INSTITUT FRANCOPHONE INTERNATIONAL

ĐẠI HỌC QUỐC GIA HÀ NỘI
ĐHQGHN VNU Since 1906

TPE : TRAVAUX PERSONNEL ENCADRÉ

RAPPORT FINAL

IFI-Promotion :23

THÈME : Deep Learning pour le marché boursier

Rédiger par :

M. LAMAH Richard

Encadrant :

Dr. HO Tuong Vinh

Table des matières

I Analyse du Sujet :

I.1 Introduction :

Les marchés financiers sont considérés comme le cœur de l'économie mondiale, dans laquelle des milliards de dollars sont échangés chaque jour. De toute évidence, une bonne prévision du comportement futur des marchés serait extrêmement utile pour les opérateurs. Cependant, en raison du comportement dynamique et bruyant de ces marchés, cette prévision est également une tâche très difficile qui fait l'objet de recherches depuis de nombreuses années. Outre la prévision de l'indice boursier, la prévision du taux de change des devises, le prix des produits de base et les crypto-monnaies telles que le bitcoin sont des exemples de problèmes de prévision dans ce domaine (Shah et Zhang, 2014 ; Zhao et al., 2017 ; Nassirtoussi et al., 2015 ; Lee et al., 2017). Les approches existantes en matière d'analyse des marchés financiers se répartissent en deux groupes principaux d'analyse fondamentale et d'analyse technique. En analyse technique, historique les données du marché cible et certains autres indicateurs techniques sont considérés comme des facteurs de prévision importants. Selon l'hypothèse d'efficience du marché, le prix des actions reflète toutes les informations les concernant (Fama, 1970), tandis que les analystes techniques estiment qu'il est possible de prévoir le comportement futur des prix sur un marché en analysant les données de prix précédentes.Pour ce là, il est important de mener des études sur ce marché d'où l'objectif de ce TPE dont le sujet est intitulé « Deep Learning pour le marché boursier »

Contexte

Étudier les méthodes, technique et outils pour le marché boursier et de construire un prototype.

I.2 Domaine d'Étude :

Notre sujet traite le Deep Learning pour le marché boursier. C'est un domaine d'étude vague qui fait appel, outre l'informatique, à des connaissances très diverses. Dans le contexte de notre sujet nous aurons principalement à utiliser les notions en Intelligence Artificielle, en programmation orientée objet, en économie,en finance et aussi en mathématiques.

I.3 LES ALGORITHMES EXISTANTS TRAITANT DU DEEP LEARNING POUR LE MARCHE BOURSIER :

Les algorithmes d'apprentissage profond sont appelés réseaux neuronaux.Ce sont des modèles mathématiques.Ils reflètent les neurones du cerveau humain.Dans le cerveau,des ensembles de neurones apprennent à reconnaître certains modèles ou phénomènes, comme des visages, des ponts ou des séquences grammaticales.Ces modèles traitant du deep learning pour le marché boursiers sont :

— DNN (Deep Neural Networks),
— CNN (Convolutional deep Neural Networks),
— RNN (Recurrent Neural Networks)

I.4 PROBLÈMES À RÉSOUDRE :

Le but du présent projet est d'étudier la modélisation du mouvement de l'action et de renforcer la tendance à prédire si les modèles de prix d'une action monte ou baisse le jour de bourse suivant.

I.5 Plate-forme et Modèle à Utiliser

Nous utiliserons le langage python pour l'implémentions du prototype.C'est un langage de programmation objet interprété, multi paradigme et multi plate-

formes. IL favorise la programmation impérative structurée,fonctionnelle et offrant des outils de haut niveau.

I.6 Problème a prévoir :

Mon projet sera , comme tout travail scientifique , sujet à des difficulté de diverses ordres. Parmi lesquelles nous pouvons citer les problèmes liés :

1. Manque de connaissances dans le domaines l'économie et des finances.
2. A l'accès à la documentation
3. A la prise en main et la maîtrise du langage python
4. A la définition des spécifications du prototype

I.7 Travaux à réaliser :

I.7.1 Travaux Théorique :

— Faire un état de l'art des techniques et outils pour l'application de Deep Learning au marché boursier et regrouper les approches en fonction des méthodes et techniques utilisées.

— Faireune synthèse des avantages et inconvénients des méthodes les plus pertinentes de l'état de l'art et proposer un prototype avec les données réelles.

I.7.2 Travaux Pratique :

— Développer le prototype

— Établir une liste des bonnes pratiques

I.8 Résultats attendus et prototype

IL est attendus au terme de ce TP de développer un prototype et établir une liste des bonnes pratiques

I.9 Définition des mot clés :

— **Deep Learning :** ou apprentissage profond est un type d'Intelligence Artificiel dérivé du machine learning (apprentissage automatique) où la machine est capable d'apprendre par elle-même, contrairement à la programmation où elle se contente d'exécuter à la lettre des règles prédéterminées.

— **Prototype :** un modèle original qui possède toutes les qualités techniques et toutes les caractéristiques de fonctionnement d'un nouveau produit. mais il s'agit aussi parfois d'un exemplaire incomplet (et non définitif) de ce que pourra être un produit (éventuellement de type logiciel, ou de type « service ») ou un objet matériel final.

— **Multi-paradigme :** Un paradigme de programmation est une façon d'approcher la programmation informatique et de traiter les solutions aux problèmes et leur formulation dans un langage de programmation approprié. Il s'oppose à la méthodologie qui est une manière d'organiser la solution des problèmes spécifiques du génie logiciel.

— **Multi plateforme :**Un logiciel multi plateforme est un logiciel conçu pour fonctionner sur plusieurs plateformes, c'est-à-dire le couple liant ordinateur et système d'exploitation

II RECHERCHE BIBLIOGRAPHIQUE

II.1 Introduction

Dans un marché efficient, les acteurs financiers font continuellement du développement de stratégies, permettant de balancer un profit maximum avec un contexte de risques définis. Afin de tendre vers leurs objectifs, ils doivent souvent anticiper les mouvements et utilisent trois approches complémentaires : l'analyse fondamentale, l'analyse technique et les prévisions de séries temporelles. Les outils mis à la disposition de l'analyste ne doivent plus être contemplatifs, mais actifs en prévenant, avec suffisamment de recul, les modifications de tendances sur la base de critères bien assimilés par son utilisateur.

II.2 L'Analyse Fondamentale

[6] consiste à récolter puis uniformiser des données financières, issues des bilans ou des comptes de résultat, afin d'évaluer la valeur intrinsèque d'une entreprise et la comparer à celle de son secteur. Les éléments d'étude récurrents ont trait à la rentabilité (bénéfice net, bénéfice par action, chiffre d'affaires, etc.) ainsi qu'à la structure des entreprises (ratio de fonds propres, besoins en fonds de roulement, part de l'actif circulant, etc.).L'objectif de l'analyse fondamentale est de positionner une valeur dans un marché en expliquant le "pourquoi" de la cote.

II.3 L'analyse technique

[6]Est une méthode visant à prédire les futures tendances des actifs des différents marchés financiers. Comme son nom l'indique, l'analyse technique ne se base que sur l'aspect technique des choses et n'utilise donc que les aspects graphiques et les données historiques mises à disposition. L'objet du chartisme est dans la mesure

du possible de prévoir "comment" une valeur va évoluer à court/moyen terme.

