de **0,867** renforce sa fiabilité en tant que classifieur robuste pour des données déséquilibrées.

## 3.3 Résultats

Les performances des modèles ont été évaluées à l'aide de plusieurs métriques : accuracy, AUC-ROC, précision, rappel, f1-score et matrice de confusion. Deux approches distinctes ont été mises en œuvre selon la disponibilité des données historiques :

- Les **modèles de machine learning classiques** (comme *XGBoost*, *MLP*, *Random Forest*, et *régression logistique*) ont été utilisés pour les entreprises ne disposant pas d'historique temporel exploitable (séquences de longueur 1), c'est-à-dire uniquement à partir de leurs états financiers statiques.
- Les **modèles de deep learning** ont été appliqués uniquement aux entreprises disposant d'un historique de longueur **supérieure ou égale à 2**, afin d'exploiter les séquences temporelles via des réseaux récurrents.

Parmi les modèles classiques, **XGBoost optimisé** a donné les meilleurs résultats avec un *ROC AUC* de 0,82, tandis que les autres modèles ont souffert soit d'un déséquilibre, soit d'un rappel très faible sur les entreprises en faillite.

Du côté des modèles séquentiels, le **BiLSTM optimisé** s'est largement distingué avec une *accuracy* de 97 %, un *f1-score* équilibré de 0,97 pour les deux classes, et un *ROC AUC* de 0,99.

## 3.4 Analyse des Résultats

L'analyse des performances nous a permis de dégager plusieurs constats :

- Les modèles de deep learning, en particulier ceux qui exploitent la structure séquentielle bidirectionnelle comme le **BiLSTM**, surpassent nettement les approches classiques pour les entreprises disposant de données temporelles.
- Pour les entreprises avec peu de données (pas de séquences), les modèles de machine learning comme **XGBoost** restent les plus adaptés, notamment grâce à leur robustesse et leur capacité d'interprétation via SHAP.
- L'utilisation de l'équilibrage des classes et des stratégies de régularisation (Dropout, early stopping, réduction du learning rate) a permis d'améliorer la généralisation sans surapprentissage.

- Le modèle BiLSTM, enrichi par une couche dense explicative et entraîné avec pondération de classes, a atteint les meilleures performances globales sur les séquences historiques financières.

En conclusion, **les modèles de deep learning sont les plus pertinents dès que des séquences sont disponibles**, tandis que les **modèles de machine learning classiques restent efficaces pour les cas sans historique** et peuvent également être utilisés à des fins explicatives.

## 3.5 Élaboration des Stratégies

### 3.5.1 Stratégies de Développement

Sur la base des résultats obtenus à travers les différents modèles, plusieurs axes de développement peuvent être envisagés afin d'optimiser l'utilisation des outils de prédiction de faillite dans le monde réel.

Il convient de noter que les **modèles de machine learning classiques** (comme XGBoost) ont été utilisés pour les entreprises ne disposant pas de séquences temporelles exploitables, tandis que les **modèles de deep learning** (comme BiLSTM) ont été réservés aux entreprises avec un historique d'au moins deux périodes (années).

#### Renforcement de la Surveillance Financière

- Intégrer les modèles prédictifs au sein des systèmes de gestion des risques des institutions financières pour détecter de manière précoce les entreprises à risque.
- Utiliser les prévisions pour orienter les décisions d'octroi de crédit ou d'investissement.

#### Déploiement dans des Environnements Réels

- Développer une interface décisionnelle (ex : tableau de bord interactif) permettant aux analystes de visualiser les résultats des prédictions.
- Adapter les modèles à d'autres contextes économiques, par exemple dans d'autres bourses ou marchés internationaux.

#### Extension des Modèles

- Enrichir les données utilisées avec des indicateurs macroéconomiques, de sentiment de marché ou de gouvernance pour améliorer la robustesse des modèles.
- Étendre les séquences temporelles utilisées par les modèles deep learning pour capter des dynamiques à plus long terme.

#### Renforcement de la Prévention des Risques

- Collaborer avec les agences de régulation pour mettre en œuvre des systèmes de signalement précoce basés sur les prédictions.



- Former les gestionnaires de portefeuilles et les analystes aux techniques d'interprétation des résultats fournis par les modèles (ex : SHAP, LIME).

### 3.5.2 Actions Recommandées

À partir des prédictions fournies par nos modèles et des performances observées, plusieurs actions concrètes sont recommandées :

- **Adoption des modèles LSTM pour la prédiction à moyen terme** (2 ans), compte tenu de leur capacité à capturer l'évolution temporelle des entreprises.
- **Utilisation des modèles classiques pour des analyses rapides** sur la faillite à court terme (année suivante), avec des modèles comme XGBoost ou la régression logistique.
- **Mise en place d'un pipeline automatisé** pour la surveillance régulière des entreprises cotées, incluant la mise à jour des données et le déclenchement de prédictions périodiques.
- **Priorisation des interventions** auprès des entreprises identifiées comme à haut risque, en combinant les scores issus des modèles avec des critères de gravité sectorielle.
- **Documentation et sauvegarde des modèles** pour permettre leur réutilisation, auditabilité, et traçabilité dans un cadre réglementaire ou industriel.

En mettant en œuvre ces stratégies de manière cohérente, il est possible de transformer les résultats prédictifs en outils d'aide à la décision performants et opérationnels dans le cadre de la prévention des faillites d'entreprises.