II.4 L'Analyse de données séries temporelles

[7]Les séries temporelles constituent une branche de l'économétrie dont l'objet est l'étude des variables au cours de temps. Parmi ses principaux objectifs figurent la détermination des tendances au sein de ces séries ainsi que la stabilité des valeurs (et de leur variation) au cours de temps. On distingue notamment les modèles linéaires et les modèles non linéaires.

II.4.1 Les différents modèles linéaires

Sont AR, ARMA, ARIMA et ses variations. Ces modèles utilisent des équations prédéfinies pour ajuster un modèle mathématique à une série temporelle uni-variée. Le principal inconvénient de ces modèles est qu'ils ne tiennent pas compte de la dynamique latente existant dans les données. Comme ils ne considèrent que des séries chronologiques uni-variées, les interdépendances entre les différents stocks ne sont pas identifiées par ces modèles. De même, le modèle identifié pour une série ne convient pas pour l'autre. Pour ces raisons, il n'est pas possible d'identifier les modèles ou les données dynamiques présentes dans l'ensemble des données.

II.4.2 Les modèles non linéaires :

Impliquent des méthodes comme ARCH, GARCH, TAR, algorithmes d'apprentissage en profondeur. Le travail proposé se concentre sur l'application d'algorithmes d'apprentissage profonds pour la prédiction du cours des marchés.

II.5 ÉTAT D'ART :

Une architecture d'apprentissage en profondeur est inspirée par les réseaux neuronaux biologiques et se compose de plusieurs couches dans un réseau neuronal artificiel composé de matériel et de GPU.L'apprentissage en profondeur utilise une cascade de couches d'unités de traitement non linéaires afin d'extraire ou de transformer les caractéristiques (ou représentations) des données. La sortie d'une couche sert d'entrée de la couche suivante. Dans l'apprentissage en profondeur, les algorithmes peuvent être supervisés et servir à classer les données, ou non supervisés et à effectuer une analyse de modèle. Parmi les algorithmes d'apprentissage machine actuellement utilisés et développés, l'apprentissage en profondeur absorbe le plus de données et a été capable de battre les humains dans certaines tâches cognitives. En raison de ces attributs, l'apprentissage en profondeur est devenu l'approche avec un potentiel significatif dans le monde de l'intelligence artificielle. La reconnaissance faciale par ordinateur et la reconnaissance vocale ont toutes deux permis de réaliser des progrès significatifs grâce à des approches d'apprentissage approfondies [1] Et aussi d'autres architecture tels que le RNN(recurrent neural network),les DNN(Deep Neural Networks), le CNN(Convolutional Neural Networks),ont été utiliser dans le domaine de reconnaissance vocale, de l'écriture, de reconnaissance d'images, la manipulation mathématique correcte pour transformer l'entrée en sortie, que ce soit une relation linéaire ou une relation non linéaire, analyse vidéo traitement du langage naturel. Suite à quelques travaux réaliser précédemment nous allons voir ce qui convient pour les marchés boursiers.

II.5.1 DNN(Deep Neural Networks) :

en français (réseau de neurones profonds) est une variété de l'ANN (Artificial Neural Network) qui ont plus d'une couche cachée. Le DNN en général, est une technologie conçue pour simuler l'activité du cerveau humain, en particulier la

reconnaissance de formes et le passage des entrées à travers différentes couches de connexions neuronales simulées. Chong, Eunsuk, Chulwoo , et Frank [2] ont présenté une analyse systématique de l'utilisation des réseaux d'apprentissage en profondeur pour l'analyse et la prédiction du marché boursier. Leur étude tente de fournir une évaluation complète et objective des avantages et des inconvénients des algorithmes d'apprentissage en profondeur pour l'analyse et la prédiction des marchés boursiers. En utilisant les rendements intra journaliers haute fréquence comme données d'entrée, ils ont examiné les effets de trois méthodes d'extraction de caractéristiques non supervisées : Principal Component Analysis (PAC), auto encodé et la machine de Boltzmann sur la capacité globale du réseau à prédire le comportement futur du marché. Leur étude offre des perspectives pratiques et des pistes potentiellement utiles pour approfondir la question de savoir comment les réseaux d'apprentissage en profondeur peuvent être efficacement utilisés pour l'analyse et la prédiction des marchés boursiers. Appliqué au marché boursier coréen, Ils ont trouvé que les DNNs fonctionnent mieux qu'un modèle auto régressif linéaire dans l'ensemble d'entraînement. On peut le voir dans les résultats qu'ils ont obtenus dans le tableau ci-dessous :

Method	Data representation for DNN			
	(RawData)	(PCA380)	(RBM400)	(AE400)
AR(10)		0.9655		
ANN	0.9937	0.9990	0.9982	0.9976
DNN	0.9629	0.9660	0.9702	0.9638
AR-DNN	0.9622	0.9625	0.9628	0.9621
	(−0.0033)	(−0.0030)	(−0.0027)	(−0.0034)
DNN-AR	0.9643	0.9650	0.9682	0.9648
	(0.0013)	(−0.0010)	(−0.0020)	(0.0010)

Figure 1 : Erreur calculé sur l'ensemble des test effectué afin d'examiner les trois

méthodes d'extraction PAC, auto encoder et la machine de Boltzmann [2]

II.5.2 Convolutional Neural Networks (CNN)

en français (réseaux de neurones convolutionnels) récemment été appliqué pour la sélection automatique de caractéristiques et la prévision du marché. Dans un article publié sur le site https ://www.researchgate.net Ehsan Hoseinzade et Saman Haratizadeh [3] ont définit un framefwork CNNpred (CNN-based stock market prediction using several data sources) sur CNN avec des CNN spécialement conçus, pouvant être appliqués à une collection de données provenant de diverses sources, y compris de différents marchés, afin d'extraire des caractéristiques permettant de prédire l'avenir de ces marchés. Leur cadre suggéré a été appliqué pour prédire la direction du mouvement du lendemain pour les indices des marchés S P 500, NASDAQ, DJI, NYSE et RUSSELL reposent sur divers ensembles de caractéristiques initiales.

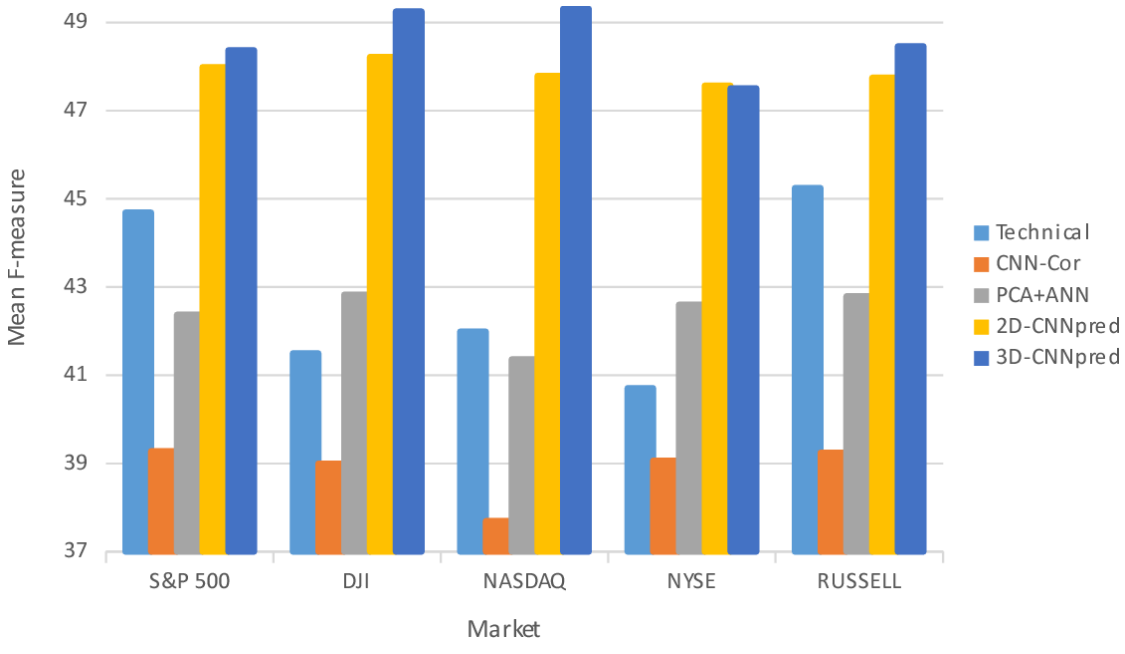


Figure 2 – F-mesure moyenne de différents algorithmes sur différents marchés [3]
Les résultats finaux (voire le schéma ci-dessus) ont montré la supériorité significative de deux versions de CNNPred par rapport aux algorithmes de base ultra-modernes. CNNpred a pu améliorer les performances de la prévision dans les cinq indices par rapport aux algorithmes de base d'environ 3% à 11%, en termes de Fmesure (méthode d'évaluation,(Gunduz et al., 2017 ; Ozgur et al., 2005). En plus de confirmer l'utilité de l'approche suggérée, ces observations suggèrent également que la conception des structures des CNN pour les problèmes de prévision des stocks est peut-être un défi fondamental qui mérite d'être approfondi.

II.5.3 Recurrent Neural Network :

Dans cet article [4] on propose un l'algorithme d'apprentissage LSTM , un type de RNN permettant de prédire le prix de fin de journée des données du stock Alcoa Corp (obtenue auprès de deux sources principales : Yahoo Finance et Google Finance.)

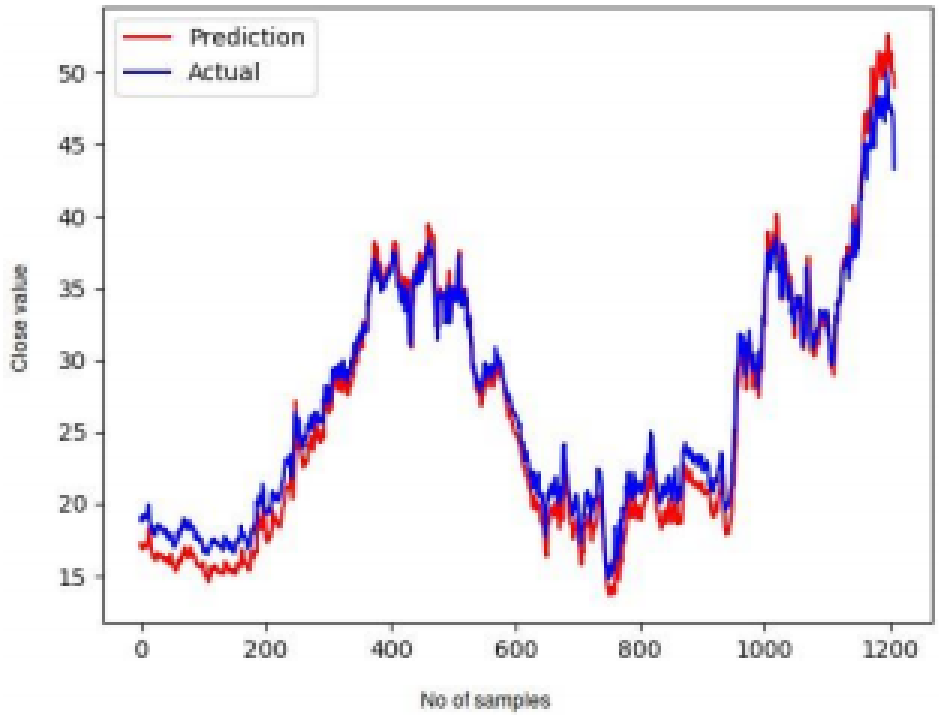


Figure 3 : Prédiction du marché de Alcoa Corpb[4]

La figure 9 présente le cours de clôture actuel et prévu de la société Alcoa Corp, un titre de grande taille. Le modèle a été formé avec une taille de lot de 512 et 50 époques, et les prédictions effectuées correspondent étroitement aux prix réels des actions, comme le montre le graphique.

Data Size	Stock Name	**LSTM (RMSE)**	**ANN (RMSE)**
Small	Dixon Hughes	**0.04**	**0.17**
Medium	Cooper Tire & Rubber	**0.25**	**0.35**
Medium	PNC Financial	**0.2**	**0.28**
Large	CitiGroup	**0.02**	**0.04**
Large	Alcoa Corp	**0.02**	**0.04**

Figure 4 – Comparaison des taux d'erreur (RMSE) de LSTM et ANN [4]

Les résultats de la comparaison entre la mémoire à long terme à long terme (LSTM) et le réseau de neurones artificiels (ANN) montrent que LSTM a une meilleure précision par rapport à ANN.L'analyse des résultats indique également que les deux modèles donnent une meilleure précision lorsque la taille de l'ensemble de données augmente. Avec plus de données, les poids des couches peuvent être mieux ajustés. La performance du système de prévision des stocks proposé, qui utilise un modèle LSTM, a été comparée à un modèle simple de réseau de neurones artificiels (ANN -Artificial Neural Network) sur cinq stocks différents de différentes tailles de données. ‘

II.5.4 Hybrid Deep Learning Model :

Le modèle [5]propose de développer un modèle d'apprentissage en profondeur hybride Hybrid Deep Neural Network (HDNN) qui sera une intégration du deep

learning et du Fuzzy logic systems. Le deep learning présente de nombreux avantages par rapport aux autres modèles d'apprentissage. Mais ils ont des limites qu'ils ne peuvent pas gérer des données incertaines. Les Fuzzy logic systems ayant la capacité de gérer des données incertaines, ils ne peuvent cependant pas apprendre des exemples. L'intégration de ces deux modèles ont permis de produire un meilleur résultat dans l'analyse prédictive

Table -2 Training Performance

	Infosys	SBI	ONGC	Tata Motors	Reliance Ind	Adani Ent	Future Retail
MSE	0.0048	0.0038	0.0041	0.005	0.004	0.0055	0.0044
RMSE	0.06	0.085	0.074	0.065	0.054	0.049	0.072
MAE	0.058	0.043	0.078	0.095	0.1	0.08	0.074

Table -3 Accuracy in Prediction

	Run 1	Run 2	Run 3	Run 4	Run 5	
	Correct (%)	Correct (%)	Correct (%)	Correct (%)	Correct (%)	Average Accu-
Infosys	96.90	97.10	97.50	97.80	96.80	97.22
SBI	98.20	97.90	98.50	98.00	97.90	98.10
ONGC	98.50	97.90	98.00	98.10	97.60	98.02
Tata Motors	97.50	96.80	97.90	97.20	96.85	97.25
Reliance Ind	97.40	98.60	98.10	97.85	98.10	98.01
Adani Ent.	98.20	98.40	97.95	97.55	98.80	98.18
Future Retail	95.30	96.10	95.90	97.00	96.95	96.25

Les figures (2-8) montrent [5] que le modèle HDNN a donné un très bon résultat de prédiction sur les 7 actions. Les chiffres ci-dessus représentent les tracés des valeurs prédites et les valeurs réelles correspondantes. Dans le tableau 3, nous avons représenté l'exactitude de la prédiction par le modèle HDNN pour les 7 actions en 5 exécutions. Le tableau -3 montre également que HDNN a donné une très bonne précision dans la prévision des prix de clôture des actions à la journée. Sa précision moyenne est toujours autour de 97% ou plus. Pour Future Retail, il a donné une précision moyenne de 96,25%, mais il est aussi un très bon résultat.