# Conclusion générale

En conclusion, ce projet de prédiction de la faillite des entreprises américaines cotées en bourse représente une contribution significative à l'analyse financière prévisionnelle. À travers les différents chapitres, nous avons exploré de manière rigoureuse les dimensions essentielles à la construction d'un système de détection anticipée du risque de faillite, allant de l'analyse contextuelle à l'implémentation de stratégies concrètes.

Dans le premier chapitre, nous avons exposé le contexte économique et financier qui motive la nécessité d'outils prédictifs dans le domaine de la finance d'entreprise. La compréhension des enjeux liés à la détection de la faillite, notamment en termes de gestion des risques, de stabilité économique, et de prise de décision stratégique, a permis de justifier les objectifs principaux de ce travail.

Le second chapitre a été dédié à la collecte et à la préparation des données. Grâce à un jeu de données réel, riche et structuré, extrait de sources publiques comme le **New York Stock Exchange** et le **NASDAQ**, nous avons pu extraire des indicateurs financiers pertinents et effectuer un prétraitement rigoureux des données. Cette étape a été cruciale pour garantir la qualité des modèles en aval.

Dans le troisième chapitre, nous avons développé et comparé plusieurs approches de modélisation : des modèles classiques de machine learning pour la prédiction à court terme (faillite l'année suivante), et des modèles de deep learning (notamment LSTM) pour la prédiction à moyen terme (deux ans). L'analyse des résultats a montré que les modèles LSTM offrent une capacité supérieure à capturer les dynamiques temporelles complexes des entreprises, tandis que les modèles classiques permettent une interprétation rapide et efficace.

Enfin, nous avons proposé une série de stratégies d'intégration des résultats dans des systèmes décisionnels, en soulignant les perspectives d'application en finance, audit, et conseil en gestion des risques. Ces recommandations visent à rendre les prédictions réellement exploitables dans des environnements professionnels.

En somme, ce projet a permis de démontrer la valeur ajoutée de l'intelligence artificielle pour anticiper les défaillances financières, et ouvre la voie à des solutions innovantes pour la prévention des risques dans le secteur économique et financier.

# Webographie

- [1] *Démographie : le spectre du vieillissement guette le Maroc* : [https://fr.le360.ma/economie/demographie-le-spectre-du-vieillissement-guette-lemaroc\\_N3GIZIZUZJC55MC375LTBQZ7DA/](https://fr.le360.ma/economie/demographie-le-spectre-du-vieillissement-guette-lemaroc_N3GIZIZUZJC55MC375LTBQZ7DA/)
- [2] *E-Santé Appel à projets innovants* : <https://ensias.um5.ac.ma/sites/ensias.um5.ac.ma/files/E-Sante%CC%81%20appel%20a%CC%80%20projets%20Version%20finale.pdf>
- [3] *E-Santé au Maroc : État des lieux et perspectives d'avenir* : <https://pharmacie.ma/uploads/pdfs/esante-best.pdf>
- [4] *Analyse du secteur de la santé au Maroc par Phénicia Conseil* : <https://recrutement-phenicia.fr/analyse-du-secteur-de-la-sante-au-maroc-par-phenicia-conseil/>
- [5] *Plan Santé 2025* : [https://extranet.who.int/countryplanningcycles/sites/default/files/public\\_file\\_rep/MAR\\_Morocco\\_Plan-de-sant%C3%A9-2025.pdf](https://extranet.who.int/countryplanningcycles/sites/default/files/public_file_rep/MAR_Morocco_Plan-de-sant%C3%A9-2025.pdf)
- [6] *Analyse PESTEL : définition, utilité et présentation des 6 composants* : <https://www.lecoindesentrepreneurs.fr/analyse-pestel-definition-outil-et-composants/>
- [7] *Le SWOT : l'outil d'analyse stratégique pour développer votre activité* : <https://bpifrance-creation.fr/encyclopedie/letude-marche/determiner-sa-strategie/swot-loutil-danalyse-strategique-developper>
- [8] *Les 5 forces de Porter : concept et mise en oeuvre* : <https://www.manager-go.com/strategie-entreprise/les-5-forces-de-porter.htm>
- [9] *Matrice du Boston Consulting Group : définition et exemple* : <https://www.manager-go.com/strategie-entreprise/matrice-bcg.htm>
- [10] *Comprendre le marketing mix ou 4P : Produit, Prix, Place, Promotion* : <https://www.manager-go.com/marketing/marketing-mix-4p.htm#:~:text=0n%20parle%20de%204P%2C%20un,d%27atteindre%20les%20objectifs%20fix%C3%A9s>
- [11] *Les 5 phases du cycle de vie produit* : <https://fr.surveymonkey.com/market-research/resources/5-stages-of-product-life-cycle/>
- [12] *TARIFS* : [http://www.directompic.ma/directinfo/documents/Tarifs\\_liste\\_2.pdf](http://www.directompic.ma/directinfo/documents/Tarifs_liste_2.pdf)
- [13] *Formalités de création d'entreprise* : <http://www.guelmiminvest.ma/formalites-de-creation-d-entreprise.php#:~:text=Enregistrement%20du%20capital%20et%20des%20statuts&text=toutes%20les%20soci%C3%A9t%C3%A9s-,Frais%20%3A,2%20Dh%20par%20signature%201%C3%A9galis%C3%A9>

- [14] *Création d'Entreprises au Maroc – Formalités Standard de la Création d'Entreprises* : <https://amde.ma/creation-dentreprise-creation-entreprises-maroc-formalites-standard-creation-entre>

Nous exprimons notre gratitude envers les auteurs et les chercheurs dont les travaux ont été une source précieuse d'information pour notre projet.