II.6 Definition des Termes Dificiles :

— Deep Learnning : en français [8]« apprentissage approfondi » est un ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction des données grâce à des architectures articulées de différentes transformations non linéaire. Ces techniques ont permis des progrès importants et rapides dans les domaines de l'analyse du signal sonore ou visuel et notamment de la reconnaissance faciale, de la reconnaissance vocale, de la vision par ordinateur, du traitement automatisé du langage. Dans les années 2000, ces progrès ont suscité des investissements privés, universitaires et publics importants, notamment de la part des GAFA (Google, Apple, Facebook, Amazon).

— Prédiction :[9]Un marché prédictif, ou marché de prédiction, est un marché sur lequel des agents s'échangent des produits dont la valeur dépend de la réalisation d'évènements comme des élections, des cours boursiers ou des résultats sportifs. Il repose sur des modèles prédictifs, algorithmes d'analyse statistique qui servent à prédire des évolutions de valeurs.

II.7 Analyse et solution Proposée :

La prévision du prix des actions ou des ventes est une approche hautement incertaine et lucrative. Il existe de nombreuses méthodes et outils utilisés à cette fin. Plusieurs méthodes d'apprentissage automatique ont été développées et sont utilisées dans l'intelligence d'affaires pour gagner plus de profit sur le marché de la bourse et vente. Ils produisent des résultats qui sont la base de la croissance de toute organisation. Dans un biotope hautement compétitif, les meilleurs résultats auront de meilleures chances de survie. Il existe plusieurs méthodes utilisées dans ce domaine, qui ne peuvent êtres efficaces. Les réseaux neuronaux récurrents sont

excellents à utiliser avec l'analyse des séries temporelles pour prédire les cours boursiers. En effet, l'analyse de séries temporelles comprend des méthodes d'analyse de données de séries temporelles afin d'extraire des statistiques significatives et d'autres caractéristiques des données. La prévision de série chronologique est l'utilisation d'un modèle pour prédire des valeurs futures basées sur des valeurs précédemment observées. Une prédiction plus précise sera obtenue en utilisant le modèle RNN-LSTM[4]

II.8 Conclusion Recherche Bibliographique :

Cette recherche bibliographique sur le sujet Deep Learning des marchés boursiers nous à permis de savoir les algorithmes utilisés sur ce domaine et lequel de ces algorithmes est adapté a notre sujet et qui donne une approche de solution adapté.

III SOLUTION PROPOSÉE

III.1 Introduction :

La prévision est souvent considérée comme l'aspect le plus problématique de la gestion, mais il est possible d'établir de bonnes prévisions (précises, fiables) grâce à des méthodes appropriées, qu'il faut avoir confiance et ne pas avoir peur de les utiliser. Ces méthodes sont des modèles du réseau de neurones artificiels qui ont été utilisés dans la prédiction du cours des actions avec leur propre force et limitation. Après l'amélioration et l'avancement dans le domaine de l'analyse prédictive, les techniques d'apprentissage en profondeur sont maintenant en vogue. Ils sont plus capables et ont de meilleures performances que les techniques conventionnelles d'apprentissage automatique. Des techniques populaires d'apprentissage en

profondeur comme des variantes de réseaux neuronaux profonds ont été utilisées dans la prévision boursière et leur résultat a été bien meilleur que les techniques conventionnelles d'apprentissage automatique ou les techniques statistiques. Dans notre cas d'étude, nous allons utiliser le RNN-LSTM

III.2 Fonctionnement du RNN - LSTM

Un modèle de séquence est généralement conçu pour transformer une sséquence d'entrée en une sséquence de sortie située dans un domaine différent. Le réseau de neurones récurrent, abréviation de « RNN », convient à cet effet et a permis d'améliorer considérablement des problèmes tels que la reconnaissance de lécriture manuscrite, la reconnaissance de la parole et la traduction automatique (Sutskever et al. 2011, Liwicki et al. 2007). Un modèle de réseau neuronal récurrent est ne avec la capacité de traiter des données sséquentielles longues et d'aborder des tâches dont le contexte s'étend dans le temps. Le modèle traite un élément de la séquence a un moment donné. Après le calcul, l'état nouvellement mis à jour est transmis au pas de temps suivant pour faciliter le calcul du prochain élément.

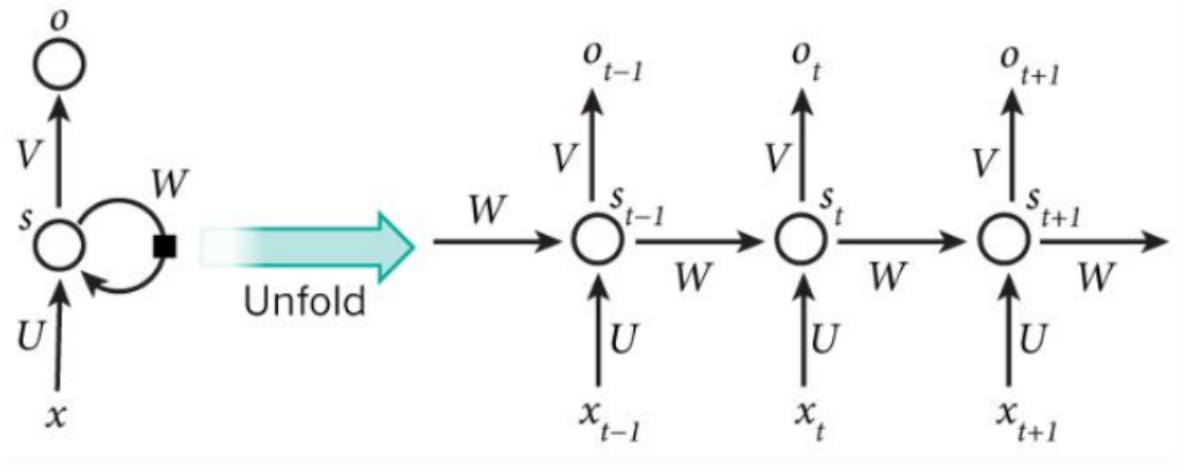


Figure 1 –

Un réseau de neurones récurrent avec une unité cachée (à gauche) et sa version déroulante dans le temps (à droite). La version déroulante illustre ce qui se passe dans le temps : st-1, st et st+1 sont la même unité avec des états différents a différents pas de temps t-1, t et t+1. (Source de l'image : Leçon, Bengio et Hinton,2015 ; Fig. 5). Cependant, les neurones de perceptron simples combinant linéairement l'élément d'entrée actuel et le dernier état d'unité peuvent facilement perdre les dépendances à long terme. Par exemple, nous commençons une phrase par « Alice travaille a... » et plus tard, après tout un paragraphe, nous voulons commencer correctement la phrase suivante par « Elle » ou « Il ». Si le modèle oublie le nom du personnage « Alice », nous ne pourrons jamais le savoir. Pour résoudre le problème, les chercheurs ont créé un neurone spécial doté d'une structure interne beaucoup plus compliquée pour la mmémorisation du contexte à long terme, appelé cellule « Mémoire à Long Terme » (LSTMs). Il est assez intelligent pour savoir pendant combien de temps il doit mémoriser les anciennes informations, quand oublier, quand utiliser les nouvelles données et comment combiner l'ancienne mémoire avec les nouvelles entrées.

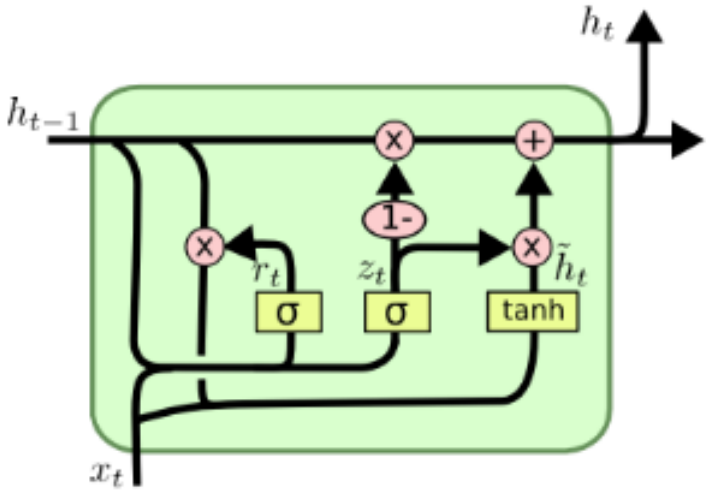


FIGURE 2 –

Figure 15 – La structure d'une cellule LSTM. (Image source : http ://colah.github.io/posts/2015-08-Understanding-LSTMs/)

III.2.1 Le module de répétition dans un LSTM contient quatre couches en interaction :

Nous allons parcourir le diagramme LSTM, étape par étape. Pour l'instant, essayons simplement de nous familiariser avec la notation que nous allons utiliser.

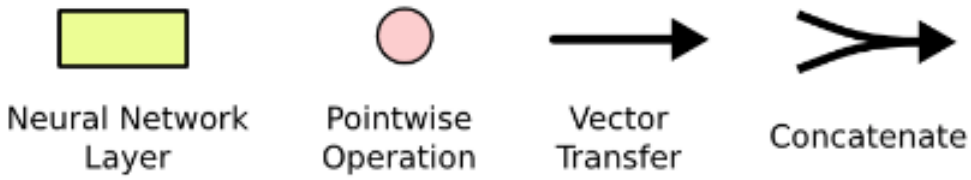


FIGURE 3 –

Figure 7 – La structure d'une cellule LSTM. (Image source : http ://colah.github.io/posts/2015-08-Understanding-LSTM

Dans le diagramme ci-dessus (figure 2), chaque ligne porte un vecteur entier, de la sortie d'un noeud aux entrées des autres. Les cercles roses représentent des opérations ponctuelles, telles que l'addition de vecteurs, tandis que les zones jaunes représentent les couches de réseau neural apprises. La fusion de lignes indique une concaténation, tandis qu'une ligne de bille indique son contenu en cours de copie et les copies envoyées à des emplacements différents.

III.3 Ensemble de données

Nous utiliserons les données historique de cinq années donc pour le prix des action de google entre 2012 et jusqu'a la fin de 2016, donc notre objectifs va être à partir de ces données des cinq années d'historique essayer de predire le prémier mois de l'année 2017 et on va essayer de prédire notamment la tendance à savoir est ce que l'action va montée ou bien elle va descendre.

Les données d'apprentissage du système disponible comprendront 80% d'échantillons. De plus, l'efficacité du système sera démontrée en utilisant 20% pour tester et validation les données.

III.4 Démarche

Nous utiliserons RNN (Recurrent neural networks) – LSTM (Long-short term memory) pour implémenter notre modèle. Les LSTM sont bien adaptés à cette tâche du fait de leur capacité à apprendre de l'expérience.

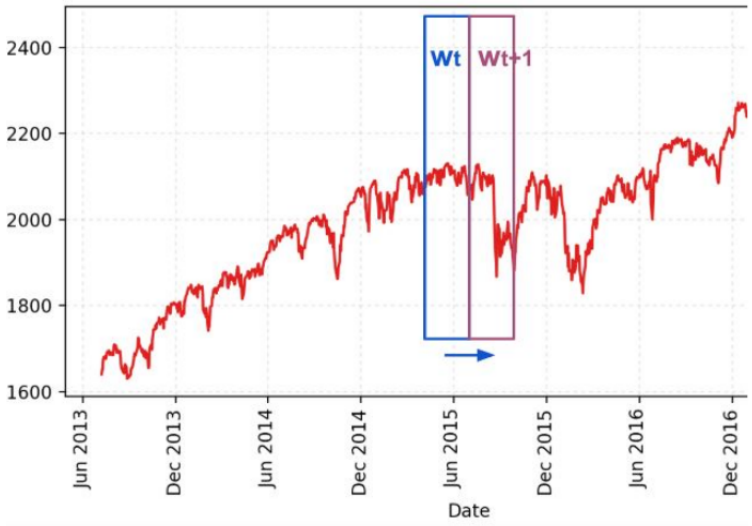


FIGURE 4 –

Le cours des actions est une série chronologique de longueur N définit comme p0,p1,..pN-1 où pi est le prix de clôture de la journée i , 0 ¡ i ¡ N.Nous utilisons du contenu dans la fenêtre courante pour faire de la prédiction pour la suivante :

$$W_{t+1} = (p_{(t+1)w}, p_{(t+1)w+1}, \ldots, p_{(t+2)w-1})$$

sachant que :

$$\begin{aligned} W_0 &= (p_0, p_1, \ldots, p_{w-1}) \\ W_1 &= (p_w, p_{w+1}, \ldots, p_{2w-1}) \\ &\ldots \\ W_t &= (p_{tw}, p_{tw+1}, \ldots, p_{(t+1)w-1}) \end{aligned}$$

FIGURE 5 –

Nous essayons d'apprendre une fonction d'approximation

$$f(W_0, W_1, \ldots, W_t) \approx W_{t+1}.$$

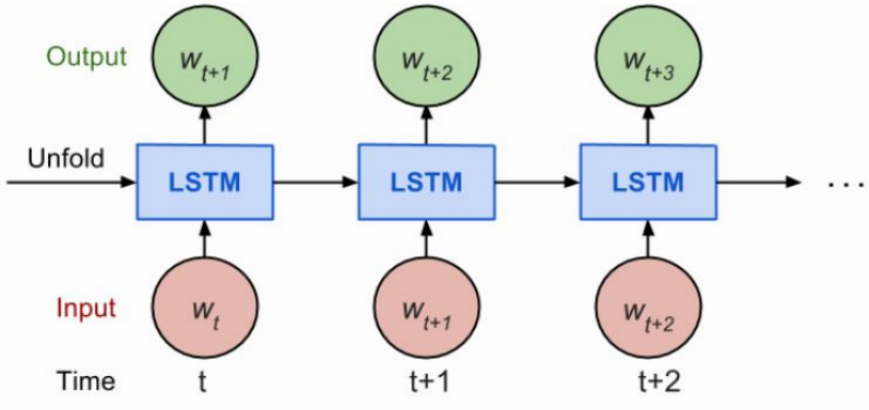


FIGURE 6 –

La séquence de prix est d'abord divisée en petites fenêtres sans chevauchement.Chacun contient des nombres d'input_size et chacun est considéré comme un élément d'entrée indépendant. Ensuite, tous les num_steps d'éléments d'entrée consécutifs sont regroupés en une entrée d'entraînement, formant une version «non déroulée » de RNN pour l'entraînement sur Tensorfow. L'étiquette correspondante est l'élément d'entr

III.5 MÉTHODOLOGIE DE RECHERCHE :

La méthodologie qui sera utilisée pour conduire la recherche proposée comprend les étapes principales illustrées sur le schema suivante :

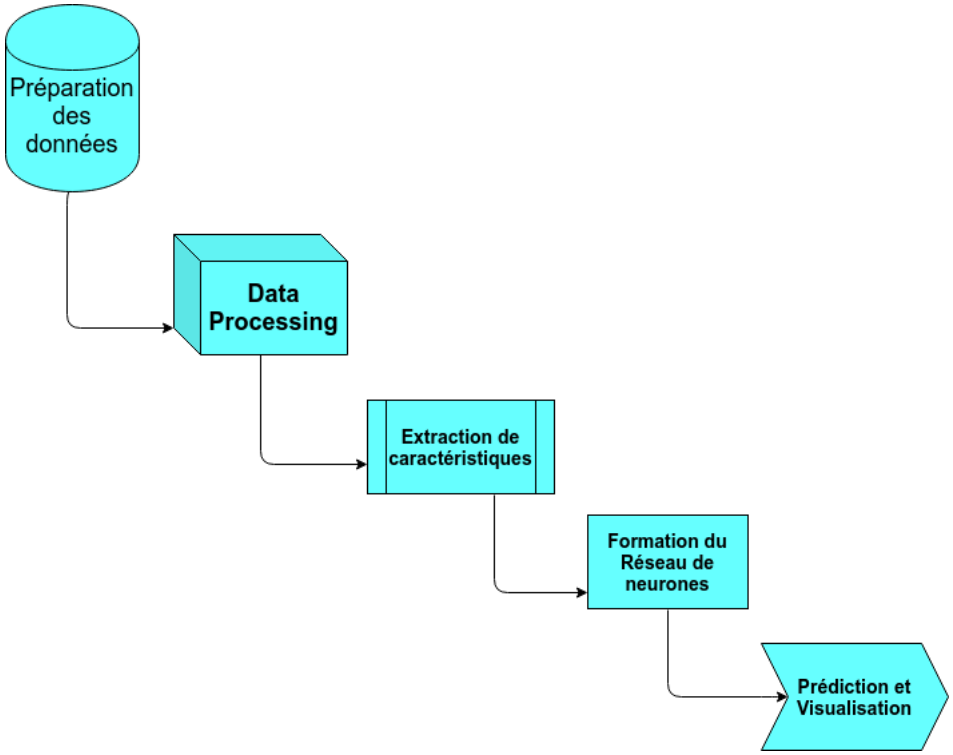


FIGURE 7 – Schema de methodologie

Notre système se compose de plusieurs étapes qui sont les suivantes :

III.5.1 Préparation des Données :

A ce niveau,nous allons collectées les données de stock historiques et ces données historiques sont utilisé pour la prédiction des prix des actions futures.

III.5.2 Data Preprocessing : L'étape de prétraitement implique :

— Discrétisation des données : partie de la réduction des données mais avec une importance particulière, en particulier pour les données numériques

— Feature Scaling : c'est l'étape de transformation de données : normalisation.

— Intégration de données : Intégration de fichiers de données.

III.5.3 Extraction de caractéristiques :

Dans cette couche, seules les caractéristiques qui doivent être introduites dans le réseau neuronal sont choisies.Nous allons choisir la fonctionnalité de Date, ouvrir.

III.5.4 Formation du Réseau de neurones :

A ce niveau, les données sont envoyées au réseau neuronal et entraînées pour la prédiction en attribuant des biais et des poids aléatoires. Notre modèle LSTM est composé d'une couche d'entrée séquentielle suivie de 2 couches LSTM et d'une couche dense avec activation ReLU et enfin d'une couche de sortie dense avec les fonctions d'activation linéaires.

III.5.5 Les données générés en sortie :

Dans cette couche, la valeur de sortie générée par la couche de sortie du RNN est comparée à la valeur cible.L'erreur ou la différence entre la cible et la valeur de sortie obtenue est minimisée en utilisant un algorithme de rétropropagation qui ajuste les poids et les biais du réseau.

III.6 OUTILS DE REALISATION

III.6.1 OUTILS :

Pour enseigner à notre machine comment utiliser les réseaux de neurones pour faire des prédictions, nous allons utiliser l'apprentissage en profondeur de TensorFlow . L'apprentissage en profondeur est un domaine d'apprentissage automatique qui utilise des algorithmes inspirés de la façon dont les neurones fonctionnent dans le cerveau humain. TensorFlow est un framework d'apprentissage automatique que Google a créé pour concevoir, créer et former des modèles d'apprentissage en profondeur. Le nom "TensorFlow" est dérivé de la façon dont les réseaux neuronaux

fonctionnent sur des tableaux de données ou des tenseurs multidimensionnels. C'est un flux ou des tenseurs, tout comme le cerveau humain a un flux de neurones ! Nous utiliserons également Keras, qui fonctionne sur TensorFlow. Et aussi Theano est une bibliothèque Python qui vous permet de définir, d'optimiser et d'évaluer efficacement les expressions mathématiques impliquant des tableaux multidimensionnels.

III.6.2 Introduction à Trensorflow :

TensorFlow est une bibliothèque de logiciels open source pour le calcul numérique haute performance. Son architecture flexible permet un déploiement aisé des calculs sur diverses plates-formes (processeurs, GPU, TPU), des ordinateurs de bureau aux clusters de serveurs, en passant par les périphériques mobiles et périphériques. Développé à l'origine par des chercheurs et des ingénieurs de l'équipe Google Brain au sein de l'organisation d'intelligence artificielle de Google, il intègre un support puissant pour l'apprentissage automatique et approfondi. Le cœur du calcul numérique flexible est utilisé dans de nombreux autres domaines scientifiques.

III.7 Conclusion Solution Proposée :

Ainsi nous essayerons de proposer une méthode efficace d'analyse de prédiction qui nous permettra d'avoir de meilleurs résultats que d'autres modèles de prédictions existants.

IV TRAVAIL PRATIQUE

IV.1 Introduction

Dans cette partie, nous présentons la prévision de prix future pour différents titres utilisant des réseaux de neurones récurrents avec Tensorflow.

IV.2 Implementation

IV.2.1 Analysé des Données :

***importation des librairies**

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import keras
import warnings
import tensorflow as tf
warnings.filterwarnings('ignore')

import os
```

FIGURE 8 – Librairies

***Chargement des Données**

Nous allons charger les données boursières obtenues afin de l'utiliser dans notre algorithme.

```
stock_data = pd.read_csv("Google_Stock_Price_Train.csv")
test_data = pd.read_csv("Google_Stock_Price_Test.csv")
```

FIGURE 9 – jeux de données

Échantillon de données

```
stock_data.tail()
```

	Date	Open	High	Low	Close	Volume
1253	2-Oct-12	382.98	383.38	375.51	378.87	2790375
1254	1-Oct-12	379.90	382.88	378.48	381.27	3168477
1255	28-Sep-12	377.45	380.03	375.95	377.63	2784091
1256	27-Sep-12	380.35	381.80	376.20	378.63	3932272
1257	26-Sep-12	375.30	381.00	370.87	377.11	5674334

FIGURE 10 – échantillon jeux de données

***Analyse**

Les données chargées comprennent les variables suivantes :

Volume :nombre de titres, d'actions échangés, au jour et à l'heure de la consultation de la fiche de l'action.

Open / High / Low / Close : Il s'agit des cours d'ouverture, le cours le plus haut et le plus bas de la journée, ainsi que le cours auquel l'action a clôturé la veille.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1258 entries, 0 to 1257
Data columns (total 6 columns):
Date      1258 non-null object
Open      1258 non-null float64
High      1258 non-null float64
Low       1258 non-null float64
Close     1258 non-null float64
Volume    1258 non-null int64
dtypes: float64(4), int64(1), object(1)
memory usage: 59.0+ KB
```

FIGURE 11 – Détail sur les variables

— **Variation des prix(Open,Close,Low, High) en fonction du temps (en jour)**

FIGURE 12 – Variation des prix(Open,Close,Low, High) en fonction du temps (en jour)

— **Variation du volume en fonction du temps (en jour)**

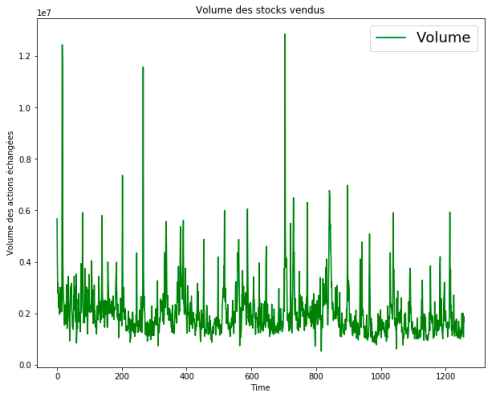


FIGURE 13 – Variation du Volume en fonction du temps (en jour)

IV.2.2 Manipulation des données (normalisation/création des données d'apprentissage, validation et test)

*** La normalisation :**L'indice Google augmente dans le temps, entraînant ce problème : la plupart des valeurs de l'ensemble de test sont hors de l'échelle des données d'entraînement et que le modèle doit donc prévoir certains nombres qu'il n'a jamais vus auparavant. Pour résoudre le problème hors échelle, on normalise les prix. La normalisation peut- être appliquée par le **min-max scaling**. Python propose pour cela une classe nommée **"MinMaxScaler"** .Pour résoudre le problème hors échelle, on normalise les prix dans chaque fenêtre glissante. La tâche devient alors la prévision des taux de changement relatifs au lieu des valeurs absolues. Dans une fenêtre glissante normalisée W't au moment t,toutes les valeurs sont divisées par le dernier prix inconnu, le dernier prix en **Wt-1**

$$W'_t = \left(\frac{p_{tw}}{p_{tw-1}}, \frac{p_{tw+1}}{p_{tw-1}}, \ldots, \frac{p_{(t+1)w-1}}{p_{tw-1}}\right)$$

***Création des données d'entraînement , de test et de validation**

```
X_train =  (1198, 50, 5)
X_valid =  (10, 50, 5)
X_test  =  (100, 50, 5)
```

IV.2.3 Modèle et validation des données

Définitions :

- **Sequential** pour initialiser le réseau de neurones
- **Dense** pour ajouter une couche de réseau neuronal densément connectée
- **LSTM** pour l'ajout de la couche de mémoire à court terme

— **lookback :**créer les données pour remonter à un nombre de jours ouvrables dans le passé

— **Callbacks** est un ensemble de fonctions à appliquer à des étapes données de la procédure de formation.

Les fonctions de **Callbacks** sont :

— **EarlyStoping :** Cela arrêtera la formation si le score du modèle n'augmente pas. Cela empêche le modèle d'overfitting. Nous devons fixer le maximum à 10 époches si cela n'augmente pas, puis nous arrêterons l'entraînement.

— **ReduceLROnPlateau :** Utilisez pour réduire le taux d'apprentissage. En 3 étapes, le score n'a pas augmenté, nous allons réduire le taux d'apprentissage pour améliorer l'entraînement.

— **ModelCheckpoint :** Utiliser pour sauvegarder le modèle uniquement lorsque le score augmente

Nous ajoutons la couche LSTM avec les arguments suivants :

1. **50 unités** qui est la dimensionnalité de l'espace de sortie
2. **return_sequences=True** qui détermine s'il faut renvoyer la dernière sortie de la séquence de sortie ou la séquence complète
3. **input_shape** comme la forme de notre ensemble de formation.

Nous ajoutons la Dense couche qui spécifie la sortie de 1 unité. Après cela, nous compilons notre modèle à l'aide du très populaire adam optimizer et définissons la perte comme mean_squarred_error. Ceci calculera la moyenne des erreurs au carré. Nous adaptons ensuite le modèle à 100 époques avec une taille de lot de 32.

Le taux d'apprentissage est défini sur **Train** pendant les premières époques **Epoch**, puis décroît à chaque époque suivante.

Constructions :

— choisir les paramètres d'entraînement : Le modèle est entraîné avec input_size= 4 et lstm= 50

— les données sont envoyées au réseau neuronal et entraînées pour la prédiction en attribuant des biais et des poids aléatoires.

— La machine va passer en revue chaque données du jeux d'entraînement. Quand le jeux de données a été parcouru complètement une fois, on appel cela une **'epoch'.**

— Le modèle est entraîné sur le jeu d'entraînement et il est ensuite optimisé en fonction des performances du jeu de validation.

— la valeur de sortie générée par la couche de sortie du RNN est comparée à la valeur cible. L'erreur ou la différence entre la cible et la valeur de sortie obtenue est minimisée en utilisant un algorithme de rétro-propagation qui ajuste les poids et les biais du réseau.

```
Train on 1198 samples, validate on 10 samples
Epoch 1/100
1198/1198 [==============================] - 6s 5ms/step - loss: 0.0449 - acc: 0.0000e+00 - val_loss: 0.0058 - val_
acc: 0.0000e+00

Epoch 00001: val_loss improved from inf to 0.00576, saving model to model.h5
Epoch 2/100
1198/1198 [==============================] - 4s 3ms/step - loss: 0.0014 - acc: 8.3472e-04 - val_loss: 7.5402e-04 -
val_acc: 0.0000e+00

Epoch 00002: val_loss improved from 0.00576 to 0.00075, saving model to model.h5
Epoch 3/100
1198/1198 [==============================] - 4s 3ms/step - loss: 0.0012 - acc: 8.3472e-04 - val_loss: 1.9095e-04 -
val_acc: 0.0000e+00

Epoch 00003: val_loss improved from 0.00075 to 0.00019, saving model to model.h5
Epoch 4/100
1198/1198 [==============================] - 4s 3ms/step - loss: 0.0011 - acc: 8.3472e-04 - val_loss: 4.3565e-04 -
val_acc: 0.0000e+00

Epoch 00004: val_loss did not improve from 0.00019
Epoch 5/100
1198/1198 [==============================] - 4s 3ms/step - loss: 0.0011 - acc: 8.3472e-04 - val_loss: 6.9702e-04 -
val_acc: 0.0000e+00

Epoch 00005: val_loss did not improve from 0.00019
Epoch 6/100
1198/1198 [==============================] - 4s 3ms/step - loss: 0.0011 - acc: 8.3472e-04 - val_loss: 1.5076e-04 -
val_acc: 0.0000e+00

Epoch 00006: ReduceLROnPlateau reducing learning rate to 0.00010000000474974513.

Epoch 00006: val_loss improved from 0.00019 to 0.00015, saving model to model.h5
Epoch 7/100
1198/1198 [==============================] - 4s 3ms/step - loss: 9.8278e-04 - acc: 8.3472e-04 - val_loss: 7.6319e-0
```

FIGURE 14 – Entrainement du modèle

IV.2.4 Prédiction :

Enfin, il faut calculer la performance finale du modèle à l'aide d'un ensemble de test indépendant : le jeu de test.

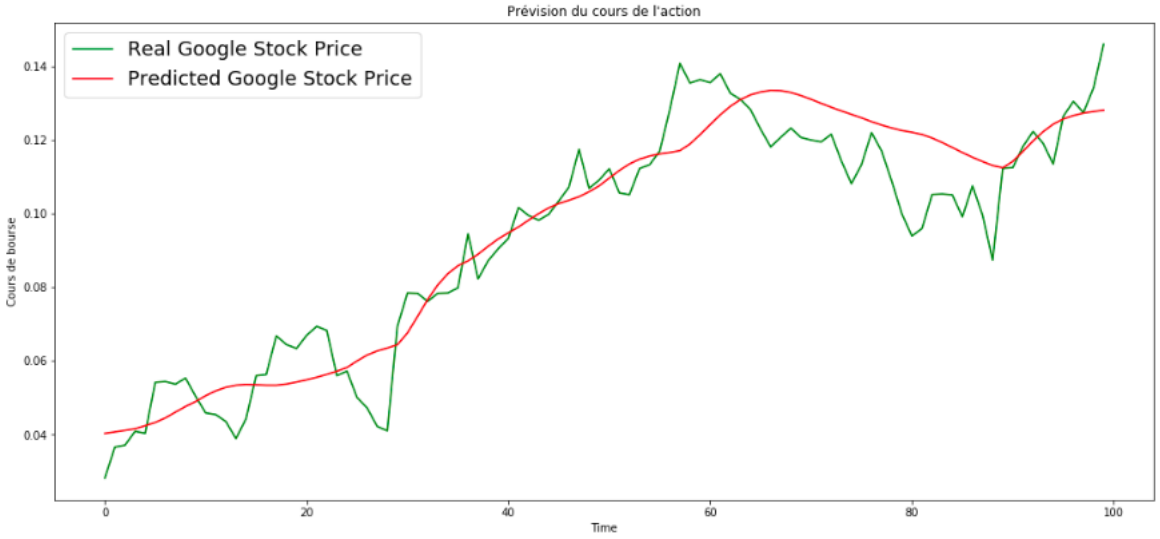
predicted_value = model.predict(X_test)

IV.3 RÉSULTATS OBTENUS :

On obtient le taux de prédiction correcte suivant :

— Taux de prediction = **50.27 %**

— MAE :mean absolute error :**0.04 degrees.**



IV.4 CONCLUSION

La mise en pratique de la solution proposé nous emmène à un résultat qui pourrait être meilleur que celui que nous avions trouvé du fait du choix des paramètres aléatoires. Ce qui nous renvoie à la question suivante : Comment régler les hyper paramètres LSTM pour les prévisions de séries temporelles ?

V CONCLUSION ET PERSPECTIVES

Nous proposons une formalisation basée sur l'apprentissage en profondeur pour la prévision du prix des actions. On voit que les architectures de réseaux neuronaux profonds sont capables de capturer des dynamiques cachées et de faire des prédictions. Nous avons formé le modèle à l'aide des données de **GOOGLE** et avons été en mesure de prévoir le cours des actions. Cela montre que le système proposé est capable d'identifier certaines relations dans les données. Par ces différents essais, on a traité exclusivement des données endogènes (pas toutes) mais il existe aussi des données exogènes ayant un impact sur la cour comme les corrélations avec d'autres actifs comme **l'euro/dollar, le pétrole,** corrélation sectorielle, etc.. Corrélation ne veux pas dire causalité, il faudra donc étudier ça de plus prêt. Une donnée exogène importante, mais qui à une grande part d'inconnu est le comportement des autres investisseurs.À la base, le marché boursier est le reflet des émotions humaines. Le calcul et l'analyse des nombres pures ont leurs limites ; Une extension possible de ce système de prévision des stocks consisterait à l'ajouter à une analyse des flux de nouvelles provenant de plateformes de médias sociaux telles que **Twitter**, où les émotions sont mesurées à partir des **articles**. Cette analyse des sentiments peut être liée au **LSTM** pour améliorer la formation des poids et améliorer la précision.

VI Références :

- ([1]) :https ://www.supinfo.com/articles/single/6041-machine-learning-introduction-apprentissage-automatique
- ([2]) :Chong, Eunsuk, Chulwoo Han, and Frank C. Park. "Deep learning net-works for stock market analysis and prediction : Methodology, data 27 representations, and case studies." Expert Systems with Applications 83(2017) : 187-205
- ([3]) :https ://www.researchgate.net/publication/328445708$_{C}NNPred_{C}NN-based_{s}tock_{m}arket_{p}rediction_{u}sing_{s}everal_{d}ata_{s}ources$

— ([4]) :Murtaza Roondiwala,Harshal Patel,Shraddha Varma.Predicting StockPrices Using LSTM.https ://www.ijsr.net/archive/v6i4/ART20172755.pdf

— ([5]) :http ://oaji.net/articles/2017/2000-1528794589.pdf

— ([6]) :https ://www.cafedelabourse.com/lexique/definition/analyse-fondamentale

— ([7]) :http ://www.sigmaplus.fr/logiciels/statistiques/analyses-statistiques/centurion/analyses-de-series-temporelles-et-previsions.html

— ([8]) :https ://fr.wikipedia.org/wiki/Apprentissage$_{p}rofond$

([9]) : $https : //fr.wikipedia.org/wiki/March\%C3\%A9\_de\_pr\%C3\%A9diction$

— https ://fr.wikipedia.org/wiki/Bourse_(%C3%A9conomie)

— https ://www.boursier.com/guide/debuter-en-bourse/comment-fonctionne-la-bourse

— https ://www.andlil.com/a-quoi-sert-la-bourse-143853.html

— https ://www.analyticsvidhya.com